Chatbot or Lifeline? Classifying Suicide Risk in User Messages with AI

Eric Martz, Isabel Berny, Ricardo Rios, and Zhenghui Chen

Department of Computer Science, Stanford University

Group 6

Stanford
Computer Science

Problem Description

- Abuse Type: As AI chatbots are increasingly used by diverse groups for various purposes, problems have also increased. A specific problem is the misuse of AI chatbots for mental health or emotional support. While chatbots can be used to mimic human conversations, they are not equipped to handle situations like suicidal or self-harming thoughts. This disconnect between users seeking help and the help chatbots are able to provide can and has led to serious harm if left unaddressed.

- Victim Profile: Victims of these problems are often vulnerable users, individuals experiencing loneliness/struggling with their mental health, who develop emotional dependencies on the chatbots without understanding or caring about their limitations.

- Our Goal: Our moderation system aims to tackle this issue by detecting and intervening when users display signs of emotional distress and suicidal thoughts by reaching out to authorities and other appropriate individuals.

Technical Backend

- We generated a synthetic dataset of 32,838 labeled sentences across three categories: no risk, moderate risk, and high risk of self-harm or suicide. Additional no risk examples were gathered from the Kaggle "Suicide and Depression Detection" dataset. Labels were informed by the QPR suicide prevention framework, which identifies key risk indicators such as:
  - ○ Access to lethal means
  - ○ Presence of a concrete plan or intent to act
  - ○ Other contributing factors (e.g., recent trauma, hopelessness, or significant loss)
- To build our classifier, we used the all-MiniLM-L6-v2 model from SentenceTransformers to convert each sentence into a vector embedding that captures the semantic meaning. We then trained a Support Vector Classifier (SVC) with a linear kernel to distinguish between the three risk levels.
- We initially considered using logistic regression and TF-IDF vectorization, but found that the BERT-based SVC model has higher performance. For simplicity and performance, we excluded deep learning models.
- The model is deployed in our Discord backend, where it classifies messages in real-time. If a message is flagged as high risk, it prompts the moderator to follow proper protocols and contact emergency services if needed. It also places a temporary ban on the user then and sends a message offering resources.
- If a moderate message is flagged, the moderator is given the option to:
  - ○ Remove the message or Place a temporary ban on the user
- A moderate message opens a user concern which will prompt the moderator to escalate to high risk after 3 messages

Policy Language

On this platform, we care deeply about the safety and well-being of our users. Therefore, while we understand AI chatbots can be helpful tools for conversation and information, they are not an actual replacement for professional mental health support.

Therefore, users on this platform may not:

1) Use chatbots to seek assistance with self-harm, suicide, or other mental health crises
2) Interact with chatbots in ways that express intent to harm oneself or others

When we detect messages that violate the terms above, we may take the following actions:

1) Escalate the situation to platform moderators for further review
2) Temporarily suspend the user's access to the chatbot
3) Contact support resources or suicide hotline information

Our goal with these responses is not to punish vulnerable users but to prevent further harm and lead them to appropriate help. Because of this, repeat or severe violations may lead to account restrictions or permanent removal from our platform.

As mentioned above, chatbots are not equipped to handle emergencies and the earlier something is caught, the better. So if you or someone you know needs support or is in danger, please report their message and contact a trusted person, mental health provider, or emergency services.

