Identify High-Risk Suicidal Posts and Psychological Risk Factors on Social Media Using a Two-Stage Deep Learning Model

Anonymous ACL submission

Abstract

Our study aims to utilize psychological risk factors to detect articles on social media that are at high risk for suicidal content. We propose a two-stage model structure: the first stage labels each sentence in an article with risk factors, and the second stage uses this information as features to predict the crisis level of the article. Our models were trained using a dataset that we developed, which consists of social media posts from Dcard. These posts were labeled by psychological professionals and will be publicly released. Our approach achieved an accuracy and F1-score of 0.96 in classifying high-crisis-level articles. Our research facilitates the automatic detection of high-crisis-level articles for further analysis of risk factors, enhancing interdisciplinary collaboration between natural language processing, deep learning, and psychology.

1 Introduction

From a psychological perspective, traditional methods of determining whether someone is at risk of suicide involve analyzing cases through questionnaires or asking participants specific psychological questions, with further assessment based on their responses. However, in this era of advanced information networks, such methods are highly inefficient. Moreover, online articles, unlike questionnaires, are mostly unstructured raw data. Therefore, it is challenging to use them for suicide prevention, especially on social media platforms like Facebook. Detecting high-risk articles using keywords has been implemented on various social platforms, yet many articles with high suicide risk do not explicitly mention words like "suicide" or "death". The suicidal intent is often hidden in the semantics.

Therefore, using deep learning for sentiment analysis is particularly suitable for predicting the level of suicide risk in articles. Additionally, assessing suicidal risk is crucial for identifying both acute and chronic factors that can be treated, as well as potential protective factors that could help manage and mitigate future suicidal behaviors. However, it's important to note that such assessments do not enable predictions of actual suicide events (**?**).

Historically, self-reported questionnaires identified high-risk populations for suicide, revealing associations between depressive and anxiety symptoms, low social support, and increased suicide risk (Scardera et al., 2020). However, traditional methods fall short in accurately predicting suicide from larger social media datasets. Recent machine learning techniques have improved predictions by analyzing big data from social media, detecting suicide ideations more effectively than older methods and providing insights into psychopathological, traumatic, and familial factors affecting youth (Tadesse et al., 2019; Miche et al., 2020). Despite the potential benefits, concerns remain about social media's role in promoting suicidal behavior among adolescents (Pourmand et al., 2019).

The unique aspect of this research is the use of manually annotated sentence labels in the training data. These human-annotated sentences are utilized to develop a sentence classification model. The article classification model then uses the results of each sentence classification as input and training data to predict the final target – the label indicating the article's level of suicide risk. The availability of sentence classification labels adds interpretability to this research. More importantly, it provides a valuable resource for experts and scholars in psychology, reducing the need for costly manual annotation.

In our research, we have successfully integrated the sentence and article classification models into a web front-end. This allows users to submit articles for prediction, and displays the results of sentence and article classifications along with relevant statistics and visualizations, creating a comprehensive online crisis article detection system for psychological professions.

2 Related Work

Psychological issues are closely linked to NLP, as text is the primary medium through which people express emotions on social media. With the advent of Transformer and language models like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-4 (Achiam et al., 2023), Llama 2 (Touvron et al., 2023), and others, NLP tasks such as sentiment analysis (Tan et al., 2023) and text mining (Hickman et al., 2022) have seen significant improvements and rapid development.

Current research using 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. 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 articles 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. We collaborate with psychology professionals to label sentences and articles. By creating sentence-level classifications, we refine the performance of article classification models. Furthermore, these classifications allow psychologists to analyze content more deeply, tracing the intentions and logical reasoning behind suicidal ideation in articles. 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 original data was collected from Dcard (https://www.dcard.tw), a popular social media platform among Taiwanese college students. We used web crawlers to gather 55,989 posts from the 2019 Mood Diaries section, representing the young generation in Taiwan. Due to the large volume of data, we initially assessed the mood intensity of these posts by calculating an average mood score—total score divided by the number of words. The score for each post was derived from the frequency of certain keywords, evaluated through statistical methods and big data analysis using another dataset (NTUSD, 2018). This score reflects the positive or negative mood of the keywords and the strength of these moods. It is important to note that this mood score is not an assessment of the post's crisis level but a preliminary step to identify relevant posts for further analysis by our professionals. We selected 1,424 posts with average scores below -1.4 for human labeling, as these are likely to contain the highest percentage of high-risk, potentially suicidal messages. Our professionals also annotated the risk factors for each sentence within these posts. The rationale behind these labels will be defined and detailed below.

