Using Natural Language Processing to predict suicidal ideation on Reddit

Author: Eugene Baraka

Background

According to the WHO (2014), suicide is one of the leading causes of death globally, with an estimated 800,000 deaths annually. That is one death every 40 seconds. A key aspect in suicide prevention is addressing suicidal thoughts and ideas before they turn into actions. With an ever-increasing number of internet users worldwide, social media platforms have become a space where people share about their daily lives, and this includes emotions and hardships. A recent US study has shown a strong positive correlation between the proportion of the tweets containing suicide risk factors with rates of age-adjusted suicide rates in the US (Jashinsky et al., 2014). This means that suicidal thoughts are posted on social media platforms and may present an opportunity to use machine learning to predict which posts are likely to be suicidal. Knowing which posts display suicidal ideation would better assist in early detection of self harm and improve public health digital surveillance (infosurveillance) to create policies that aim to reduce suicide. Multiple studies (e.g. Morese et al., 2022, Ueda et al., 2014, and many more) have explored the use machine learning for early detection of suicidal thoughts and the progress has been impressive over the past decade

In this analysis, the start-of-the-art machine learning techniques to classify suicidal and non-suicidal thoughts and compare to corroborate the research that's already been done in the field. We compare various models and decide which one performs the best taking into account the use computational resources, interpretability, and overall complexity.

About the data

This analysis relies on the data collected from Reddit platform, which leverages group chats known as sub-reddits to discuss various topics. Each subreddit is allocated to a specific topic and people around the world interested in that topic can join the group to enhance discussions. The data used in this analysis was collected from "SuicideWatch" (contains 390k users as of October 2022) and "Teenagers" (contains 2.9 million users as of October 2022) subreddits using PushShift API. Since subreddits are topic-centric, it was assumed that posts scraped from SuicideWatch subreddits are suicidal or include some form of self-harm, depression, or anxiety of some way, while posts in Teenagers subreddits were not suicidal and contain just general daily life conversations. The data from both subreddits is a representative sample of the posts that start from December 16, 2008 to January 1, 2021. Please find the public dataset here

Data description and cleaning

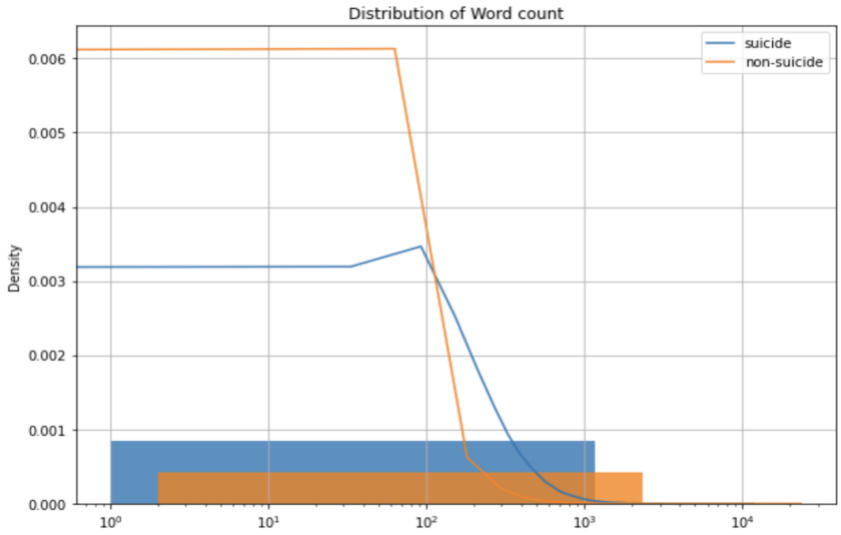
The dataset consists of 232,074 unique posts and each post is classified as either suicidal or non-suicidal. There are no missing values in the dataset and the percentage of each class is 50%.

