Screening for Suicidality in Social Media Using Semi-Supervised Learning and Sentiment Analysis

Sagar Saxena
University of Maryland, College Park
ssaxena1@umd.edu

Sathvik Ravi
University of Maryland, College Park
sathrav5@terpmail.umd.edu

Sriv Srinivasan
University of Maryland, College Park
sriv1999@umd.edu

Abstract

Social media has popularized the anonymous sharing of thoughts and users can use these platforms to talk about their problems and ask for help. On the SuicideWatch subreddit, users who are struggling with suicidal thoughts can find support from peers. Although forums like SuicideWatch may not always exist or be accessible on social media platforms, user posts can provide insight into whether they are at high risk of suicidality. To help identify users who are at high risk for suicidality, we have created a semi-supervised learning approach that uses user Reddit posts to classify their risk of suicidality. Our approach uses a DistilBERT base model and sentiment features from pretrained sentiment analysis models. This approach achieves an F1 score of 0.79 on identifying users for their risk of suicidality.

1 Introduction

In an age where social media is increasingly becoming a place to share one's thoughts anonymously, the posts on these social media platforms have varying levels of seriousness. Though most posts tend to be informational, humorous, or neutral in connotation, posts regarding suicidality are becoming increasingly common. In many instances, individuals who are increasingly suicidal tend to post negatively on social media. These posts range from describing negative struggles (ex. "I lost my job, I don't know how I'm going to pay my bills") to stronger feelings of hatred in one's life (ex. "I hate my life" or "What's the point of life") to an actual suicide note.

On one specific platform, Reddit, there exists a channel or "subreddit" known as SuicideWatch which acts as a forum for individuals with suicidal thoughts to vent and seek support from professionals and peers online in an anonymous manner. Though the SuicideWatch subreddit is explicitly defined as a place for individuals with suicidal thoughts to go to, there often exist posts from suicidal individuals on other non-SuicideWatch subreddits. Therefore, we believe that a system to screen non-SuicideWatch subreddits for suicidal users would be significant in helping professionals identify such users and help them through their struggles. Specifically, we propose a semi-supervised learning approach to identifying whether a user's post can be classified as low risk or severe risk for suicide.

2 Background

Creating models to detect varying levels of suicide and working towards suicide prevention is a topic that has been explored for several years now. As described by (Resnik et al., 2020), basic NLP techniques have been used to identify users at risk based on their social media posts. Many of these basic models, however, use certain word or phrase popularity to determine whether a user is at high risk. As discussed later in this paper's exploratory data analysis, we explore why using such word popularity techniques may be misleading.

Social media company Meta's suicide prevention program (de Andrade et al., 2018) iterates on these baselines to provide a much more challenging and large-scale system. Instead of relying completely on a machine learning approach that is trained on Facebook posts that are labeled by internal human employees for severity, Meta maintains a human-in-the-loop method. Even when a user may be flagged by their machine learning system as at risk, it goes to a human reviewer to determine whether they should act on it. However, with such a system, acting on a flag can be delayed as the pipeline is bottlenecked by a human evaluation. Since speed is an important factor in suicide prevention, this could inhibit the system's real-world use.

When dealing with data associated with suicide, privacy and ethics tend to be major factors when constructing these prevention systems. Due to the

sensitive nature of these posts, it is of significant importance to ensure that individuals are not explicitly identified and that data is ethically collected. Work has been done to create better collection methodologies for suicide prevention data. Specifically, (MacAvaney et al., 2021) highlights their approach to collecting Twitter posts, involving information about suicide attempts, in a secure computational environment.

3 Exploratory Data Analysis

The data we use to train and test our system is a collection of Reddit users' posts, stemming from SuicideWatch and Non-SuicideWatch, subreddits accumulated by Dr. Philip Resnik at the University of Maryland. The posts are classified, either by experts or crowd-sourced, based on whether they are of low, severe, or moderate risk using a letter designation (either A, B, C, or D) where A refers to no to low risk, B and C refer to moderate risk, and D refers to severe risk.

