Deep Learning for Identifying Language Features that Differentiate Mental Health Communities on Reddit

Inhwa Song, David Gomez, Ray Hung, Charles Nimo

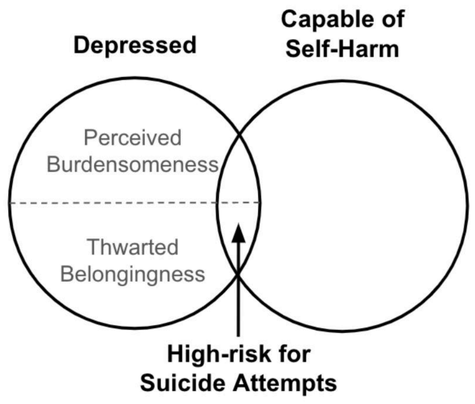
Georgia Institute of Technology

INTRODUCTION

In the domain of suicide research, it is important to distinguish between those who are ***depressed***, those who have the ***capacity for self-harm***, and those who are at ***high-risk for suicide attempts***. In this work, we use the tools of Deep Learning to understand what language features differentiate these modes of suicidality. Specifically, we use Reddit communities r/depression, r/StopSelfHarm, and r/SuicideWatch, as language proxies for those who are depressed, capable of self-harm, and suicidal, respectively. We train a spectrum of deep learning models in a single-label, multi-class context to predict to which subreddit a post belongs. Then we conduct a feature importance study to identify the language features that were most useful in making predictions, which we interpret as the language features that differentiate the stated modes of suicidality. Success will be defined in two parts: 1) we must demonstrate that we can classify posts better than a naivee or baseline model. 2) We must be able to extract which textual features from the posts were most useful during classification. Our final result will be sets of textual features that were the strongest predictors that a post belong a to particular subreddit.

DATA & METHODS

Interpersonal Theory of Suicide

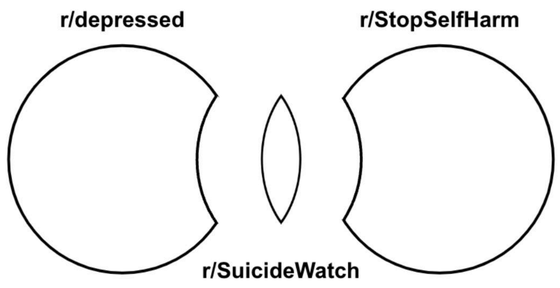


Interpersonal Theory of Suicide posits that three ingredients are required for suicide attempts:

- a sense of Thwarted Belongingness
- a sense of Perceived Burdensomeness
- Capacity for Self-Harm.

It has been observed that the presence of both TB and PB is often associated with depression, which in severe cases, may lead to suicidal desires; however, one must also possess the capacity for self-harm to be at risk for suicide attempts.

Corresponding Reddit Communities



The Reddit communities r/depression, r/SuicideWatch, and r/StopSelfHarm, serve as supportive communities, and thus constitute language proxies for those who are depressed, at-risk for suicide attempts, and capable of self-harm, respectively.

RQ1 Can we use textual features from a post to determine to which subreddit it belongs?

RQ2 If so, what features in the language distinguish the subreddits?

RESULT: PERFORMANCE & INTERPRETABILITY

**DEP: r/Depression, SW: r/SuicideWatch, SSH: r/StopSelfHarm*

MLP

Metric	DEP	SW	SSH	Macro
Precision	0.7095	0.7578	0.9242	0.7972
Recall	0.7489	0.7210	0.9197	0.7965
F1 Score	0.7287	0.7390	0.9220	0.7965

LSTM

Metric	DEP	SW	SSH	Macro
Precision	0.955	0.8554	0.9973	0.9359
Recall	0.8268	0.9650	0.9869	0.9262
F1 Score	0.8863	0.9069	0.9921	0.9284

Distil BERT

Metric	DEP	SW	SSH	Macro
Precision	0.7732	0.7478	0.9469	0.8226
Recall	0.7247	0.7992	0.9361	0.8200
F1 Score	0.7482	0.7726	0.9415	0.8208

RoBERTa

Metric	DEP	SW	SSH	Macro
Precision	0.7601	0.7791	0.9564	0.8319
Recall	0.7721	0.7755	0.9448	0.8308
F1 Score	0.7661	0.7773	0.9506	0.8312

Multi-class Classification Performance per Model

Interpretability Attribution Scores & Strong Predictors per Model

x-axis: value (SHAP value for MLP & LSTM / Captum value for DiistilBERT)

y-axis: token

CONCLUSIONS

Through the development and evaluation of several deep learning architectures, we were able to successfully identify to which subreddit a posts belongs (RQ1). Additionally, via a thorough model interpretabiltiy study, we were also able to identify the language features that differentiate our three subreddits (RQ2). Our key finding is that the three subreddits can indeed be differentiated via their language use, which reinforces the notion that these modes of suicidality are distinct.

FUTURE WORK

We aim to enhance the interpretability of our models by implementing lemmatization on the text data. Lemmatization will help capture semantic similarities by reducing words to their base or root forms, thereby treat- ing variants like ”cut” and ”cutting” as the same idea. This approach can provide a more consolidated and nuanced understanding of the linguistic patterns associated with mental health discussions on Reddit.

Georgia Tech