Can the suicidal intentions of young people be predicted using the language analysis of Reddit posts?

Data Description:

This report describes the initial analysis and visualization of the "Reddit dataset for Multi-task NLP" on Kaggle. This dataset is a collection of 226704 posts from the "SuicideWatch" subreddits of the Reddit platform. The posts are collected using the Pushshift API. All posts that were made to "SuicideWatch" from December 16, 2008, until January 2, 2021, are labelled as "suicidal". Posts from r/teenagers (normal conversations) are labelled as "non-suicidal". The "SuicideWatch" subreddit is a platform where a significant portion of the posts on this will involve discussions and expressions related to suicidal ideation.

Figure 1 displays the dataset with the target variable in the "Suicidal_label" column, labelled as suicidal (1) and non-suicidal (0). One feature is "Sentiment labelling,", a type of emotion analysis classified as "negative", "neutral" and "positive". From the context extracted from user posts ("Post" column), this report will extract the "Length" feature and the "Deciding Words" feature to see how much the length of each post and the frequency of some important words affect the target variable and whether the suicidal intention is always related to negative feelings, to predict whether a post from social media has a suicidal intention or not.

	Post	Suicidal_label	Sentiment_label
0	Ex Wife Threatening SuicideRecently I left my ...	1	0
1	Am I weird I don t get affected by compliments...	0	1
2	Finally is almost over So I can never hear ...	0	0
3	i need helpjust help me im crying so hard	1	0
4	I m so lostHello my name is Adam and I ve b...	1	0
...	...	...	...
226948	I sound like a dudebro but I can t handle my f...	1	0
226949	Fuck my sister She is such I fucking bitch and...	0	0
226950	I ve been suicidal for years and no one knowsT...	1	1
226951	My boyfriend is sick so I took some Polaroids ...	0	0
226952	What would happen to my dog M What would happ...	1	0

226953 rows × 3 columns

Figure 1: A data frame with 226953 instances and the target variable column

Figure 2 shows the count of posts, the average word count (mean), the spread of word counts (standard deviation), the minimum and maximum word counts, and the range of lengths at different percentiles of the data.

```
count    226953.000000
mean        152.421744
std         253.073394
min           1.000000
25%          29.000000
50%          69.000000
75%         179.000000
max       16427.000000
Name: Length, dtype: float64
```

Figure 2: A statistical analysis of word lengths with count, mean, min, max and quantities

Clustering/Patterning, and Visualisation

1.1."Length" feature and "Deciding words" feature with "Suicidal label" target variable

Figure 3 shows an equal number of posts from each category ('suicidal' and 'non-suicidal'). The dataset used for analysis might be intentionally balanced to ensure equal representation of both categories and prevent any bias towards one category during analysis or modelling.