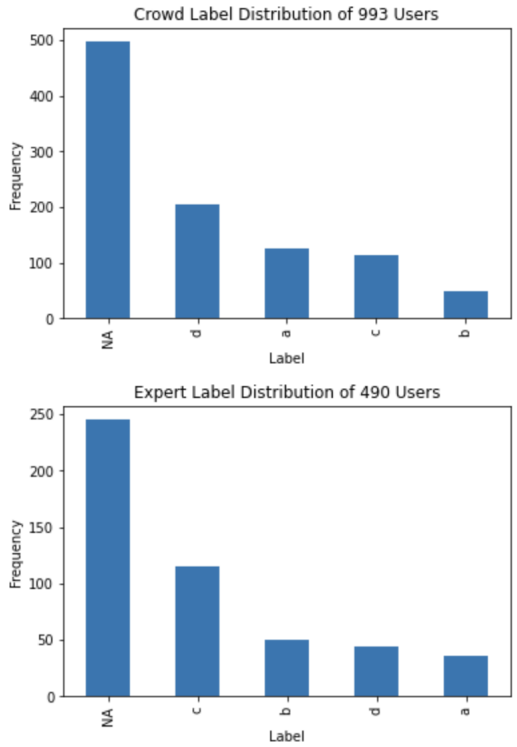

Figure 1: Distribution of suicide risk across data

From performing a quick exploratory data analysis of our dataset, we notice that of a total of 1 483 users, 993 users' posts are labeled through crowd-sourcing means while the remaining 490 users' posts are done by experts. Figure 1 highlights the distribution of the users across their labeled risk.

Furthermore, Figure 1 shows that crowd-sourced labelers tend to be more likely to use the extreme labels, C or D, while expert labelers are more likely to use the less severe labels, A or B, to classify users' suicidality. This may mean that a model trained on crowd data may classify expert-labeled data at a more extreme severity.

Since our proposed system is focused on binary classification between severe and lower suicidal risk, we restructure this data into two classes where all C and D posts are of severe risk and all A and B posts are of low risk. We can witness a difference in distribution in Figure 2, where a label of 1 refers to severe risk and a label of -1 refers to low risk:

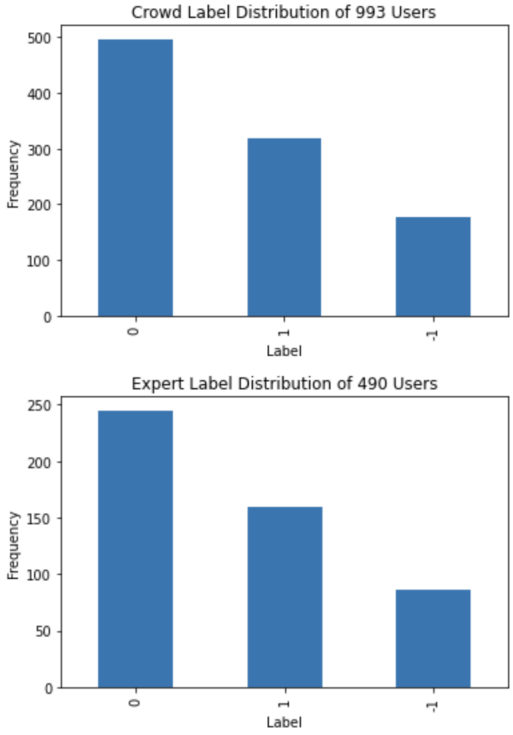

Figure 2: Distribution of suicide risk across data after grouping into low vs. severe risk

We can see that by applying this transformation, the expert label distributions and crowdsourced label distributions are more similar. As such, it may make sense to use this transformation in our future model.

Next, we explored each user's post at a more granular level. In the crowd-sourced labeled dataset, we have 2 038 753 posts and 49 417 posts in the expert labeled dataset. To determine if there is a balanced distribution across both dataset in terms of which subreddits the posts come from, we compared the top 20 most popular subreddits

from each dataset in Figure 3. We can see that the number of posts we have on each subreddit varies drastically. If we were to include this feature in our model, there may be significant bias towards subreddits with more data (or very little representative data).

