# 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
Atlanta, Georgia
inhwa.song@kaist.ac.kr, {dgomez32, ruizehung, nimo}@gatech.edu

## Abstract

*In the domain of suicide research and clinical practice, 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.*

## 1. Introduction

### 1.1. Motivation

Suicide constitutes a global public health burden. In United States, over 48,000 people committed suicide in 2021. Another 1.7 million attempted suicide, and roughly 12.3 million had serious suicidal thoughts. In clinical practice, 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**, as they have different treatments and rehabilitation processes. Within this context, our project's research questions (RQs) can be stated as follows:

> **Goal**: To identify the language features that differentiate those who are depressed, from those who have the capacity for self-harm, from those who are at high-risk for suicidal attempts.

To accomplish this objective, we use communities on Reddit to operationalize these modes of suicidality, and we use the tools of Deep Learning (DL) to identify distinct language features between them.

In particular, we use Reddit communities r/depression, r/StopSelfHarm, and r/SuicideWatch, as proxies for those who are depressed, have the capacity for self-harm, and are at high risk for suicide attempts, respectively. We frame the problem as a single-label, multi-class classification problem, and we train a spectrum of DL models to predict to which subreddit a post belongs. We then conduct a series of model explainability studies to identify the language features that differentiate the stated modes of suicidality.

Success will be defined in two parts: first we must demonstrate that we can classify posts better than a naivee or baseline model (in our case, a random classifier). Second, we must be able to extract which textual features from the posts were most useful during classification. Our final artifact will be three sets of textual features (e.g., tokens) that were the strongest predictors of a post belong to each subreddit.

With this background our research questions can stated as follows:

> **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?

Our project's success may offer critical insights for mental health professionals, enhancing their ability to accurately identify and address various mental health conditions, including depression, self-harm, and suicidal tendencies. This could result in more personalized and effective treatments. Additionally, our findings could aid online communities and platforms in improving monitoring and support for individuals showing signs of mental distress, potentially leading to the creation of automated tools for early detection and intervention.

## 1.2. Background

### 1.2.1 Interpersonal Theory of Suicide

In the domain of suicide research, one prevalent theory for why people die by suicide is Interpersonal Theory of Suicide (ITS) [19], which posits that three ingredients are required for suicide attempts: a sense of Thwarted Belongingness (TB), a sense of Perceived Burdensomeness (PB), and the 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. Suicide is a complex phenomena, and no psychological theory is perfect, but this theory provides us with three modes of suicidality that are important to be able to distinguish as they have different treatments and rehabilitation protocols.

### 1.2.2 Reddit

Reddit is a social media platform that functions as a collection of user-generated communities, known as subreddits, where individuals can share content and engage in discussions on various topics. In the context of mental health support, Reddit serves as a valuable space for people to connect with others facing similar challenges. Numerous mental health-related subreddits exist, providing a platform for users to share their experiences, seek advice, and offer support to one another. These communities often foster a sense of empathy and understanding, enabling individuals to discuss mental health issues openly and anonymously. Users may share coping strategies, seek guidance from peers, or simply find solace in knowing they are not alone in their struggles. The Reddit communities used in this work, r/depression, r/StopSelfHarm, and r/SuicideWatch, fall into this context of supportive communities, and thus serve as language proxies for those who are depressed, capable of self-harm, and at-risk for suicide attempts, respectively. See Figure (1).

We acknowledge two potential limitations in our approach and offer solutions. First, we recognize that individuals with thoughts of self-harm may express themselves on forums like r/depression. To address this, we selected posts based on their popularity score in each subreddit, ensuring we focus on more representative content within each community. Second, we justify using these subreddits as language proxies for different aspects of suicidality. Subreddits represent communities made up of individuals, and a post on platforms like r/depression reflects the language of someone who identifies with that community. While our dataset is precise in capturing language typical of these communities, it is not exhaustive, as it is limited to the demographics of the Reddit platform.