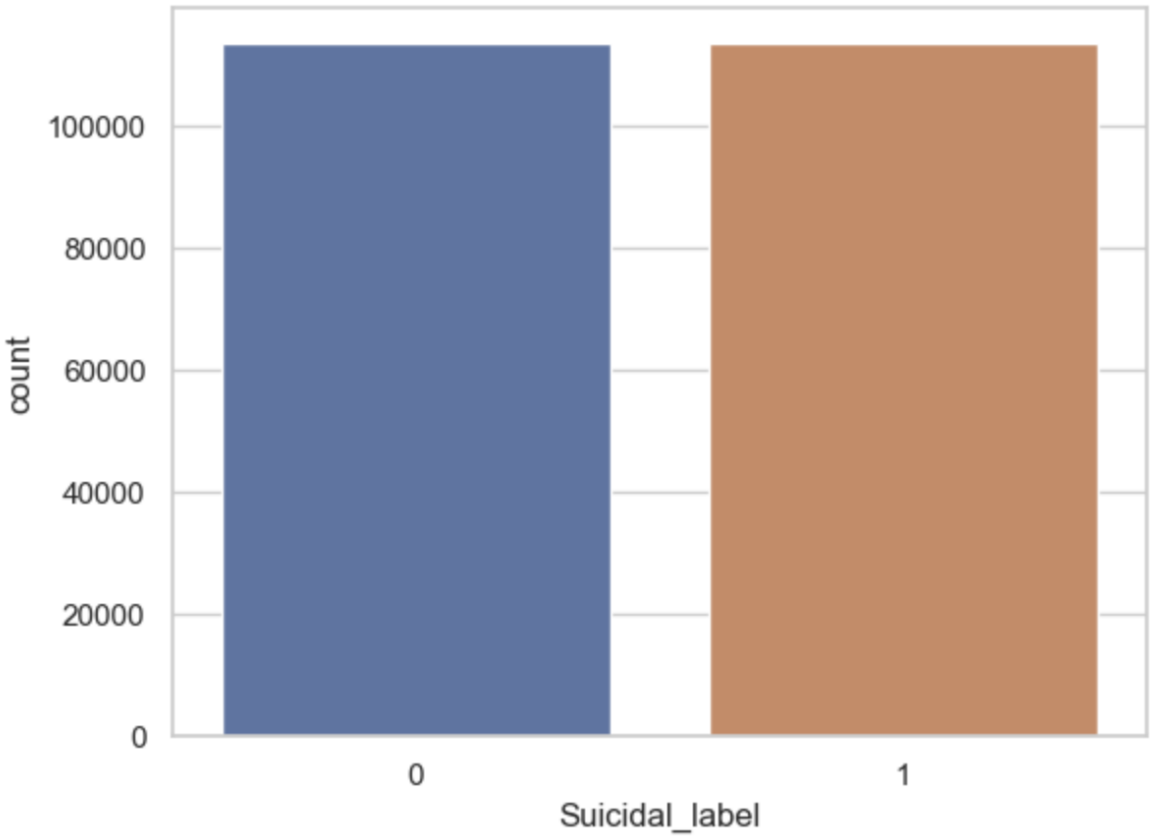

Figure 3: The initial number of posts from each class

We cut off the post length at 317 words to easily extract meaningful insights within a reasonable length range when extremely long posts might be outliers or contain irrelevant or repetitive content. Figure 4 shows the graph after cutting off the length of the posts. The number of posts

in the 'suicide' category becomes less than the number of posts in the 'non-suicide' category, so the removal of longer posts disproportionately affected the 'suicide' category.