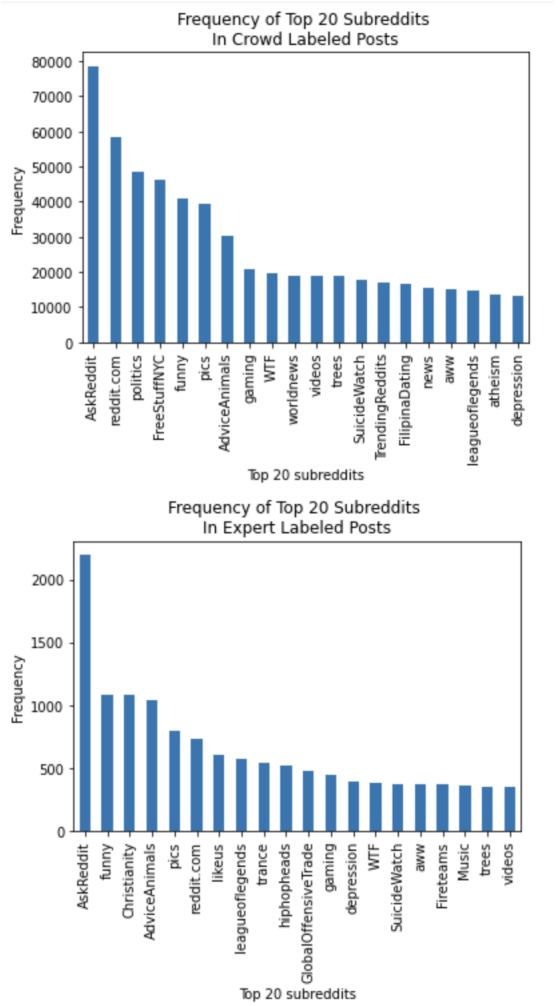

Figure 3: Top 20 most popular subreddits across both crowd-sourced and expert labeled datasets

Similarly, the number of posts that users post vary greatly as well. The top 100 most active users are shown in Figure 4. Even in the top 100 users, the number of posts can vary from less than 100 to over 1 400 in the expert labeled dataset. This indicates that some users may be overrepresented in the dataset. If our system was to be implemented by a simple model, it may be biased in classifying these users correctly.

Diving deeper into each of the posts, we notice that we have a very large vocabulary size for each dataset. Specifically, there are 1 051 222 unique words in the posts' titles and 2 336 311 unique words in the posts' body for the crowd-sourced

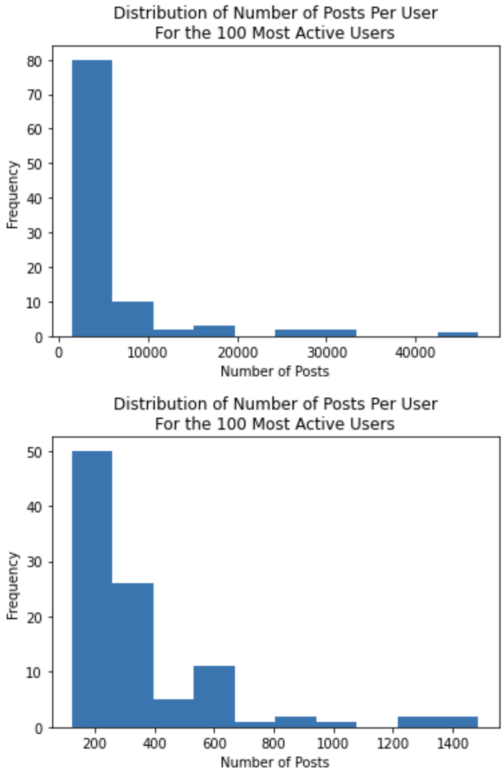

Figure 4: Distribution of the number of posts per user for top 100 most active users

labeled dataset. Similarly, there are 78 582 unique words in the titles and 174 417 unique words in the body for the expert labeled dataset.

Due to this large vocabulary size, treating each word as an individual token could be very costly and may not generalize well to unseen vocabulary. It may make more sense to convert each word into a vector representation using *word2vec*.

Lastly, before we start model development, there are two main questions that we propose in helping to solidify our model implementation direction: 1) Do users who are classified as negative for suicidality post on certain threads more than others? and 2) are certain words used more by users who are labeled positively for suicidality?

For the former question, by plotting the popularity of subreddits for both positive and negative users across both datasets (as shown in Figure 5), we can see that certain threads like *depression* and *offmychest* show up more often for users that have a positive label than those that have a negative label. It may indicate that users that post on the SuicideWatch subreddit also tend to post on these subreddits.

