Suicidal Posts Detection System Incorporating Psychological Risk Factors

Chih-Ning Chen ANDREWMAN71@GMAIL.COM
Department of Physics, National Tsing Hua University, Taiwan

Chieh-Jou Lin LINCHIEHJOU@GMAIL.COM
Department of Physics, National Tsing Hua University, Taiwan

Kun-Hua Lee KUNHUALE@GMAIL.COM
Institute of Educational Psychology and Counseling, National Tsing Hua University, Taiwan

Yu-Ping Ma YUPINGMA@GMAIL.COM
Department of Computer Science, National Tsing Hua University, Taiwan

Kuo-Liang Ou KUOLIANGOU@GMAIL.COM
Institute of Learning Sciences and Technologies, National Tsing Hua University, Taiwan

Daw-Wei Wang DWWANG@PHYS.NTHU.EDU.TW
Department of Physics, National Tsing Hua University, Taiwan
Center for the Applications and Developments of AI in Humanity and Social Sciences, National Tsing Hua University, Taiwan

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Our study aims to utilize psychological risk factors to detect posts on social media that contain high-risk suicidal content in Mandarin. We propose a two-stage model structure: the first stage labels each sentence in a post according to risk factors, while the second stage uses these labels as features to predict the crisis level of the post. Our models were trained using a dataset developed from social media posts on a popular Mandarin-speaking platform, labeled by psychological professionals. Our approach achieved an accuracy and F1-score of 0.96 in classifying posts with high crisis levels. Furthermore, we developed a frontend webpage system to apply our model, designed for use by psychological professionals as an aid. This system not only helps psychological professionals detect and address high-risk posts but also offers them the opportunity for psychological analysis based on risk factors. By integrating expertise from psychology with advanced natural language processing and deep learning techniques, our system bridges the gap between technical models and psychological insights.

Keywords: suicide detection; psychological risk factors; deep learning; natural language processing; social media analysis; Mandarin

1. Introduction

Recent deep learning techniques have utilized data from social media to detect suicide ideation (Sawhney et al., 2018; Un Nisa and Muhammad, 2021; Boonyarat et al., 2024). These studies have introduced innovative methods for modeling emotional context over time and incorporating social network information, significantly improving the process for

identifying at-risk users. However, these advancements often present challenges in further analytical exploration, making it difficult for consultants and psychological professionals to understand underlying causes due to the opaque, "black-box" nature of these models, which lacks interpretability (Von Eschenbach, 2021). This interpretability gap highlights a crucial disconnect between practical applications in psychology and cutting-edge natural language processing (NLP) techniques, posing challenges for deeper research into suicide detection.

To address the challenges outlined above and develop a practical suicidal detection system for psychological analysis, we proposed a two-stage model that predicts psychological risk factors at the sentence level and then determines the suicidal crisis level for entire posts. A distinctive feature of our research and system is the training approach: we not only train the detection model with labeled posts but also employ manually annotated psychological risk factors as sentence labels to develop a classification model. This method is rare in related research and systems because manually labeling sentences is extremely time-consuming, and it is debatable whether sentence labels significantly enhance the performance of suicidal detection models. However, our experimental results show that our system achieves high accuracy in post classification, reaching 0.96 with the aid of risk factors. Furthermore, we demonstrate that the accuracy improves significantly, by 0.09, with the inclusion of risk factors. These results indicate a clear correlation between psychological risk factors and suicidal crisis levels, highlighting the value of incorporating psychological knowledge into feature extraction.

We have successfully integrated sentence and post classification models into a web frontend, allowing users to submit posts for prediction. The system displays the results of sentence and post classifications along with relevant statistics and visualizations, creating a comprehensive online crisis post detection system for psychological professionals. Thanks to the efficiency of the BERT-based model (Devlin et al., 2018), which we chose as the pre-trained model for our system, the inference time is notably brief compared to large language models (LLMs). This efficiency offers the potential for integrating our system into social media platforms. Our contribution can be summarized as follows:

1. We developed a two-stage BERT-based model that predicts psychological risk factors and suicidal crisis levels in Mandarin, achieving a high performance with an accuracy of 0.96.

2. We have established a mature system integrated into a frontend webpage, designed specifically for psychological professionals to conduct further analysis.

3. Our experimental results demonstrate that psychological risk factors significantly impact the detection of suicidal content in Mandarin social media posts, underscoring the importance of incorporating psychological knowledge into suicidal detection, sentiment analysis, and natural language processing.