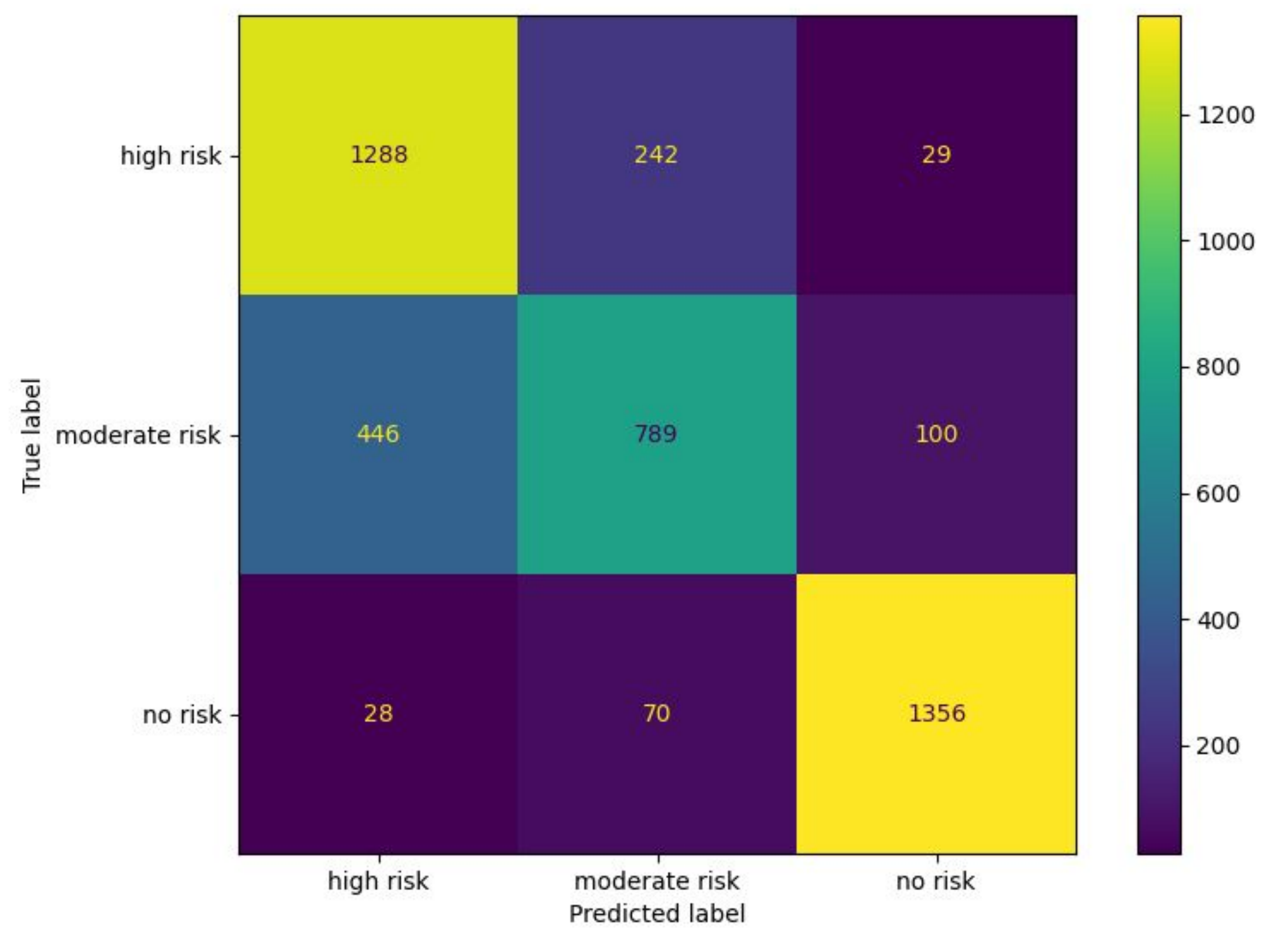
Looking Forward

Given more time and resources, we would improve our bot in the following ways:

- Provide more data to moderators, including an individual's stated means and methods of self-harm, the imminence of the threat, and user data that would be helpful in getting them immediate support.
- Give moderators more options to support the user, including connection to mental health resources and encouraging responses.
- Expand the dataset used to train the classification model to account for punctuation and spelling errors.
- Be able to classify messages in different languages other than English.
- Allow users to appeal/ provide context to moderators when false positives occur.

We believe our implementation is imperative for preventing abuse of LLMs and chatbots as therapists or crisis professionals. This technology enables moderators to intervene when an individual is in danger, with the goal that users at risk of suicide or self-harm can be connected to emergency services and mental health resources.

Evaluation



```
              precision    recall  f1-score   support

   high risk       0.73      0.83      0.78      1559
moderate risk      0.72      0.59      0.65      1335
     no risk       0.91      0.93      0.92      1454

    accuracy                           0.79      4348
   macro avg       0.79      0.78      0.78      4348
weighted avg       0.79      0.79      0.79      4348
```

- Prioritized recall because it is more dangerous to have false negatives than positives in this case—we want to make sure we identify the high-risk suicidal cases correctly.
- False Positive:
  "I would die for my dog"
  Despite not actually having suicidal undertones, the bot flagged this message as "high risk" due to its discussion of the user wanting to end their life for a particular reason.
- False Negative:
  "im gonna do it i cant stand missing her if i cant be with her whats the point"
  Though this message shows hopelessness and an intent to go through with suicidal actions, the bot classified this as no risk. The lack of punctuation as well as the generality of the terms used seem to mess up the bot. However, when proper punctuation is used, it is classified as moderate risk. A more comprehensive dataset may be able to reduce these inaccuracies due to small changes.
- Overall, our bot does fairly well in classifying the level of risk of suicide from messages, erring on the side of caution. One potential downside to this is users may be undeservedly placed on a ban or have their messages deleted.