Our initial dataset was sourced from Dcard (https://www.dcard.tw), a social media platform favored by Taiwanese college students. We employed web crawlers to extract 55,989 posts from the 2019 Mood Diaries section to represent Taiwan's young generation. 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, 2018). These scores indicate the overall positive or negative mood conveyed by the keywords and their intensity, rather than measuring the posts' crisis levels. This step helped us to preliminarily identify posts for more detailed analysis by our professionals.

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. We also labeled the crisis level of another 1240 posts (denoted by A2), which have average scores between -1.4 and -1.2, for the test data and augmentation.

3.2 Article Description

Crisis Level	A1 article	A2 article
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: Article data statistics

The crisis level was divided into four groups from level 0 to level 3. Level 0 means people did not have any ideas about suicide and no problem at present; Level 1 means people subjectively report suicidal thoughts and some crisis events existed; however, the participants still could tolerance the bothering of suicidal thoughts; Level 2 means the participants reported suicidal thoughts and challenging to deal with the disturbance of suicidal thoughts; Level 3 means the participants reported vivid suicide thoughts and suicide attempts and they could not tolerate the suffering anymore. Among these annotated online articles, they can be divided into two types: A1 and A2 articles. Articles in A1 have undergone both article and sentence annotations, while those in A2 have only been annotated at the article level. The statistical data for the labels of A1 and A2 articles are as table 1.

3.3 Sentence Description

Sentence Label	A1 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 the human annotated sentences of the training data, they are classified into seven psychological risk factors: Suicidal thoughts and depression(SD), Negative cognition (NC), Negative emotion (NE), Suicidal attempts (SA), Medical condition and treatments (MT), Positive emotion (PE) , Neutral. All the originally annotated sentences are labeled under one of these seven categories.

Suicidal thoughts and Depression (SD): The posts mentioned any depressive symptoms, including loss of energy, lower mood, lack of confidence, inability to feel any positive emotion or agitation, wanting to injure themselves, wishing to leave alone, etc. Example: “I cannot hold on without my family’s support now.”

Negative cognition (NC) (Hopelessness and Helplessness): The contents of the posts mentioned frustrations and a 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): The posts mentioned anxiety, agitation, loneliness, and other negative emotions. Example: “A little messy and resentful; be careful.”

Suicidal Attempts (SA): The contents of the posts mentioned the behaviors of self-harm, self-injury, killing themselves, etc. Example: “Overdose makes me dizzy.”

Medical condition and Treatments (MT): The contents of the posts mentioned the experiences of somatic complaints, physical discomfort, seeking help, psychotherapy, therapy, medicine, etc. Example: “I feel my heart beating fast.

Positive emotion (PE): The contents of the posts mentioned having the confidence to solve their problem, never giving up, cheering or encouraging themselves or Thanksgiving, etc. Example: “Just wanna say it, make yourself feel better.”

Neutral: The sentences that are not categorized by the categories above.

From Tables 1 and 2, it can be observed that the training data, apart from being of a limited scale, suffer from a severe imbalance. In the article data, the number of articles decreases sharply with increasing levels of crisis; in the sentence data, neutral sentences account for over 70%, while intuitively, sentences indicating suicide behavior, which should be influential in predicting the level of suicide risk, constitute only 0.3%.