2. Related Work

Current suicide detection systems on social media are adept at calculating high performance in identifying potential suicidal content through various machine learning and deep learning techniques (Chen et al., 2020; Metzler et al., 2022; Akintoye et al., 2024) or proposed novel

models designed for suicide detection (Mishra et al., 2019; Sawhney et al., 2021). However, these systems often fall short in real-world applications as they lack the capability to analyze underlying causes or provide further insights into the psychological states of the users. Most systems only quantify performance without analyzing the context and psychology necessary for deeper applications, limiting their use beyond basic detection and failing to provide actionable insights for psychological research (Kirtley et al., 2022).

Some approaches attempt to enhance analytical depth by examining users' posting habits—such as language use (Sinha et al., 2019), posting frequency, or timing (Sawhney et al., 2020)—or by integrating questionnaire data (Martınez-Castano et al., 2020) and other online records (Wiest et al., 2024) to infer psychological states. While these methods aim to provide a more nuanced understanding, they struggle in real-world scenarios where consistently accessing such comprehensive data can be impractical. The use of LLMs can provide a deeper understanding of the semantics behind suicidal ideation (Alhamed et al., 2024). However, LLMs require relatively more resources and time for thorough analysis, which poses challenges for applications.

Our system extends beyond mere suicide detection by leveraging identified psychological risk factors to conduct in-depth analyses of users' mental states. This dual capability not only enhances the system's accuracy but also enriches its application across disciplines, providing valuable system for psychological researchers. By analyzing the content of high-risk posts in conjunction with risk factors, our system offers not just high accuracy but also efficient and practical psychological applications.

Current research using the Deep Learning model and train or apply on social media in general tasks reaches incredible performance (Chen et al., 2020). Our work focuses on suicide detection and further analysis in Mandarin. Previous research has explored various aspects of suicide detection, employing machine learning approaches (Azim et al., 2022; Tadesse et al., 2019; Ji et al., 2020). Recent trends show a shift towards deep learning techniques such as LSTM (Azim et al., 2022; Tadesse et al., 2019), BERT (Ji et al., 2020; Castillo-Sánchez et al., 2020), GPT (Bernert et al., 2020), and LLM (Izmaylov et al., 2023; Tanaka and Fukazawa, 2024). A primary challenge in this research is data labeling—professionally or psychologically classifying large volumes of sentences and posts is difficult. Additionally, these detection models often lack transparency, a common issue in NLP known as the 'black-box' phenomenon, which complicates their use in psychological analysis and research.

Our research focuses on suicide detection through psychological feature engineering in Mandarin. We collaborate with psychology professionals to label sentences and posts. By creating sentence-level classifications, we refine the performance of post classification models. Furthermore, these classifications allow psychologists to analyze content more deeply, tracing the intentions and logical reasoning behind suicidal ideation in posts. Our work integrates NLP, deep learning, and psychological expertise to advance suicide detection and support psychological research.

3. Dataset Description

3.1. Data source

Our initial dataset was obtained from Dcard, a popular Mandarin-speaking social media platform. We signed a formal contract with Dcard, securing authorization to crawl post data

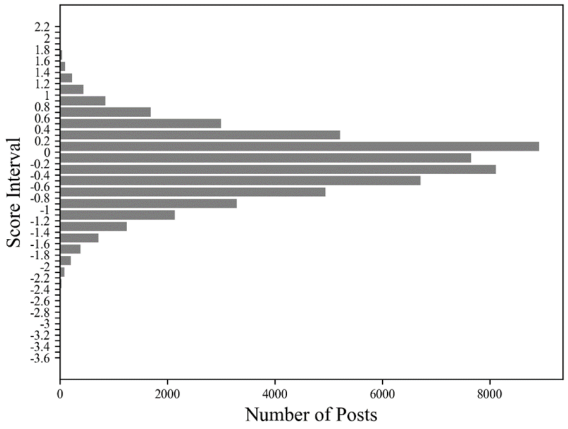

Figure 1: Distribution of posts by average mood scores from Dcard in 2019. The x-axis represents average mood scores (negative values indicate more negative content), and the y-axis shows the number of posts. Posts with scores below -1.4 were selected for human annotation as they exhibit the highest probability of containing high-crisis content. This threshold-based selection strategy ensures efficient annotation of the most relevant posts for suicide risk detection.

under a confidentiality agreement that prohibits the disclosure of any personal information in our research. In compliance with this agreement, we anonymize all data, removing any identifiers such as author names and corresponding IDs, before using it to train our models. We extract 55,989 posts from the 2019 Mood Diaries section to represent the young generation in the Mandarin-speaking community. Given the extensive data volume, we first gauged the mood intensity of these posts by calculating an average mood score—derived by dividing the total score by the number of words. The scores, based on the frequency of specific keywords, were analyzed using statistical methods and big data techniques alongside another dataset (NTUSD, Ku and Chen, 2007). These scores indicate the overall positive or negative mood conveyed by the keywords and their intensity with lower scores indicating more negative content.