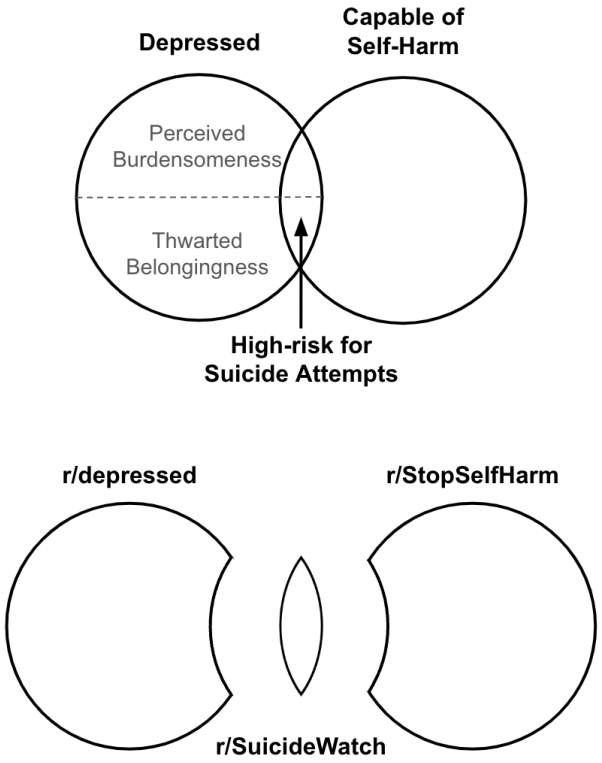


Figure 1. Interpersonal Theory of Suicide [19] and corresponding Reddit communities. Note that these diagrams are not drawn to scale. The subset of those who are capable of self-harm is much smaller than those who are depressed.

## 1.3. Related Works

Recently, researchers have used natural language processing (NLP) to study mental health discussions on social media, analyzing Reddit with traditional and deep learning methods.

In [1] researchers examined unorganized user data from Reddit, identifying and categorizing prevalent mental health issues like depression, anxiety, bipolar disorder, ADHD, and PTSD. Their analysis utilized conventional machine learning, deep learning, and transfer learning methods to effectively detect mental disorders in social media texts. Through comprehensive experiments, they showcased how these techniques can supplement clinical procedures in predicting mental health disparities between those seeking help and those unaware of their condition.

While our main goal aligns with previous studies that highlight the usefulness of deep learning and transfer learning in spotting and diagnosing mental health issues through Reddit data, our approach goes a step further. Specifically, we're concentrating on using deep learning techniques to strengthen the connection between precise subreddit classification—a crucial measure in our analysis—and

the broader aim of pinpointing individuals who might be at risk of self-harm or displaying high-risk behaviors.

In [4], researchers present SDCNL, a novel deep learning technique tailored for distinguishing suicide from depression cases. Utilizing data primarily sourced from Reddit's r/SuicideWatch and r/Depression subreddits, they apply an innovative unsupervised label correction method that doesn't depend on prior error distribution information. By evaluating various embedding models and classifiers, they demonstrate the superior efficacy of SDCNL, aiming to bolster its robustness for text-based data and labels through an unsupervised label correction strategy. This approach enables the utilization of abundant online content typically deficient in annotations.

In [5], researchers developed two specialized pre-trained masked language models, MentalBERT and MentalRoBERTa, tailored for mental healthcare. They fine-tuned these models using the final layer's special token embedding as the primary feature, utilizing data from Reddit and Twitter posts. Employing the CLPsych 2015 dataset from Johns Hopkins University, originally aimed at identifying depression and PTSD from Twitter, the authors assessed the models and different versions of pre-trained language models on various mental disorder detection benchmarks. Their findings highlighted the substantial performance improvement in mental health detection tasks with domain-specific pre-trained language representations.In our study, we prioritize connecting accurate subreddit identification with our overarching goal of identifying individuals at risk. Using explainability techniques, we delve into the inner workings of our model to enhance interpretability, which aligns seamlessly with our objective of early detection and intervention for individuals in distress. In our work, we highlight the essential connection between accurately pinpointing relevant subreddits and our main goal of identifying people who might be at risk. Taking this a step further, we further this work by building fine-tuned models adapted for the mental health domain. In contrast, to their approach of building pre-trained models.