The first step in this analysis was to clean and preprocess the data before any exploratory data analysis and modelling. Each post was cleaned by removing any sort of punctuation, URLs, stopwords (e.g. I, to, ...), expanding contractions (e.g., "I'm" -> "I am") and changing all text to lowercase. The cleaned posts were saved in a new column in the dataset. Some feature engineering was also performed on the original text to extract the number of words, characters, and sentences as well as the average word and sentence lengths for later comparisons between the two classes.

Exploratory data analysis

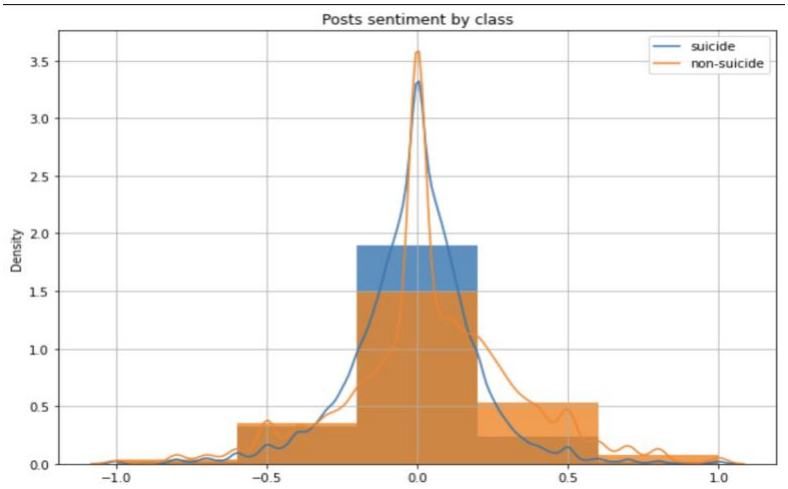
The next step was to analyze and compare posts with respect to their class. The distribution of the word count tended to be higher in the non-suicidal posts but the large areas of both distributions overlap heavily, so the difference is not very significant (*Figure1*). Looking at other characteristics including the number characters, the average word and sentence length, there seem to be difference between suicidal and non-suicidal posts.

Figure 1. Distribution of word count by post class



Sentiment analysis was also determined using the *vader* algorithm to compare the sentiment between the posts attributed to both classes. The distribution of the sentiment for both suicidal and non-suicidal posts are normally distributed, centered at 0 (i.e. the majority of the posts in both categories tend to have a neutral sentiment) and they both overlap covering (*Figure 2*). Generally, there seem to be no difference between the sentiments between the two categories.

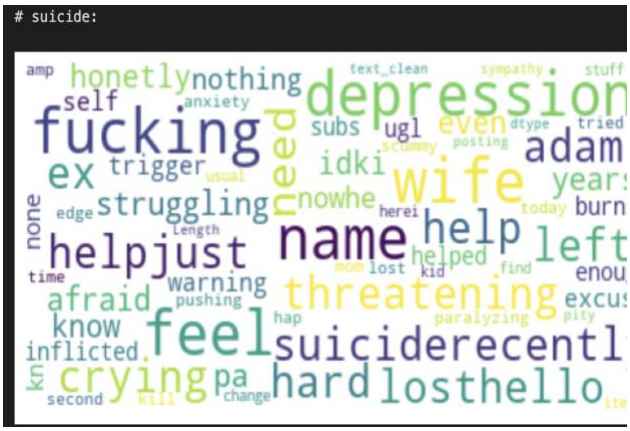
Figure 2. Sentiment analysis by post class



The most common words (from 1 n-gram to 3-ngrams) were also extracted to compare both groups. Overall, it was found that the suicidal posts category contained higher proportion of feeling words (e.g. want, like, feel) and more negative words, some related to suicide (suicidal thoughts, want kill, want die, feel like shit, want live anymore, ...).

The last step in the EDA stage was to summarize the content of each category using WordClouds. As expected, and seen in the graph above, the suicidal group contains more negative and suicide-related words than the non-suicidal posts category (*Figure 3 below*).