3.2. Post Data Description

Crisis Level	A1 post	A2 post
level 3	95(7%)	72(6%)
level 2	200(14%)	118(9%)
level 1	457(32%)	312(25%)
level 0	672(47%)	738(60%)
total	1424(100%)	1240(100%)

Table 1: Post data statistics

We selected 1,424 posts with average scores below -1.4 for human evaluation (denoted by A1), as these likely contained a high percentage of messages with potential suicidal risk. Our team further annotated each sentence within these posts to identify and categorize risk factors. Additionally, we labeled the crisis level of another 1,240 posts (denoted by A2), which have average scores between -1.4 and -1.2. These posts served as test data and were not previously annotated with risk factors.

The crisis levels are categorized into four distinct groups based on the severity of suicidal risk: **Level 0** indicates no suicidal thoughts and no current problems; **Level 1** represents individuals who subjectively report suicidal thoughts and some crisis events but can still tolerate the disturbance caused by these thoughts; **Level 2** indicates individuals experiencing suicidal thoughts that are challenging to manage due to significant disturbance; and **Level 3** represents the highest level of suicidal crisis, where individuals report vivid and persistent suicidal thoughts alongside suicide attempts, indicating an inability to tolerate the suffering any longer.

Table 1 presents the statistics of the posts for each crisis level, including A1 and A2 posts.

3.3. Sentence Data Description

Sentence Label	Sentences
Neutral	34599(74.3%)
Suicidal thoughts and depression(SD)	3443(7.4%)
Negative cognition (NC)	279(0.6%)
Positive emotion (PE)	209(0.5%)
Negative emotion (NE)	7362(15.7%)
Medical condition and treatments (MT)	557(1.2%)
Suicidal attempts (SA)	139(0.3%)
Total	46588(100%)

Table 2: Sentences statistics in A1 posts

We split sentences from A1 posts to develop the sentence classification model for predicting psychological risk factors. In the human-annotated data, sentences are classified into seven psychological risk factors based on their content and emotional characteristics:

- **Suicidal thoughts and Depression (SD)**: Posts mentioning depressive symptoms, including loss of energy, lower mood, lack of confidence, inability to feel positive emotions, or agitation, wanting to injure themselves, or wishing to leave alone. Example: "I cannot hold on without my family's support now."
- **Negative cognition (NC)**: Posts expressing hopelessness and helplessness, including frustrations and lack of motivation to act or solve problems in the future. Example: "Recently, negative things have exploded one by one. I feel very pain but do not know what to do."

- **Negative Emotion (NE)**: Posts mentioning anxiety, agitation, loneliness, and other negative emotions. Example: "A little messy and resentful; be careful."
- **Suicidal Attempts (SA)**: Posts mentioning behaviors of self-harm, self-injury, or killing themselves. Example: "Overdose makes me dizzy."
- **Medical condition and Treatments (MT)**: Posts mentioning somatic complaints, physical discomfort, seeking help, psychotherapy, therapy, or medicine. Example: "I feel my heart beating fast."
- **Positive emotion (PE)**: Posts expressing confidence to solve problems, never giving up, cheering or encouraging themselves. Example: "Just wanna say it, make yourself feel better."
- **Neutral**: Sentences that do not fall into any of the above categories.

Table 2 displays the statistics of risk factors in the sentences.

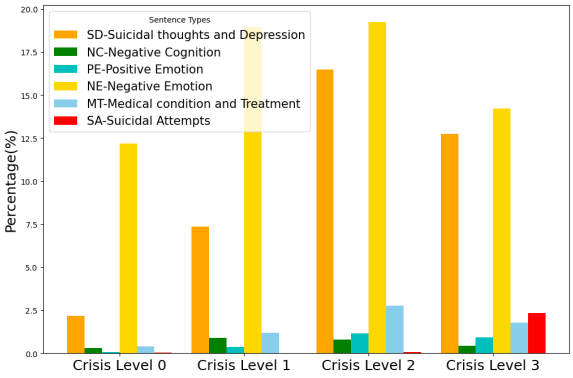

Figure 2: Distribution of psychological risk factors across crisis levels in posts. The bar chart shows the percentage of non-neutral sentence categories (SD, NC, PE, NE, MT, SA) across four crisis levels (0-3). Key observations: (1) SD and NE sentences show increasing prevalence in higher crisis levels, (2) SA sentences are significantly more frequent in crisis level 3 posts, and (3) this distribution validates the correlation between specific risk factors and suicidal crisis severity, supporting our feature engineering approach.