## 2. Technical Approach

At a high level, our approach involves collecting Reddit posts from the r/depression, r/SuicideWatch, and r/StopSelfHarm subreddits and preprocessing this data. We then select subsets of these posts to train a spectrum of DL models for two distinct classification tasks: (1) binary classification where we predict whether the post came from r/SuicideWatch or not, and (2) multi-class classification where we predict to which subreddit the post belongs. We evaluate these models using the standard classification performance metrics of accuracy, precision, recall, and F1 score. Furthermore, we delve into model interpretability, aiming to uncover language features that differentiate modes of suicidality as discussed in the Introduction.

| Subreddit | Count |
|---|---|
| r/depression | 4,920 |
| r/SuicideWatch | 4,583 |
| r/StopSelfHarm | 4,178 |

Table 1. Our Reddit posts dataset after preprocessing

### 2.1. Data Collection and Preprocessing

Our dataset, provided by The Social Dynamics and Well-Being (SocWeB) Lab at Georgia Tech, consists of 1,479,218 posts from r/depression, 863,684 from r/SuicideWatch, and 9,262 from r/StopSelfHarm. We preprocessed the data to remove deleted posts, those with empty titles or bodies, and posts containing survey links (e.g., Qualtrics or Google Forms) as these are likely requests from researchers for survey participation. Our preprocessing efforts also included normalization tasks such as converting text to lowercase and excising Unicode characters. Detailed methodology will be available in our supplementary source code documentation.

Following these preprocessing steps, the dataset was reduced to 898,096 posts from r/depression, 460,835 from r/SuicideWatch, and 6,302 from r/StopSelfHarm. We selected an equal number of posts (6,302) from each subreddit based on the highest engagement scores—calculated by subtracting the number of downvotes from upvotes. Additionally, sentiment analysis was employed to exclude posts with positive sentiment. This exclusion criterion aligns with our focus on posts with negative sentiment, which are more likely to exhibit suicidal elements described in the Interpersonal Theory of Suicide [19].

Finally, we split the dataset into train, validation, and test subsets using a 70/15/15 split.

See Tab. 1 for the amount of posts in each subreddit in the final dataset we use to train our models.

### 2.2. Model Selection Rationale and Training Strategy

#### 2.2.1 Model Selection

Our approach for identifying language features that differentiate mental health communities on Reddit involved the development and evaluation of several models, including MLP, LSTM, DistillBERT, RoBERTa, and XLNet. We selected these models based on their proven effectiveness in natural language problems. As we will show, we eventually narrowed our scope down to three models: MLP, LSTM, DistillBERT.

The MLP (Multi-Layer Perceptron), with its simple architecture and bag-of-words-style features, is effective for

capturing linear relationships and can provide insights into basic patterns in the text. The Bi-LSTM (Bidirectional Long Short-Term Memory), being a type of recurrent neural network with sequential features, excels in capturing sequential dependencies and understanding context, making it adept at recognizing nuanced relationships within the text. The transformer-based models DistillBERT ([14]), RoBERTa ([9]), and XLNet ([21]) are proficient in capturing complex contextual information and semantic relationships, which allows them to discern intricate patterns and dependencies within language and makes them well-suited for nuanced analyses of mental health discussions.

The choice of employing this diversity of models stems from the recognition that each model possesses unique strengths and perspectives when interpreting textual data. These different perspectives allow for a more thorough examination of the language features that differentiate our mental health communities.

#### 2.2.2 Model Development and Training Strategy

We conducted our analysis incrementally in two phases. First, to establish data processing pipelines and general model architectures, we trained our models to solve a binary classification problem whereby we predict whether a post came from r/SuicideWatch or not. Once we were satisfied with these results, we progressed to the multi-class classification setting, whereby we retrained our models predict to which of our three subreddits a post belongs.

All models were trained using the (binary or categorical) cross-entropy loss function. This specific loss function is commonly used in classification tasks, where the goal is to predict the correct category from two or multiple classes. The Cross-Entropy Loss measures the dissimilarity between the predicted probabilities and the actual target class, providing a clear signal for the model to adjust its parameters during training. This choice aligns with the nature of our task and helps guide the model toward accurate subreddit identification.

For the MLP, we used a single hidden layer with 16 neurons and dropout regularization ($\alpha = 0.5$ and trained the model with the RMSProp optimizer with learning rate = $0.001$, and batch size of 16.

For the LSTM, we used a Bidirection layer with unit size 32 with a smaller classifier head. The model was trained with an ADAM [24] optimizer with learning rate $0.0001$ and batch size 16.

We fine-tuned the DistilBERT [13] model from Hugging Face using its Transformers library. For this model, we used learning rate = $2e^{-6}$, weight decay = 0.5, batch size = 32, and stopped early if no improvement after 3 epochs.