Figure 1 presents statistics on the distribution of risk factors across different crisis levels in articles. It reveals that sentences associated with Suicidal Thoughts and Depression (SD), as well as Negative Emotion (NE), constitute a significant proportion, particularly in articles 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

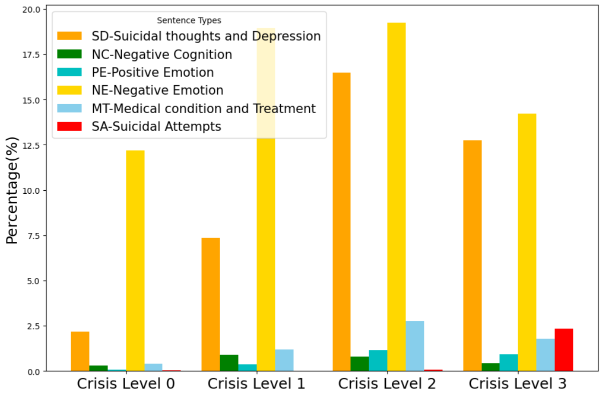

Figure 1: Distribution of Sentence Types Across Crisis Levels. The bar chart illustrates the percentage distribution of various non-neutral sentence categories across four crisis levels. Each bar represents the relative frequency of sentence categories, including suicidal ideation and depression (SD), feelings of helplessness or hopelessness (NC), positive expressions (PE), other negative expressions (NE), medical or physiological responses (MT), and suicide-related actions (SA).

level 3 articles 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 articles.

The primary goal of extracting risk factor features is to enhance the article classification model's ability to identify critical sentence labels, thus enabling the effective prioritization of important sentence label types. Based on the observations in Figure 1, we identified the key risk factors for high suicidal risk articles as: Suicidal Thoughts and Depression (SD), Suicidal Attempts (SA), and Negative Emotion (NE). Given that Neutral sentences constitute the majority of content in articles, their consideration is crucial to preserve the article's integrity. Consequently, we consolidated other risk factors into these principal categories. Negative Cognition (NC) and Medical Condition and Treatments (MT) were merged into Negative Emotion (NE), and Positive Emotion (PE) was incorporated into Neutral sentences. After this consolidation, we extracted four main risk factor features: SD, SA, NE, and Neutral.

4 Method

4.1 Structure

Figure 2 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 determine the crisis level of the post. CNN can help us effectively capture spatial hierarchies and patterns within the text, allowing for a deeper understanding of contextual relationships that are critical for accurate crisis level assessment.

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 Article Augmentation

Since the article dataset contains many articles of type 0 (No Crisis) and C (Low Crisis), which still include many 'Neutral' and 'Suicide and Depression Emotion' sentences, this study uses a portion of these 0 and C articles to create an augmentation dataset. Then, 'Neutral' and 'Suicide and Depression Emotion' sentences from these articles are extracted and swapped with corresponding sentences from other articles. The rationale for this method is that swapping 'Neutral' and 'Suicide