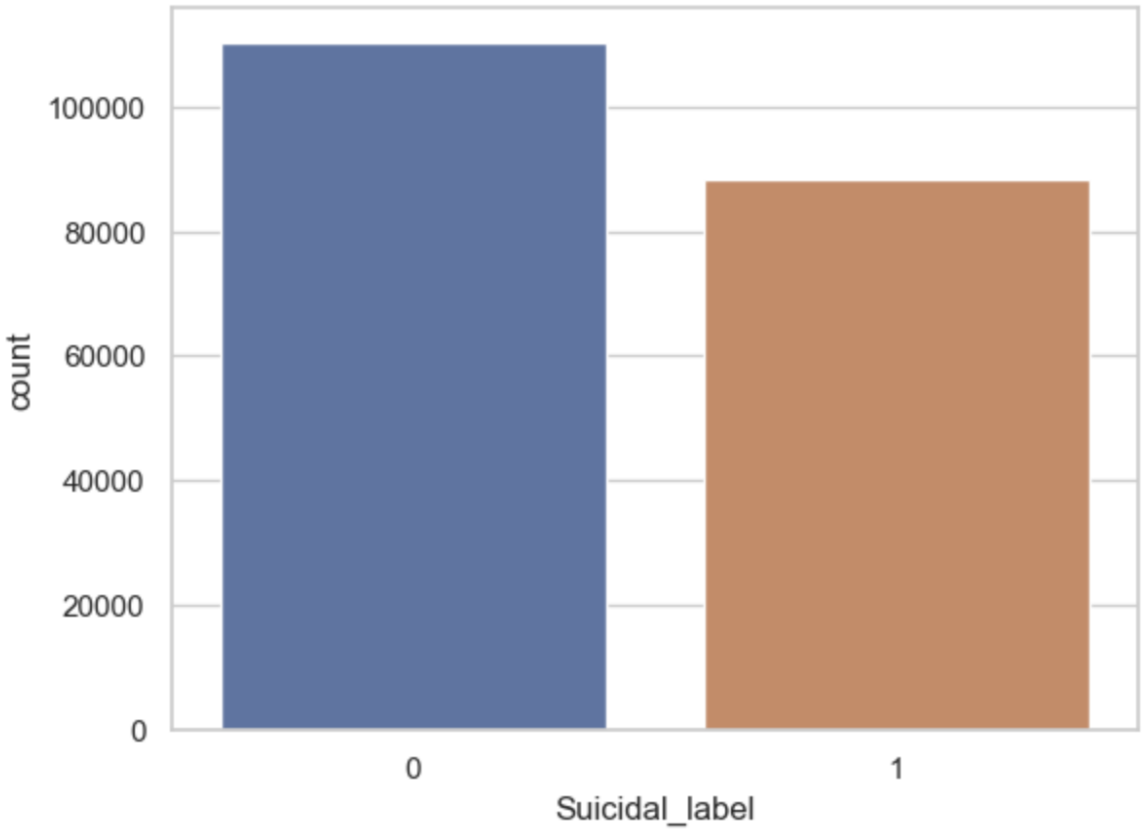

Figure 4: The number of posts from each class after cutting off the post length

The report will calculate the frequency of words. Word frequency provides insights into the importance and relevance of words within a given text and identifies the overall distribution of

word occurrences. However, figure 5 shows the lists of frequency words that are not suicidal-related, which might stop words and require more data filtering.

	Word	Frequency
28		1606277
4	I	1082859
18	TO	498740
16	AND	412877
89	THE	312340
...	...	...
215	THINK	37775
162	HE	37102
219	THERE	36610
27	HER	35517
361	GOING	34677

Figure 5: Word frequency at the limited post length of 317 words

The report continues to cut off the word frequencies below 31905 and above 3615. By removing such words, we focus on a subset of words that have a moderate frequency, allowing for a more balanced representation of the vocabulary and potentially capturing more relevant and distinctive terms. Figure 6 shows more suicide-related words.

	Word	Frequency
775	BROTHER	3665
1751	PROBLEM	3663
160	TOGETHER	3652
1807	HOSPITAL	3648
393	HEALTH	3645

Figure 6: Word frequency below 31905 and above 3615

1.2. Result:

Figure 8 visualises the differences in the average length of posts based on the category. It shows that the length of suicidal-related posts is longer than that of non-suicidal posts. Suicidal posts might include more contextual information, personal experiences, or detailed descriptions of their situation, contributing to longer text lengths.