We also fine-tuned the Roberta [9] model from Huggingface. For this model, we configured the hyperparameters as such - weight decay = $0.5$, batch size = 8, learning rate $1e^{-6}$

### 2.3. Interpretability: Connecting Linguistic Patterns to High-Risk Behavior

In order to fortify the connection between our primary metric—accurate subreddit identification—with our overarching goal of discerning individuals who might be at risk of self-harm or showing high-risk behaviors, we used explainability techniques. We recognized that just looking at the subreddit choice might not capture all the complexities of mental health expressions. So, we aimed to uncover the most impactful language features in the posts. By using explainability techniques, particularly focusing on feature importance, we dived into the detailed language patterns that play a big role in what our models predict. Understanding these subtleties is crucial for dealing with concerns about potential factors that might affect our results, like, for example, the varied content in the "R/StopSelfHarm" subreddit.

When we pinpointed words with high feature importance, we directly linked language elements to our goal of spotting people at risk of self-harm. For example, our analysis might show that certain words or phrases are more important in predicting high-risk content. This not only makes our models easier to understand but also gives a solid basis for explaining why we use subreddit identification as a proxy measure.

Explainability is a useful tool for refining our focus by helping us zero in on the language elements that closely match our study's goals. This not only helps deal with potential issues but also gives us a way to keep improving and adjusting our methods based on the important features we identify.

Using explainability to find words with high feature importance makes the link between our chosen measure and the wider research goal stronger. It gives us a detailed understanding of language patterns, checks our assumptions, and fine-tunes our focus on finding individuals at risk of self-harm in the complex world of mental health communities on Reddit.

We have conducted an analysis of interpretability using Feature Attribution-based explanation techniques.Feature attribution aims to measure the impact of a particular feature on the model's prediction. This process is crucial for understanding how much influence each feature has in shaping the overall prediction made by the model. Several techniques for computing and visualizing feature attribution have been explored in literature. Some of the simple feature attribution methods include occlusion [8] [22]. This works by estimating the importance of a group of features by setting it to zero and measuring the resulting decrease in prediction accuracy. While some other popular feature attribution techniques include gradient based techniques

[2, 3, 11, 15, 20, 23], integrated gradients [18], saliency maps [17], and deep lift [16]. The Captum library, a robust tool for enhancing the interpretability of PyTorch models, implements a variety of attribution algorithms [6]. In our attribution analysis, we leveraged this open-source library, developed by Meta, to delve into the inner workings of our models [6]. Captum provides an array of tools designed to make models more interpretable, and one such technique we embraced is Integrated Gradients. This approach allowed us to gain nuanced insights into the contribution of different features to our model predictions, enhancing the overall transparency and comprehensibility of our findings. In one of our experiments, we utilized the SHAP technique [10] to enhance the interpretability of text-based models by visualizing influential words in posts. SHAP provides a framework that allows models to quantify the contribution of each feature towards the predicted outcome, thereby explaining the model's output in a comprehensive manner. In a subsequent experiment, we extended SHAP to transformer models, deepening our understanding of linguistic patterns.

Integrated Gradients [18], in particular, has proven effective in interpreting the behavior of deep learning models, especially for complex models like those based on the Transformer architecture. This method attributes the prediction of a deep network to its input features, providing a more nuanced understanding of the model's behavior.

To enhance the clarity and relevance of our attribution analysis, we have implemented preprocessing steps that involve the removal of stop words, punctuation, and non-English words. This ensures that our analysis focuses on the most impactful and meaningful words in the dataset, thereby providing a clearer picture of the linguistic characteristics that differentiate posts from various subreddits.

### 2.4. Evaluation Metrics

The performance of the trained classifiers was evaluated using multiple metrics including F1 score, precision, recall, and accuracy. All of these metrics provide a meaningful perspective into the model's classification capabilities, and its ability to make accurate predictions across multiple classes. Additionally, the evaluation process incorporated advanced techniques for feature attribution explainability, leveraging SHAP values and the Integrated Gradients methodology. This holistic evaluation framework not only underscores the classifier's overall performance but also demonstrates the interpretability of the model through the lens of feature importance and gradient-based explanations.

## 3. Experiments and Results

These phases and results are discussed next.