Figure 2 presents statistics on the distribution of risk factors across different crisis levels in posts. It reveals that sentences associated with Suicidal Thoughts and Depression (SD), as well as Negative Emotion (NE), constitute a significant proportion, particularly in posts classified under crisis levels 2 and 3. This notable increase suggests a strong correlation between these risk factors and higher crisis levels. Additionally, the proportion of Suicidal Attempt (SA) sentences is markedly higher in crisis level 3 posts compared to those in levels 0, 1, and 2. This observation underscores the importance of SA sentences as a critical risk factor in identifying high suicidal risk posts.

The primary goal of extracting risk factor features is to enhance the model's ability to identify critical sentence labels, thus enabling the effective prioritization of important sentence label types. Based on the observations in Figure 2, we identified the key risk factors for high suicidal risk posts as: SD, SA, and NE. Given that Neutral sentences constitute the majority of content in posts, their consideration is crucial to preserve the post's integrity. Consequently, we combined other risk factors into these principal categories. NC and MT were merged into NE, and PE was incorporated into Neutral sentences. After this combination, we extracted four main risk factor features: SD, SA, NE, and Neutral.

4. Method

4.1. Model Structure

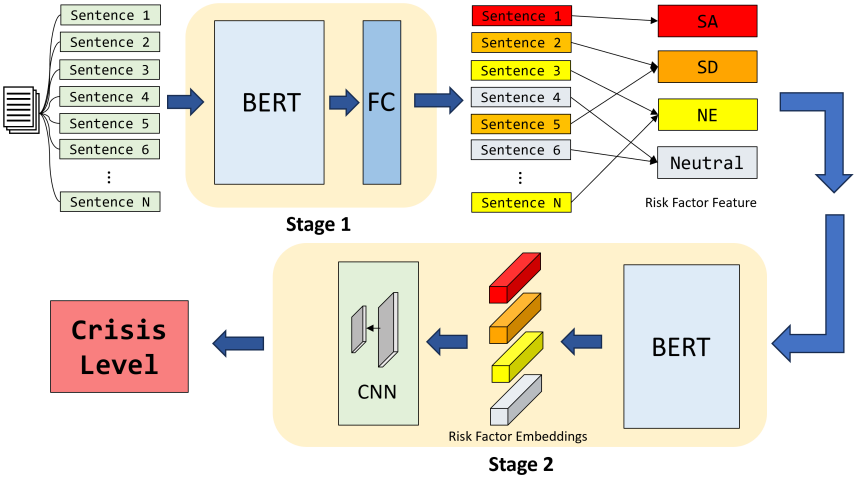

Figure 3: Two-stage architecture for suicidal post detection. Stage 1: BERT-based sentence classification model processes individual sentences to predict psychological risk factors (SD, SA, NE, Neutral). Stage 2: Sentences are grouped by risk factors into paragraphs, which are then processed by a CNN to extract spatial features and classify the overall crisis level of the post. This hierarchical approach enables both fine-grained risk factor analysis and holistic crisis assessment.

Figure 3 illustrates the structure of our research, which involves a two-stage model. The first stage (Stage 1) aims to predict risk factor labels for individual sentences. In this stage, we employ a BERT-based model to obtain embeddings for the sentences, which are then processed through a fully connected layer to generate predictions of risk factors. Once sentences are labeled by the Stage 1 model, they are concatenated into paragraphs based on their assigned risk factors.

Following the completion of Stage 1, each risk factor is associated with a corresponding paragraph. The second stage (Stage 2) of the model focuses on extracting features from these paragraphs. Subsequently, it utilizes these risk factor features to classify the crisis

level of the post. We utilize a BERT-based model to derive features from the embeddings of the corresponding paragraphs. After extracting these risk factor features, we employ a convolutional neural network (CNN) (O'shea and Nash, 2015) to process these aggregated risk-factor paragraphs. The CNN is designed to identify "patterns" in the co-occurrence and frequency of risk factors that are indicative of a specific crisis level, thereby providing a more nuanced understanding of the psychological state.

4.2. Data Augmentation

4.2.1. Sentence Augmentation

Due to the abundance of neutral sentences in the sentence dataset, this study segments a portion of these neutral sentences to create an augmentation dataset. Then, the number of sentences in less frequent categories is increased to match the size of the augmentation dataset. Randomly selecting 5 characters from the neutral sentences in the augmentation dataset, these are concatenated with the original sentences to form new ones. This method is based on the rationale that adding five neutral characters to a sentence does not affect its emotional label, whether judged by a human or AI. It's important to note that the data after augmentation should only be used for training and not for testing. Therefore, the test dataset should be kept separate and independent.

4.2.2. Post Augmentation

Since the post dataset contains many posts of type 0 (no crisis) and 1 (low crisis), which still include many 'Neutral' and 'Suicide and Depression Emotion' sentences, this study uses a portion of these type 0 and 1 posts to create an augmentation dataset. Then, 'Neutral' and 'Suicide and Depression Emotion' sentences from these posts are extracted and swapped with corresponding sentences of the same type from other posts. The rationale for this method is that swapping sentences of the same type (e.g., a 'Neutral' sentence with another 'Neutral' sentence) preserves the overall post-level crisis label.