For the latter question, by plotting the popularity of words used in positive, negative, and neutral

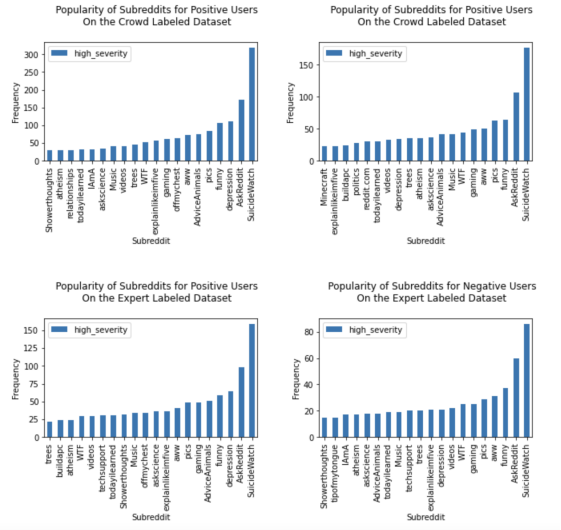

Figure 5: Popularity of subreddits for users who are positive and negative for suicidality

users, we can see (in Figure 6) that there are certain words that are more popular for both positive and negative labeled (1 and -1) users than control users (0). This may indicate that there are words that can help a model separate titles that belong to SuicideWatch from other threads. This would provide an easy, albeit incorrect, signal to a trained model that a user may be positively labeled for suicidality.

We can also see that between positive and negatively labeled users, there are differences in the most popular words. "Feel" and "love" are ranked higher for positively labeled users; "hate" and "live" also show up quite often as well.

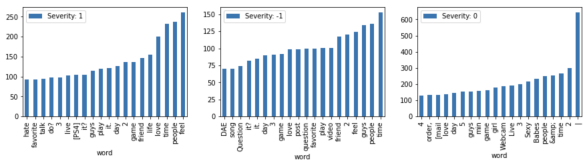

Figure 6: Popularity of words used for users who are positive or negative for suicidality

While these may be words that are used more often by a positively labeled user, a model that is trained on this dataset may be tricked by well-crafted text such as:

> *I currently **live** in College Park, Maryland! I **love** the campus but **feel** that the weather is quite unpredictable. I **hate** when it gets super cold, but **hate** it even more when it flips between hot and cold on a daily basis.*

We believe, in our unprofessional opinion, that a user who posts this message is not suicidal and should be labeled as -1. However, a potential model, especially one that places high weights on word token popularity, could possibly label this user as suicidal with a label of 1.

4 Methods

Keeping in mind the insights that we have acquired from our exploratory data analysis, we have implemented two different models. The first is a baseline model that takes a large-language model (LLM) and trains it against our dataset. We then use this baseline and improve on it.

4.1 Baseline

For our baseline implementation, we fine-tuned a pre-trained DistilBERT model on SuicideWatch and Non-SuicideWatch posts. We prepared the training data by assigning each post a weak suicidality label based on the ground-truth label assigned to that user by the experts/crowdsource workers. We trained our model for 1 epoch on crowd-sourced examples and evaluated on expert examples. The output of this model can be interpreted as a suicidality score for each post.

To obtain user-specific predictions, we represent each user with the maximum suicidality score across all posts. Then we fit a simple decision tree of depth 1 on these aggregated scores to find the threshold at which a user should be identified as suicidal. By taking the maximum across all posts, we filter out predictions on noisy posts for each user.

4.2 Training with Sentiment

To include sentiment features, we used two pretrained models: Flair (Akbik et al., 2018), an LSTM-based approach that identifies positive vs negative sentiment, and a pretrained DistilBert model for fine-grained emotion classification (anger, fear, joy, love, sadness, surprise) (Saravia et al., 2018). Neither of these models was pretrained or fine-tuned on Reddit data.

To include sentiment features in our model, we precomputed a 7-dimensional sentiment vector for each post. During training, we concatenated this representation to the DistilBERT hidden layer output and used the new post embedding as an input to the final output layer. Figure 7 shows a simple diagram with this architecture. Then we repeated the

process from the baseline to aggregate the predictions for each user and fit a decision tree to identify users who may be suicidal.