### 3.1. Binary Classification Task

In this section, we show results derived from an evaluation of different model performances dedicated to distinguishing between suicidal and non-suicidal Reddit posts on a binary classification task. The findings, shown in Table 2, highlight the superior performance of the Long Short-Term Memory (LSTM) model in accomplishing this task. A deeper analysis suggests that the stronger performance of the LSTM model might stem from its inherent capacity to excel in scenarios characterized by limited data availability and comparatively shorter sequences.

Building on these promising outcomes, our upcoming experiments will involve the application of advanced explainability techniques to uncover the dynamics behind the LSTM's remarkable performance, offering insights into the specific factors contributing to its edge over the Multi-Layer Perceptron (MLP) and transformer-based models in our experimental setup.

| Metric | MLP | LSTM | DistilBERT | RoBERTa |
|---|---|---|---|---|
| Accuracy | 0.804 | 0.631 | 0.831 | 0.782 |
| Precision | 0.788 | 0.602 | 0.736 | 0.703 |
| Recall | 0.783 | 0.470 | 0.830 | 0.684 |
| F1 | 0.786 | 0.503 | 0.780 | 0.760 |

Table 2. Comparative Model Performance Metrics for Suicidal Classification. Results reported are macro-scores.

### 3.2. Multi-Class Classification Task

#### 3.2.1 Model Performance

We now present our results for the multi-class setting, where we try to predict to which subreddit (r/depression, r/SuicideWatch, and r/StopSelfHarm) a post belongs. (Note that the performance of the Bi-LSTM model for the multi-class setting were suspiciously high. An error was indeed found, but the underlying issue could not be resolved given our time constraints. Please ignore LSTM performance and interpretation.)

We selected a diverse set of models, including MLP, Bi-LSTM, Distill-BERT, and RoBERTa, for text classification due to their unique strengths in interpreting textual data. MLP provides simplicity and efficiency in capturing linear patterns, Bi-LSTM excels in grasping sequential dependencies, Distill-BERT balances performance and efficiency, and RoBERTa is optimized for deep contextual understanding. This ensemble approach recognizes that different models may excel in recognizing distinct linguistic nuances, enhancing overall performance and interpretability. The inclusion of these varied models ensures a comprehensive examination of the data, allowing the final model to benefit from the strengths of each constituent model.

Shown in Tables 3-6 are the performance results on the multi-class classification problem for each model. Note we abbreviate r/depression, r/SuicideWatch, and r/StopSelfHarm as DEP, SW, and SSH, respectively. Also note that a baseline (random) classifier, would achieve scores of roughly 0.33 for all metrics.

Two salient findings from the presented tables are noteworthy. First, all models exhibit robust performance surpassing the baseline, with macro F1-scores of 0.80, 0.93, 0.82, and 0.83 for MLP, Bi-LSTM, DistilBERT, and RoBERTa, respectively. In stark contrast, the random classifier yields a score of approximately 0.33, underscoring the success of our models in effectively discerning the source of Reddit posts in the dataset. Second, a consistent observation emerges, wherein class-specific metrics for r/StopSelfHarm (SSH) consistently surpass those for r/depression (DEP) and r/SuicideWatch (SW) by 15-20 percentage points. This discrepancy suggests that the language employed in r/StopSelfHarm is more distinctive, rendering posts from this subreddit more readily classifiable than those from the other two subreddits.

| 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 |

Table 3. MLP Performance for Multi-Class Classification

| 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 |

Table 4. LSTM Performance for Multi-Class Classification. Note: These values were suspiciously high, but the underlying issue could not be resolved given our time constraints.

| 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 |

Table 5. DistilBERT Performance for Multi-Class Classification

| 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 |

Table 6. RoBERTa Performance for Multi-Class Classification

#### 3.2.2 Interpretability

As discussed previously, we leverage SHAPley values [10] and Captum [6], both powerful tools for explaining model behavior, to gain a deeper comprehension of the features used to make our model's predictions.

Importantly, we should acknowledge a limitation we experienced with SHAP(ley): computation time. The algorithm for computing SHAP values took much longer than compared to Captum (and would often crash our Colab kernel). To get around this, for SHAP (e.g., for the MLP and LSTM models) we had to restrict our analysis to 10 samples from the test set. This means that the features for MLP and LSTM are obtained from a smaller sample set and thus may exhibit higer variability. In contrast, the features obtained with Captum were extracted from the *entire* test set, and thus should be relied on more heavily than the other two models. Because of this, we will focus on the features obtained via DistilBERT as those were extracted via Captum.