Model Settings	**Accuracy**	**Precision**	**Recall**	**F1-Score**
hfl/chinese-bert-wwm-ext	$96.88_{0.78}$	$96.74_{0.89}$	$96.88_{0.78}$	$96.70_{0.93}$
hfl/chinese-roberta-wwm-ext	$93.26_{3.99}$	$95.54_{0.91}$	$93.26_{3.99}$	$93.92_{2.79}$
bert-base-chinese	$92.06_{3.68}$	$95.02_{1.06}$	$92.06_{3.68}$	$93.00_{2.59}$

Table 3: Performance comparison for the 2-class (3/2,1,0) post classification with augmentation across different pre-trained models.

Model Settings	**Accuracy**	**Precision**	**Recall**	**F1-Score**
7-class	$68.90_{0.69}$	$80.42_{0.28}$	$68.88_{0.69}$	$72.24_{0.63}$
4-class	$75.82_{0.38}$	$82.02_{0.34}$	$75.84_{0.38}$	$77.72_{0.33}$

Table 4: Performance of sentence classification.

Model Settings	Accuracy	Precision	Recall	F1-score
4-class w/o Aug.	$68.06_{0.83}$	$68.06_{0.83}$	$68.06_{0.83}$	$68.04_{0.83}$
4-class w/ Aug.	$75.82_{0.38}$	$82.02_{0.34}$	$75.84_{0.38}$	$77.72_{0.33}$

Table 5: Performance comparison of sentence classification with or without data augmentation (Aug.).

5. Evaluation

5.1. Setup

For both Stage 1 and Stage 2 models, we selected `hfl/chinese-bert-wwm-ext` (Cui et al., 2020, 2019) as the pre-trained model because it outperformed the other BERT-based models we tested, as shown in Table 3. This pre-trained model contains approximately 1 million parameters. The parameter settings for our models are: 8 epochs, a batch size of 32, a learning rate of 2e-5, and a sequence length of 128. In the CNN model, the CNN section includes two convolutional layer sequences: conv1 and conv2.

The hyperparameters for the conv1 layer sequence are as follows: the input channels are set to 1, output channels to 16, kernel size at 3x3, stride of 1, and padding of 1. This convolutional layer has a total of 160 parameters. The batch normalization layer features 16 channels, accounting for 32 parameters. In total, the conv1 layer sequence contains 192 parameters. For the conv2 layer sequence, the configuration includes input channels of 16, output channels of 4, kernel size of 2x2, stride of 1, and padding of 1. This convolutional layer contains 132 parameters. The batch normalization layer features 4 channels, which adds up to 8 parameters. Consequently, the conv2 layer sequence totals 140 parameters.

5.2. Sentence Classification of Risk Factors

Type of Sentence Labeling	Model Settings	Accuracy	Precision	Recall	F1-Score
Type 1: Human Labeling	4-class	$59.56_{2.63}$	$61.82_{3.09}$	$59.36_{2.52}$	$59.26_{3.55}$
	2-class(3/2,1,0)	$96.88_{0.78}$	$96.74_{0.89}$	$96.88_{0.78}$	$96.70_{0.93}$
	2-class(3,2/1,0)	$84.58_{2.16}$	$86.36_{1.08}$	$84.58_{2.16}$	$85.14_{1.82}$
Type 2: Stage-1 model Labeling	4-class	$58.10_{2.43}$	$64.44_{3.09}$	$58.10_{2.43}$	$59.56_{2.29}$
	2-class(3/2,1,0)	$93.04_{2.54}$	$94.52_{1.32}$	$93.04_{2.54}$	$93.64_{2.01}$
	2-class(3,2/1,0)	$82.58_{1.36}$	$85.62_{1.43}$	$82.58_{1.36}$	$83.52_{1.13}$
No Sentence Labeling	4-class	$59.84_{1.69}$	$63.28_{3.19}$	$59.84_{1.69}$	$60.46_{1.59}$
	2-class (3/2,1,0)	$87.16_{1.37}$	$91.60_{0.82}$	$87.16_{1.37}$	$89.04_{0.97}$
	2-class (3,2/1,0)	$84.22_{1.39}$	$84.62_{1.21}$	$84.22_{1.39}$	$84.32_{1.22}$

Table 6: Performance comparison of post classification models with sentences labeled by human psychologists, automated systems, and no sentence label.