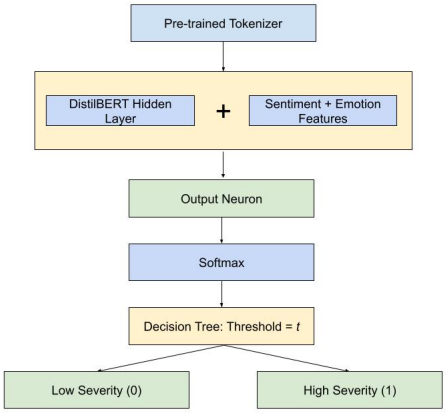

Figure 7: DistilBERT + Sentiment Model Architecture

5 Results and Discussion

Approach	F1-Score	Accuracy
DistilBERT (Baseline)	.78	.66
+ Sentiment	.79	.65

Table 1: F1-Scores and Accuracy of Baseline and Proposed Approaches

Table 1 shows the comparison of the original DistilBERT model and the improved model with input sentiment features (positive/negative, anger, fear, joy, love, sadness, surprise). Using sentiment features, we were able to get a marginal 1% improvement in F1-Score. This indicates that we were not able to significantly improve the baseline by using sentiment features in our model architecture.

We believe that there are two main reasons for this new model only improving marginally:

- the pre-trained models that we used for sentiment analysis were not trained on Reddit data which may have resulted in noisy sentiment features, and
- our approach reduces each user's history of posts to only the maximum suicidality score which may have ignored other useful signals from the predictions and data.

Given limited datasets and time, we were only able to use pre-trained models for sentiment analysis. This may have led to inaccurate and noisy sentiment features generated for each post and the model may have ignored these input features.

We also used a decision tree of depth 1 to search for a threshold at which users would be identified as at risk of suicidality. While this approach was effective in providing a single value for each user, it ignores most of the data collected for each user and is susceptible to outliers. For example, a user may have 1 post that is incorrectly classified as highly at risk of suicide (and above the threshold). Our approach would classify this user as at risk even if the user had 100 posts suggesting otherwise. This approach also ignores how the user may have evolved (through timestamps of posts). For example, a user may be indicating more and more negative sentiment as they post (which may indicate at-risk of suicide) or they may have a 1 negative sentiment post followed by 100 positive posts (which would indicate at low risk of suicide).

6 Future Work

In our work, we presented a simple approach to identify users at risk of suicide with a DistilBERT model and sentiment features generated from a pre-trained model. We believe that future researchers can take the following approaches to further this work:

- use sentiment analysis tools that are trained on Reddit data or robust to out-of-distribution data to generate improved sentiment features,
- shift sentiment features from inputs to weak labels for the model to train on to avoid overfitting on sentiment features, and
- take advantage of the linear history of a user's posts rather than treating each post as an independent data point.

One major weakness of our current approach is that the sentiment models we used are not pre-trained on Reddit data. By addressing this, researchers should be able to see a more modest improvement in the F1-Score and accuracy.

However, this may also lead to issues with overfitting (for example a model may ignore all textual features and only use the sentiment data). To address this, it is also possible that we could use sentiment data as labels rather than as inputs to the model. This would be a less noisy label than suicidality (which we only have for each user) and

would allow us to create improved deep representations of post text (as they would be sensitive to the sentiment present in the text). We leave this approach for future researchers to investigate and evaluate.

Finally, in this work, we treat each post as an independent data point. While this simplifies the problem (classifying suicidality based on a single post), it ignores the rich data we get from looking at an entire history of a user's posts. Future researchers should investigate the best way to leverage this history of posts for each user to identify users as at risk of suicide.

Acknowledgements

We would like to thank Dr. Jordan Boyd-Graber and Dr. Philip Resnik for collecting and providing the datasets that we used for training and testing our models.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and artificial intelligence: Suicide prevention on facebook. *Philosophy &amp Technology*, 31(4):669–684.

Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Association for Computational Linguistics.

Philip Resnik, April Foreman, Michelle Kuchuk, Katherine Musacchio Schafer, and Beau Pinkham. 2020. Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior*, 51(1):88–96.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.