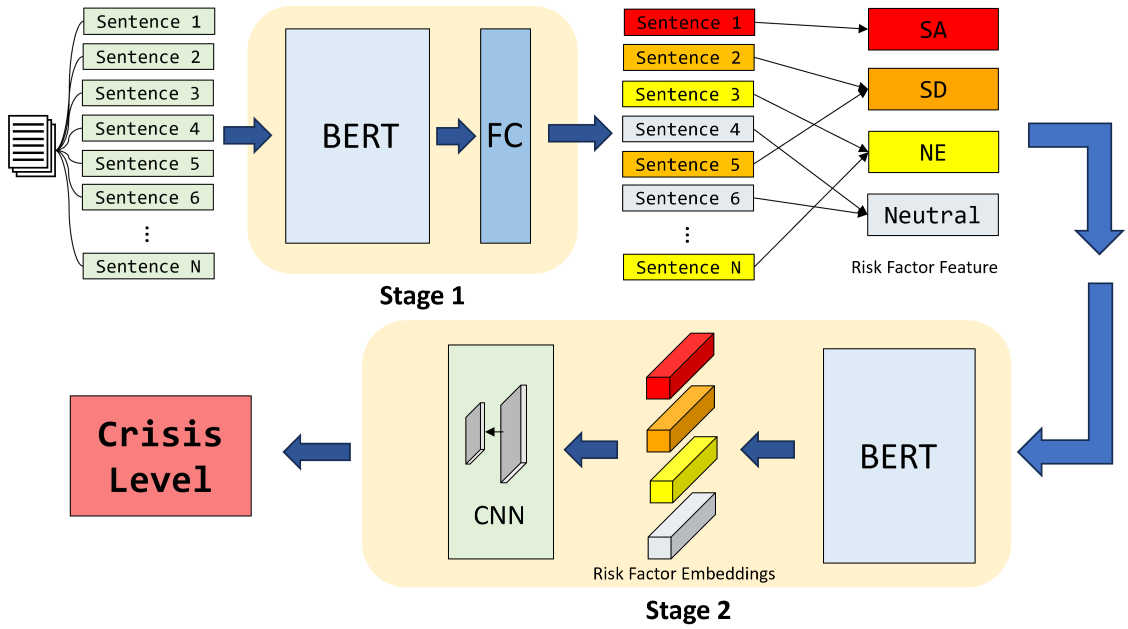

Figure 2: Structure of the suicidal detection model: Stage 1 uses a BERT-based model to generate risk factor labels for sentences, which are then grouped into paragraphs. Stage 2 extracts features from these paragraphs using a CNN to classify the crisis level of the post.

and Depression Emotion' sentences in an article shouldn't affect the overall crisis level of the article, as the labels of the sentences remain the same.

5 Experiments & Results

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 4. 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

Table 3 displays the performance of the sentence classification model, highlighting variations differentiated by the use or absence of data augmentation. We present sentence classification results for both 7-class and 4-class risk factor models. Observations from Table 3 indicate that augmentation significantly improves the performance of the sentence classification model. The data shows that the best performance is achieved with data augmentation, where the precision reaches 0.82 and the F1-score approximately 0.76—a commendable achievement for a 4-class classification task.

With the robust performance of the sentence classification model, pooling the embedding vectors of each sentence class can effectively represent the original article, which in turn enhances the performance of the subsequent article classification model. However, it's important to note that the performance of the sentence classification model is not our ultimate objective.

Model Settings	Accuracy	Precision	Recall	F1-Score
7-class w/o Aug	$56.95_{1.20}$	$78.60_{0.30}$	$56.96_{1.20}$	$63.10_{0.96}$
7-class w/ Aug	$68.90_{0.69}$	$80.42_{0.28}$	$68.88_{0.69}$	$72.24_{0.63}$
4-class w/o Aug	$68.04_{0.83}$	$80.34_{0.20}$	$68.06_{0.83}$	$71.80_{0.64}$
4-class w/ Aug	$75.82_{0.38}$	$82.02_{0.34}$	$75.84_{0.38}$	$77.72_{0.33}$

Table 3: Performance of Sentence Classification.

Type of Sentences Labeling	Model Settings	Accuracy	Precision	Recall	F1-Score
Type 1: Human Labeling	4-class w/o Aug	$58.56_{3.24}$	$60.56_{2.98}$	$58.56_{3.24}$	$57.58_{3.48}$
	4-class w/ Aug	$59.56_{2.63}$	$61.82_{3.09}$	$59.36_{2.52}$	$59.26_{3.55}$
	2-class(3/210) w/o Aug	$96.96_{0.92}$	$96.82_{1.02}$	$96.96_{0.92}$	$96.78_{1.07}$
	2-class(3/210) w/ Aug	$96.88_{0.78}$	$96.74_{0.89}$	$96.88_{0.78}$	$96.70_{0.93}$
	2-class(32/10) w/o Aug	$83.92_{2.20}$	$87.54_{1.37}$	$84.86_{1.88}$	$84.86_{1.88}$
	2-class(32/10) w/ Aug	$84.58_{2.16}$	$86.36_{1.08}$	$84.58_{2.16}$	$85.14_{1.82}$
Type 2: Stage-1 model Labeling	4-class w/o Aug	$55.34_{4.35}$	$62.48_{2.24}$	$55.34_{4.35}$	$56.70_{3.46}$
	4-class w/ Aug	$58.10_{2.43}$	$64.44_{3.09}$	$58.10_{2.43}$	$59.56_{2.29}$
	2-class(3/210) w/o Aug	$90.92_{3.77}$	$94.12_{0.94}$	$90.92_{3.77}$	$92.12_{2.67}$
	2-class(3/210) w/ Aug	$93.04_{2.54}$	$94.52_{1.32}$	$93.04_{2.54}$	$93.64_{2.01}$
	2-class(32/10) w/o Aug	$75.90_{3.97}$	$86.36_{0.93}$	$75.90_{3.97}$	$78.42_{3.37}$
	2-class(32/10) w/ Aug	$82.58_{1.36}$	$85.62_{1.43}$	$82.58_{1.36}$	$83.52_{1.13}$
No Sentence Labeling	4-class	$59.84_{1.69}$	$63.28_{3.19}$	$59.84_{1.69}$	$60.46_{1.59}$
	2-class (3/210)	$87.16_{1.37}$	$91.60_{0.82}$	$87.16_{1.37}$	$89.04_{0.97}$
	2-class (32/10)	$84.22_{1.39}$	$84.62_{1.21}$	$84.22_{1.39}$	$84.32_{1.22}$