Table 4 showcases the performance of the sentence classification model with results for both 7-class and 4-class risk factor models. The data highlights a precision that reaches

0.82 and an F1-score of approximately 0.76, which is a commendable achievement for a 4-class classification task. As shown in Table 5, the comparison of sentence classification performance with and without data augmentation strongly proves this strategy's effectiveness. The model using data augmentation clearly performed better across all key evaluation metrics than the model without it.

With the robust performance of the sentence classification model, it is appropriate to incorporate the same risk factors into paragraphs as features. Although there is room for improvement in the sentence classification model's performance, it is important to note that enhancing model performance is not our ultimate objective. We aim to further investigate whether the labeling of risk factors can improve the performance of suicidal detection.

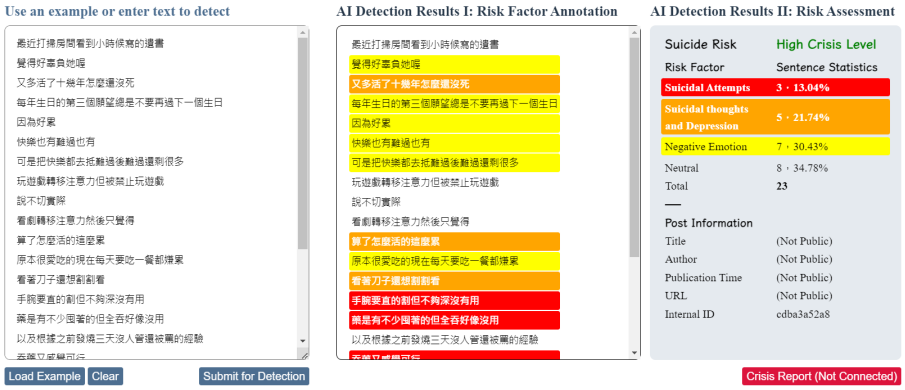

Figure 4: User interface of the suicidal post detection system. The system features a three-panel layout: (1) Input panel (left) for post submission, (2) Sentence analysis panel (middle) displaying color-coded risk factor classifications for each sentence, and (3) Post-level prediction panel (right) showing the overall crisis level assessment. This design enables psychological professionals to understand both the detailed risk factors and the holistic crisis assessment, supporting informed decision-making.

5.3. Post Classification of Suicidal Crisis Levels

Table 6 outlines the performance of post classification models trained with three different types of sentence labels. The first type utilizes models that are trained on risk factor features labeled by humans. The second type employs models trained on risk factor features labeled by the stage-1 model. The last type is used for an ablation study, which involves naive classification using entire original posts without utilizing any risk factor features. For each model, we established three classification methods. The first method categorizes according to the original four-class labeling of the posts. The second method is a binary classification that distinguishes between crisis levels 3 and 2,1,0. The third method differentiates between crisis levels 32 and 10. We also applied data augmentation for the first two types of sentence label type to observe the impact of augmentation on model performance.

The results from the 4-class model show that the best performance, reaching about 0.6 across all metrics with augmentation, is not particularly strong. This modest outcome is primarily due to the difficulty in distinguishing between crisis levels 1 and 2 in posts. We can also see that naive classification performs better than the model utilizing risk-factors. This result sounds frustrating and may make us wonder: Are risk-factors really helpful for post classification? However, since our primary objective is to detect high-risk suicidal posts, we now focus on the 2-class model with the model settings of 3/2,1,0.

With the 3/2,1,0 2-class model settings, both the F1-score and accuracy reach approximately 0.96, demonstrating the model's efficacy in identifying posts at crisis level 3. This capability not only aids in pinpointing high-risk suicidal posts but also effectively filters out a large volume of low-crisis and non-crisis posts, significantly enhancing efficiency in practical applications. Compared to the baseline with no sentence labeling, accuracy improved from 0.87 to 0.96, marking a substantial enhancement and highlighting the significance of risk factors.

To explore the impact of risk factors on suicidal detection, we examined sentence labeling type 2, which utilizes features derived from the initial model in post classification. Comparing this with type 1, we observed a decline in performance, underscoring how sentence classification enhances the effectiveness of post classification by refining information extraction. This finding is crucial as it validates the significant role psychological risk factors play in identifying high-risk posts. Additionally, the 2-class model focusing on crisis levels 3 and 2,1,0 achieved an F1-score and accuracy of 0.93, reflecting its applicability in real-world scenarios where sentences on social media are not manually labeled.

6. System Implementation and Real-World Deployment

6.1. System Architecture and Performance

We have developed an online system with a publicly accessible frontend webpage for application and testing[1]. In the demonstration system of our model, we chose a four-category risk factor classification model and a binary (3/2,1,0) suicidal crisis level model, detecting the highest crisis level posts. As shown in Figure 4, our system allows users to input posts on the left side. After pressing the "Submit for Detection" button, the sentence classification model first predicts and displays the results in the middle column, marking them with different colors to visualize the classification results. On the right side, the system displays the prediction results of the suicidal crisis level. In addition, it provides simple sentence classification data statistics and basic posting information, which are not displayed until the system is fully integrated with social media platforms and receives the necessary authorization.