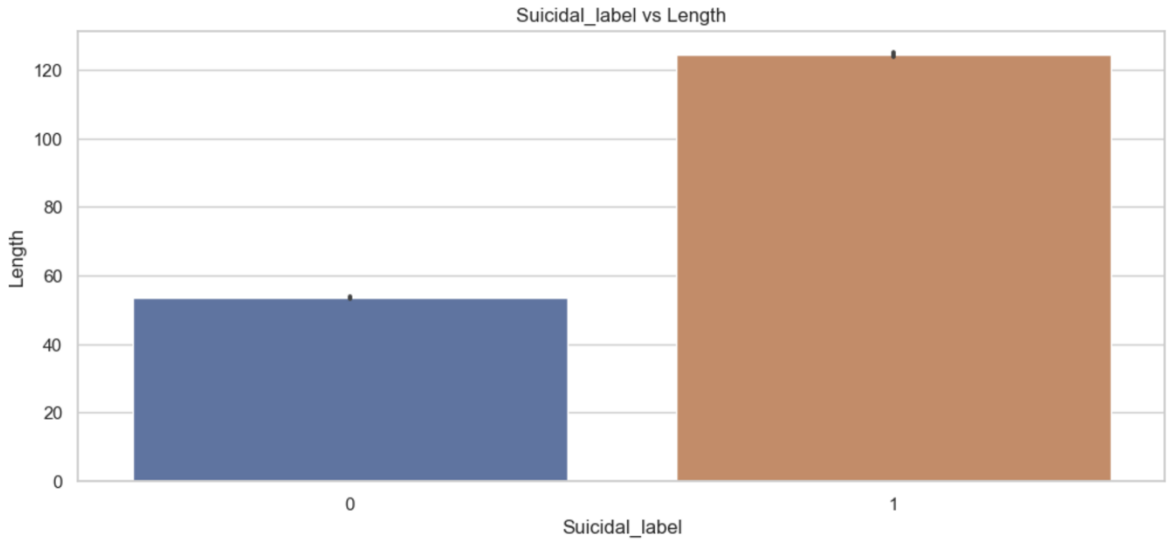

Figure 8: Suicidal class has a length of more than 120 compared to the non-suicidal class with approximately 50

Figures 9 and 10 measure the strength and direction of the linear relationship between each feature (word) and the target variable (suicidal label) by calculating the correlation coefficient. The words with a higher positive correlation, like "suicide", "kill", "die", and end," are positively associated with the target variable (suicidal intent), which means that the text containing those words will help to predict suicidal behavior with a higher accuracy rate. However, we can see the correlations are not high considering the non-linear relationship between the variables. This is because suicidal prediction datasets often involve analyzing emotional expressions and sentiments. Emotions are inherently complex and can manifest in diverse ways within a text.

	Feature	Correlation
1	Length	0.461558
45	SUICIDE	0.307711
34	KILL	0.307260
20	ANYMORE	0.304422
25	DIE	0.300119
36	END	0.271914
48	LIVE	0.248508
52	FAMILY	0.247769
116	SUICIDAL	0.241300
46	EVERYTHING	0.230240

Figure 9: The table shows the correlation of some "deciding words" with the target variable

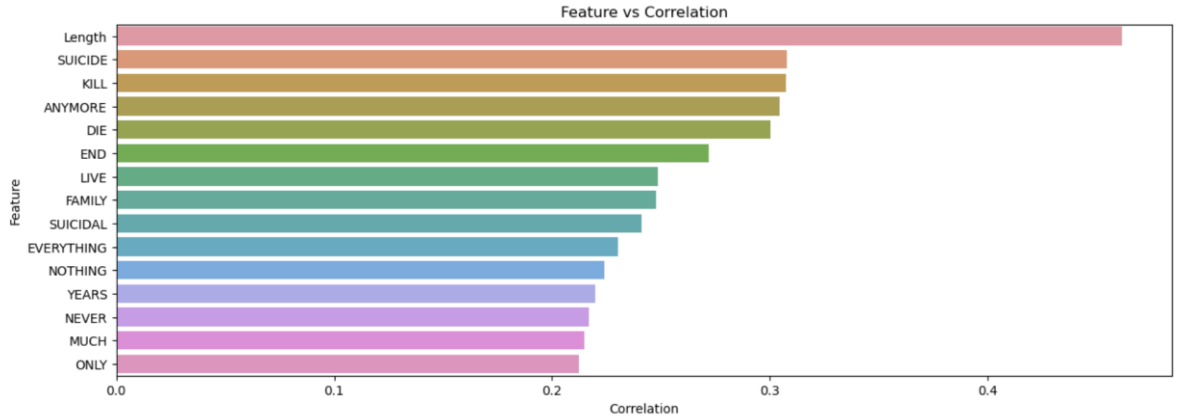

Figure 10: The graph visualises the correlation of some "deciding words" with the target variable

2. "Sentiment_label" feature with "Suicidal label" target variable

Figure 11 observes the relationship or pattern between the sentiment expressed in the posts (negative, neutral, or positive) and the presence of suicidal intent. Surprisingly, suicidal intent also appears in positive posts (1), with the percentage of suicidal intention (1) nearly equal to non-suicidal intention (0). Accordingly, analyzing the presence of suicidal intent related to negative feelings can indeed provide a more serious warning about suicide compared to suicidal intent related to positive or neutral feelings. Suicidal intent associated with negative feelings, such as despair, hopelessness, sadness, or extreme distress can reflect a more significant struggle with mental health issues and more vulnerable to suicidal thoughts and actions.