Shown Figure 2 are the attribution scores for the top 10 features by model and by subreddit that were most useful in making predictions.

Consider the r/depression class and focus on the DistilBERT model. At the very top we have "depression", which is far and away the most indicative feature that a post is from the r/depression, as compared to the others. We also see words related to feelings or moods: "feel/feeling/happy/depressed". Interestingly, we see the filler word "like" as strong predictor.

Next, consider r/SuicideWatch, and focus on the DistilBERT model. We again see that one feature stands out over the others. In this case, "suicide" is the most indicative feature of a post belonging to r/SuicideWatch. Other important features are "suicidal", "kill", "killing", "die", and interestingly both "life" and "death".

Finally, consider r/StopSelfHarm and focus on DistilBERT. We see six of the top ten features are related to cutting (e.g., "cut", "cutting", "cuts", "scars", "blade", and "knife"). We also see the proverbial *title drop* here: "self, harm" as well as the word "clean" which in this context refers to abstaining from committing self-harm. These results align with the prior work [7] which finds the large majority of self-harm behavior to be of the cutting-type. It also explains why the class-specific performance was in general better for self-harm class: a majority of posts in

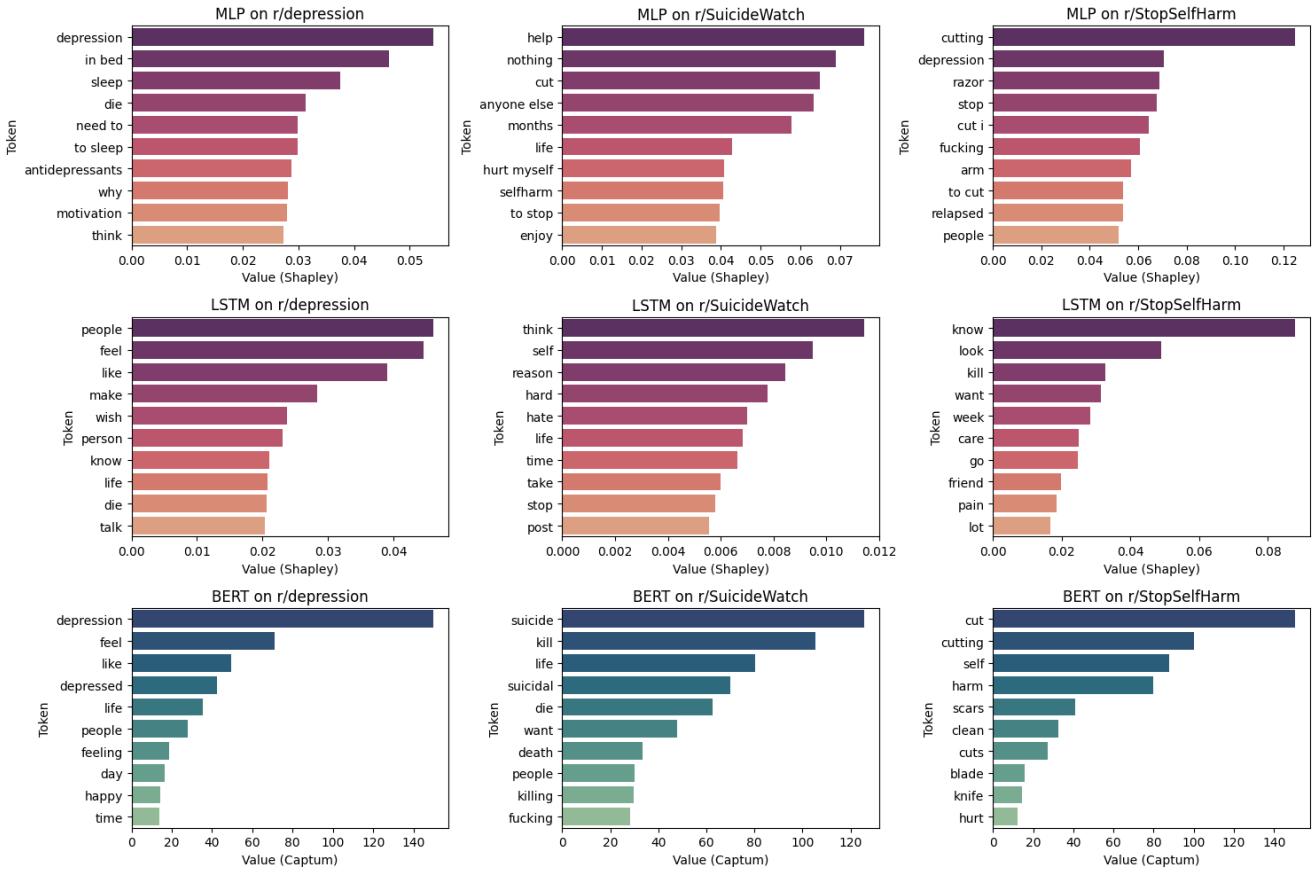


Figure 2. Model Interpretability via Attribution Scores for each model (rows) and subreddit (columns). The red-ish hues are attribution values computed via SHAPley and the blue-ish with Captum. Each subpanel represents the top 10 features most useful for distinguishing between r/depression, r/StopSelfHarm, and r/SuicideWatch.

r/StopSelfHarm discuss cutting, which is a topic not likely to be in the other subreddits.

It is important to note that the features identified these subreddits are not necessarily the most use words by their respective communities, but rather, *these are the features that are most useful to differentiate them*. For example, it is well established [12] that depressed persons use more first-personal-singular ("i") and negative valence language, but these features are not shown in DistilBERT's (more reliable) r/depression features because "i" words are also be used in the other subreddits and would not be useful for differentiating them.

## 4. Conclusion

We set out to identify language features that differentiate mental health communities on Reddit (r/depression, r/SuicideWatch, and r/StopSelfHarm) using deep learning. Through the development of several deep learning architectures, we were able to successfully identify to which subreddit a posts belongs (RQ1). Additionally, through a thorough model interpretabiltiy study we were also able to identify the language feature that differentiate our three subreddits.