Table 4: Performance comparison of models with sentences labeled by human psychologists, automated systems, and no sentence label. Articles are categorized into crisis levels 0, 1, 2, and 3, with level 0 indicating the least severe crisis and level 3 indicating the most severe.

5.3 Article Classification

Table 4 outlines the performance of article 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 articles 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 articles. The second method is a binary classification that distinguishes between crisis levels 3 and 210. 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 articles. We can also see that naive classification performs better than the model utilizing risk-factors. This result sounds frustrated and may make us wonder: Are risk-factors really helpful for article classification? However, since our primary objective is to detect high-risk suicidal articles, we now focus on the 2-class model with the model settings of 3/210.

With the model settings of 3/210 2-class model, both the F1-score and accuracy approximate 0.97, demonstrating the model's effectiveness in distinguishing whether an article pertains to crisis level 3. This capability not only helps in identifying high-risk suicidal articles but also efficiently filters out a large volume of low-crisis and non-crisis articles, significantly saving time in practical applications. Ultimately, this allows for the subsequent tracing of authors of high-risk articles, providing them with counseling and support as part of mental health interventions.

To explore the impact of sentence-level classifications on article-level classifications, we refer to second sentences label type, which displays the performance of an article classification model trained using risk factor features derived from stage-

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 5: Summary of mean and standard deviation performance metrics for the 2-class (A/BC0) settings with augmentation across different models.

1 model. A comparison between the first type and second type reveals a decrease in performance. This observation demonstrates that sentence classification aids the article model in extracting information, thereby enhancing the performance of article classification. This finding is pivotal to our research as it confirms the significant role of psychological risk factors in the detection and analysis of high-risk articles. Furthermore, the 2-class model with the settings of 3/210 achieves an F1-score and accuracy of 0.93, which closely aligns with real-world scenarios where sentences are not labeled by humans on social media.

6 Demonstration

In demonstration of our model, we chose a four-category sentence classification model and a binary (3/210) article classification model. As shown in Figure 2, our system allows users to input articles 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 article classification model. In addition, it provides simple sentence classification data statistics and basic posting information, with some information not disclosed due to privacy concerns.

Although the system currently operates by inputting articles, it can also integrate web crawling to form an automatic labeling system for online crisis articles for professional use, aligning with actual needs and assisting more students. In the future, we do not rule out collaborating with external entities or application platforms to enhance the system's effectiveness.

7 Conclusion

Our research has successfully integrated NLP, deep learning, and psychology across various aspects, including data labeling, feature engineering, result analysis, and demonstration. We have introduced public datasets that feature professional psychological labeling of both sentences and articles. Utilizing this dataset, we developed models for classifying sentences and articles to detect suicide risk. Our comprehensive methodology spans word, sentence, and article levels, establishing a benchmark for the dataset we proposed. Our human-labeled datasets will be released to the public when the paper is accepted. Among all models settings tested, the 2-class(3/210) model performed the best, achieving high score in every metric, which is crucial for practical applications. With risk factors labeled within the articles, the results can be interpreted and analyzed from a psychological perspective.

Future work will utilize transfer learning (Pan and Yang, 2009) to enhance the performance of article 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 articles 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.

