Name – Devarshi Dwivedi

Topic – NLP for suicide detection.

Dataset - https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

Github - https://github.com/devarshi167/Suicide-detection111

Drive - https://drive.google.com/drive/folders/123DXfzrxASqRidKsxbXR-K9SrC72wGpz?usp=sharing

Steps Followed –

Data Cleaning and Data Wrangling

To begin cleaning the data | ensured that the dataset did not have any duplicate posts or posts with null values (no text). Next, | began the process of normalizing the text in order to convert the messy social media posts into a neat machine-readable format.

The first steps in the process of text normalization include:

- Converting all text to lowercase
- Converting all hyperlinks and urls to standard text
- Converting all emojis and emoticons to text
- Removing punctuation and numerals
- Removing white spaces
- Ensuring all posts are written in the English language
- Expanding contractions

Dataset consisted of 232,074 social media posts from Reddit

Then we performed EDA and Modeling –

Three different machine learning models:

1) Multinomial Naive Bayes

2) Random Forest

3) Logistic Regression

Total of 12 models analyzed

Summary –

- Evaluated 12 Models
- Best model was Logistic Regression (TF-IDF Vectorizer - Text Only)
- Model seems to perform well enough to be useful.
- Additional input on performance from potential stakeholders.

Challenges faces and improvements –

- Extremely High ROC-AUC Score. Consider analysis between more similar groups .
- Enhance the engineered features.
- False negatives: Look into ways to correct text prior to prediction.
- More robust hyperparameter optimization.