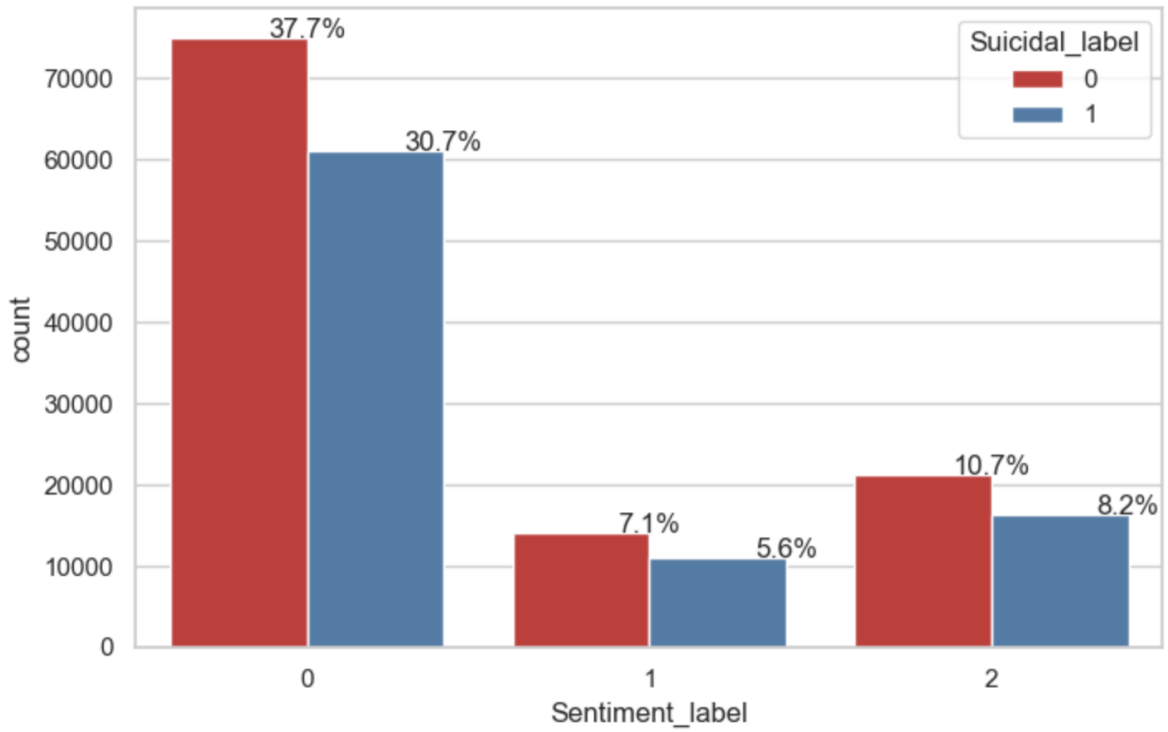

Figure 11: The graph shows the relationship between the sentiment expressed in the posts and the target variable

III. Problem refinement and summary

After analyzing the relationships between the "Length" feature, "Deciding words" feature, and "Sentiment_label" feature with the target variable, the report will build a Random Forest model to continue to answer the question "Can the suicidal intention of young people be predicted using language analysis of Reddit posts?". By analysing the feature's importance, Random Forest can identify the words that contribute the most to the prediction of suicidal-related posts. In addition, Random Forest can capture complex nonlinear relationships between input features and the target variable, allowing for more accurate predictions.

Referencing list:

1. Goyal, A 2021, *Reddit dataset for Multi-task NLP*, kaggle, 21 June, viewed 12 June 2023, https://www.kaggle.com/datasets/amangoyl/reddit-dataset-for-multi-task-nlp?select=Dataset_Suicidal_Sentiment.csv.
2. Komati, N 2021, *Suicide and Depression Detection*, kaggle, 19 May, viewed 12 June 2023, https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch.
3. Yadhu, A 2022, *Predicting Suicide and Word Analysis*, kaggle, 2 July, viewed 4 July 2023, https://www.kaggle.com/code/yadhua/predicting-suicide-and-word-analysis.