8 Limitations

The actual determination of a suicide crisis is a complex task that should be carried out by qualified mental health professionals. Our models serve primarily as a warning system; they are not equipped to make definitive diagnoses. The reliance on algorithmic assessments without human expertise can lead to misinterpretations or oversights. Therefore, our models are intended to support, not substitute, the critical judgments made by human experts in clinical settings. This highlights the necessity of integrating our tools with professional psychological evaluation to ensure accuracy and safety in high-stakes scenarios.

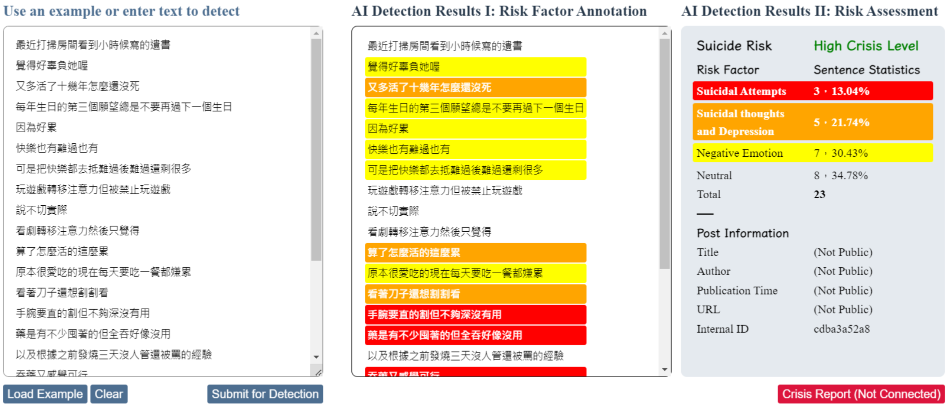

Figure 3: Screenshot of the system.

9 Ethics

We invited graduate students with backgrounds in psychology and counseling to annotate our data, compensating them as official part-time research assistants within our university.

The annotators undergo comprehensive training and education prior to the annotation task, with regular online discussions held throughout the process. As a result of this meticulous approach, consensus in annotation can be effectively achieved upon completion of the task.

The data were gathered from an openly accessible and anonymous social media platform, devoid of any personal identifiers such as names, IDs, or photos. This situation is regarded as exempt from ethical review procedures.

All data were gathered within the context of Taiwanese society, and our annotators also originate from this cultural milieu.

Throughout the preparation of this manuscript, ChatGPT was utilized for writing support, with all content thoroughly examined by the authors for accuracy and coherence.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tayyaba Azim, Loitongbam Gyanendro Singh, and Stuart E Middleton. 2022. 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.

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

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

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

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. 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. Association for Computational Linguistics.

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

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

Louis Hickman, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. 2022. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146.

Daniel Izmaylov, Avi Segal, Kobi Gal, Meytal Grimland, and Yossi Levi-Belz. 2023. Combining psychological theory with language models for suicide risk detection.

In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2430–2438.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marcel Miche, Erich Studerus, Andrea Hans Meyer, Andrew Thomas Gloster, Katja Beesdo-Baum, Hans-Ulrich Wittchen, and Roselind Lieb. 2020. Prospective prediction of suicide attempts in community adolescents and young adults, using regression methods and machine learning. *Journal of affective disorders*, 265:570–578.

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

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

Ali Pourmand, Jeffrey Roberson, Amy Caggiula, Natalia Monsalve, Murwarit Rahimi, and Vanessa Torres-Llenza. 2019. Social media and suicide: a review of technology-based epidemiology and risk assessment. *Telemedicine and e-Health*, 25(10):880–888.

Sara Scardera, Léa C Perret, Isabelle Ouellet-Morin, Geneviève Gariépy, Robert-Paul Juster, Michel Boivin, Gustavo Turecki, Richard E Tremblay, Sylvana Côté, and Marie-Claude Geoffroy. 2020. Association of social support during adolescence with depression, anxiety, and suicidal ideation in young adults. *JAMA network open*, 3(12):e2027491–e2027491.

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

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7):4550.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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