For the multi-class task, diverse models (MLP, Bi-LSTM, Distill-BERT) demonstrated robust performance, consistently outperforming a baseline random classifier. Notably, r/StopSelfHarm excelled compared to r/depression and r/SuicideWatch, each revealing distinct linguistic patterns. SHAPley values and Captum analysis highlighted emotional expressions and the term "depression" in r/depression, explicit language related to suicide in r/SuicideWatch, and cutting-related terms with nuanced language markers in r/StopSelfHarm, emphasizing community-specific language.

In future work, we aim to improve interpretability by implementing lemmatization on text data, capturing semantic similarities by reducing words to their base forms. Exploring domain-specific embeddings or pre-trained language models and expanding the dataset to include diverse sources could enhance model performance and interpretation, contributing to cross-domain applications.

## References


[1] Iqra Ameer, Muhammad Arif, Grigori Sidorov, Helena Gòmez-Adorno, and Alexander Gelbukh. Mental illness classification on social media texts using deep learning and transfer learning, 2022. 2

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks, 2018. 5

[3] Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management*. ACM, 2021. 5

[4] Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction, 2021. 3

[5] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare, 2021. 3

[6] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. 5, 6

[7] Aviva Laye-Gindhu and Kimberly A Schonert-Reichl. Nonsuicidal self-harm among community adolescents: Understanding the "whats" and "whys" of self-harm. *Journal of youth and Adolescence*, 34:447–457, 2005. 6

[8] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure, 2017. 4

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. 4

[10] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. 5, 6

[11] Ian E. Nielsen, Dimah Dera, Ghulam Rasool, Ravi P. Ramachandran, and Nidhal Carla Bouaynaya. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4):73–84, 2022. 5

[12] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18 (8):1121–1133, 2004. 7

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 4

[14] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. 4

[15] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2): 336–359, 2019. 5

[16] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019. 5

[17] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. 5

[18] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. 5

[19] Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. The interpersonal theory of suicide. *Psychological review*, 117(2):575, 2010. 2, 3

[20] Hanjing Wang, Dhiraj Joshi, Shiqiang Wang, and Qiang Ji. Gradient-based uncertainty attribution for explainable bayesian deep learning, 2023. 5

[21] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020. 4

[22] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013. 4

[23] Jindi Zhang, Luning Wang, Dan Su, Yongxiang Huang, Caleb Chen Cao, and Lei Chen. Model debiasing via gradient-based explanation on representation, 2023. 5

[24] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018. 4