It is crucial to emphasize the efficiency of our model. Unlike LLMs that require gigabytes of backend storage and several seconds to process and produce results, our model achieves remarkable efficiency, analyzing a 400-word post in Mandarin in just 0.13 seconds. This includes sentence segmentation, predicting risk factors for each sentence, and determining the suicidal crisis level of the post. Moreover, each stage of our model consumes only 0.45 gigabytes of backend memory. These efficient characteristics not only align with actual

1. Our system is publicly accessible at: https://hssai-smcrisis.phys.nthu.edu.tw/english.html

needs, assisting more students but also demonstrate our model's suitability for platforms with limited resources, offering potential for real-time suicidal detection on social platforms. In the future, we plan to enhance the system's effectiveness by potentially collaborating with external entities or application platforms, aiming to improve prediction and labeling efficiency and create a positive feedback loop with more data in the psychology and NLP domains.

6.2. Real-World Application and Impact

To demonstrate the practical impact of our system, we have deployed our model to analyze real posts from Dcard. Figure 5 shows the monthly distribution of posts across different crisis levels from January 2020 to June 2024, as automatically classified by our model. Our system demonstrates consistent performance in identifying high-risk posts, with an average of 3.2 posts per day being flagged as crisis level 3 (highest risk) across the entire dataset period. This detection rate (1.5% of posts classified as crisis level 3) is particularly ideal for practical applications. If too many posts were flagged as high-risk, it would be challenging for psychological professionals to provide meaningful attention and care to each case. The high accuracy of our model in identifying class 3 posts (0.96 accuracy) ensures that nearly all flagged posts genuinely require immediate attention and intervention, maximizing the efficiency of limited mental health resources.

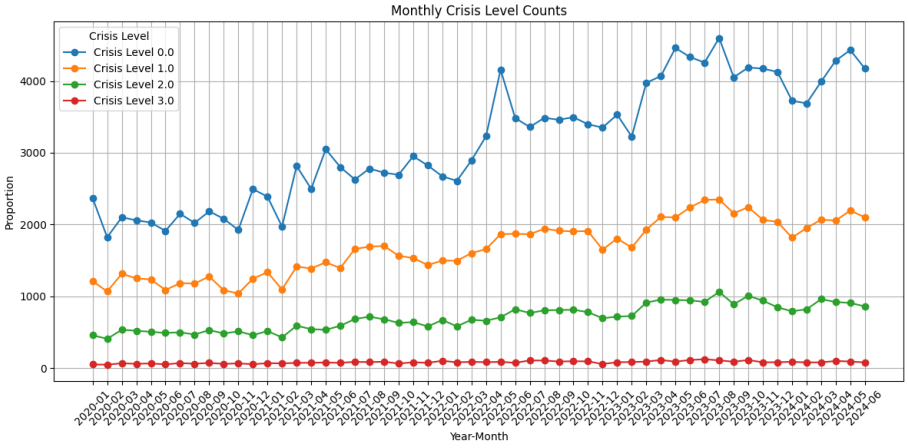

Figure 5: Monthly distribution of social media posts by crisis level (Jan 2020–Jun 2024), as classified by our model. The trend demonstrates stable and consistent identification of high-crisis posts over time, validating the model's reliability in real-world applications.

7. Conclusion

Our research integrates NLP, deep learning, and psychology, covering data labeling, feature engineering, and analysis. In collaboration with Dcard, we developed models to classify

risk factors in texts for suicide risk detection. Our 2-class model achieved 0.96 accuracy for high-crisis posts, enhancing practicality and efficiency. We successfully developed a frontend webpage system to efficiently apply our model. By labeling risk factors, we've improved model performance and provided insights from a psychological perspective, underscoring the importance of integrating psychological knowledge into suicide detection and NLP applications.

Future work will utilize transfer learning (Pan and Yang, 2009) to enhance the performance of post classification models. Additionally, label propagation (Zhu and Ghahramani, 2002) will be considered as part of the semi-supervised learning process. We also plan to deploy this system to automatically label high-crisis level posts while continuing to collaborate with psychological professionals and groups. On one hand, more human-labeled data will assist in training and improving our models. On the other hand, by leveraging this system, we aim to potentially save lives by identifying and addressing high-risk suicidal content on the internet.

Limitations

The determination of a suicide crisis is complex and should be managed by qualified mental health professionals. Our models act as a warning system and are not designed to make definitive diagnoses. They are intended to augment, not replace, the critical judgments of human experts, emphasizing the need to integrate these tools with professional evaluations for accuracy and safety in high-stakes scenarios.