Figure 3. The most common words by post category (left: suicidal, right: non-suicidal)



Preprocessing before modeling

The class column was numerically encoded with suicidal posts mapped to 1 and 0 if otherwise. The dataset was split into train and test sets in 7:3 ratio and both unlabelled sets were stemmed using NLTK's PorterStemmer and vectorized using TFIDF vectorizer.

Text Classification (MODEL EXPLORATION PHASE)

In total, 6 classifiers (logistic regression, KNN, Decision tree, XGBoost, Random Forest, and SVM) were explored individually to choose which ones fit the data the best. The classification report that produced the precision, recall, and F1 score for each model was produced and the models with the highest F1 scores passed this phase to be explored further through parameter optimization. The table below shows how the model compared to each other.

Table 1: Comparison between the accuracy of six classifiers

Model	class	Precision	Recall	F1 score
Logistic regression	0	0.86	0.88	**0.87**
	1	0.88	0.85	
KNN	0	0.88	0.75	0.82
	1	0.78	0.89	
Decision Trees	0	0.80	0.82	0.81
	1	0.82	0.79	
XGBoost	0	0.85	0.87	**0.86**
	1	0.87	0.85	
Random forest	0	0.86	0.88	**0.87**
	1	0.88	0.85	
SVM	0	0.86	0.88	**0.87**
	1	0.88	0.85	

The four models with the highest F-1 score were optimized to choose the model with the highest accuracy score and after optimization, logistic regression ended up being the best model with 90% test accuracy score.

Transfer Learning using DistillBERT

DistilBERT is a small, fast, cheap and light Transformer model that approximates the BERT architecture, the largest pretrained language architecture to-date. I chose to use DistillBERT to train my neural network since it is fast and small, so it would be computationally cheaper to run and give almost the same results as the ones I would've obtained using BERT. Since the model has its own text cleaning and preprocessing modules, the process was straightforward. The model was trained on 5 total epochs with two input layers (note that a the number of input layers has been reduced to accomodate

for a very small computation power). The F1 score for the DistillBERT model was 92% as shown in the figure below:

Figure 4: DistillBERT model accuracy score

```
F1 score 0.9232954545454547
Classification Report
              precision    recall  f1-score   support

non-suicidal       0.97      0.88      0.92       361
    suicidal       0.88      0.97      0.92       336

    accuracy                           0.92       697
   macro avg       0.92      0.92      0.92       697
weighted avg       0.93      0.92      0.92       697
```

Discussion and conclusion

The overall goal of this project was to leverage natural language processing to predict suicidal ideation on Reddit. Since this was a classification problem, a total of 6 classifiers were initially chosen and four models were selected to be optimized to select one final model. The model with the highest accuracy (F1 score) was logistic regression. Transfer learning using DistillBERT was also performed and the accuracy improved by 2% compared to that of logistic regression. Given the computational power required to train neural networks and their lack of interpretability, logistic regression maybe preferred in this case. This is because it is easy to know which words predict suicidality in reddit posts better in a logistic regression than in a NN. The findings in this analysis agree with the literature that machine learning has the power to predict suicidality and prevent suicide rates across the globe, if detected early.

Although the accuracy of the models trained in this task and those presented in the literature are incredibly high, the applicability of suicide prediction using machine learning still faces multiple challenges. First of all, as is in this task, the data quality is still very low. This is associated with high data labelling costs. For example in this analysis, all posts from the "SuicideWatch" subreddit were labelled as suicide, which is very likely innacurate; hence, in this case, the predictive accuracy obtained might be for the two subreddits used in the dataset instead of the actual suicidal versus non-suicidal posts. If this is true, the generalizability of the model would be very poor. Another limitation is that there is no formal definition of suicidal thoughts because it can span a very big spectrum (i.e., anxiety → mild depression → strong depression → suicidal thoughts). Given these limitations, it is

important to work on maximizing the data quality by involving human psychologists and psychotherapists.

To conclude, machine learning is a very powerful tool that can be used to predict suicidal thoughts and ideation and although in its infancy, with high quality data, it can help reduce suicide risk by early detection on social media platforms.