Ethics

We engaged graduate students with backgrounds in psychology and counseling as part-time research assistants at National Tsing Hua University to annotate our data. These annotators received comprehensive training and participated in regular online discussions to ensure the quality of the annotation process. Furthermore, the data used were under a confidentiality agreement with Dcard, ensuring that all information was private and stripped of any personal identifiers such as names, IDs, or photos, adhering to strict ethical guidelines.

Acknowledgement

We would like to express our gratitude to Dcard for providing access to the anonymous article data from their social networking platform. We thank Prof. Yu-Kuang Hsu and Prof. Yi-Shin Chen for the valuable discussions. We also thank Ms. Yun-Yun Ho, Mr. Cheng-Shun Yeh, Ms. Wen-Yu Liang, and Ms. Shan Chung for coding original article data. This work is supported by the National Center for Theoretical Sciences, the Higher Education Sprout Project funded by the Ministry of Science and Technology, and the Ministry of Education in Taiwan.

References

Oluwole Akintoye, Nathan Wei, and Qingzhong Liu. Suicide detection in tweets using lstm and transformers. In *2024 4th Asia Conference on Information Engineering (ACIE)*, pages 22–27. IEEE, 2024.

Falwah Alhamed, Julia Ive, and Lucia Specia. Using large language models (llms) to extract evidence from pre-annotated social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 232–237, 2024.

Tayyaba Azim, Loitongbam Gyanendro Singh, and Stuart E Middleton. Detecting moments of change and suicidal risks in longitudinal user texts using multi-task learning. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 213–218, 2022.

Rebecca A Bernert, Amanda M Hilberg, Ruth Melia, Jane Paik Kim, Nigam H Shah, and Freddy Abnousi. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *International journal of environmental research and public health*, 17(16):5929, 2020.

Panchanit Boonyarat, Di Jie Liew, and Yung-Chun Chang. Leveraging enhanced bert models for detecting suicidal ideation in thai social media content amidst covid-19. *Information Processing & Management*, 61(4):103706, 2024.

Gema Castillo-Sánchez, Gonçalo Marques, Enrique Dorronzoro, Octavio Rivera-Romero, Manuel Franco-Martín, and Isabel De la Torre-Díez. Suicide risk assessment using machine learning and social networks: a scoping review. *Journal of medical systems*, 44(12): 205, 2020.

Liang-Chu Chen, Chia-Meng Lee, and Mu-Yen Chen. Exploration of social media for sentiment analysis using deep learning. *Soft Computing*, 24(11):8187–8197, 2020.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*, 2019.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.58.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Daniel Izmaylov, Avi Segal, Kobi Gal, Meytal Grimland, and Yossi Levi-Belz. Combining psychological theory with language models for suicide risk detection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2430–2438, 2023.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226, 2020.

Olivia J Kirtley, Kasper van Mens, Mark Hoogendoorn, Navneet Kapur, and Derek De Beurs. Translating promise into practice: a review of machine learning in suicide research and prevention. *The Lancet Psychiatry*, 9(3):243–252, 2022.

Lun-Wei Ku and Hsin-Hsi Chen. Mining opinions from the web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, 58(12):1838–1850, 2007.

Rodrigo Martınez-Castano, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. Early risk detection of self-harm and depression severity using bert-based transformers. *Working Notes of CLEF*, 16, 2020.

Hannah Metzler, Hubert Baginski, Thomas Niederkrotenthaler, and David Garcia. Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach. *Journal of medical internet research*, 24(8):e34705, 2022.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. Snap-batnet: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: student research workshop*, pages 147–156, 2019.

Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 167–175, 2018.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697, 2020.

Ramit Sawhney, Harshit Joshi, Rajiv Shah, and Lucie Flek. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies*, pages 2176–2190, 2021.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 941–950, 2019.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7, 2019.

Rika Tanaka and Yusuke Fukazawa. Integrating supervised extractive and generative language models for suicide risk evidence summarization. *arXiv preprint arXiv:2403.15478*, 2024.

Qamar Un Nisa and Rafi Muhammad. Towards transfer learning using bert for early detection of self-harm of social media users. *Proceedings of the Working Notes of CLEF*, pages 21–4, 2021.

Warren J Von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34(4):1607–1622, 2021.

Isabella C Wiest, Falk Gerrik Verhees, Dyke Ferber, Jiefu Zhu, Michael Bauer, Ute Lewitzka, Andrea Pfennig, Pavol Mikolas, and Jakob Nikolas Kather. Detection of suicidality through privacy-preserving large language models. *medRxiv*, pages 2024–03, 2024.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. *ProQuest number: information to all users*, 2002.