Initial question: Can we develop a predictive model to analyze linguistic patterns and sentiment indicators in social media texts that accurately predict depression and suicide risk?

Problem description:

Suicide is a severe mental health issue, especially among young individuals, and social media often becomes a platform for expressing negative emotions and even suicidal thoughts. Many who deal with mental disorders consider their struggles as private problems and/or decide not to seek help because they do not want to be grouped as psychiatric patients. This project aims to identify language usage, sentiment patterns, and contextual cues from social media posts as early indicators of suicidal ideation. Through statistical and sentiment analysis, we extract features from textual datasets and employ supervised machine-learning techniques for accurate suicide detection.

Datasets:

Dataset 1: This dataset has 700,000 rows of text data from Reddit with 5 columns, including title, content, and post time. Approximately 34% of the posts are classified as borderline personality disorder (BPD).

Dataset 2: The Twitter dataset has 20,000 rows and 11 columns, collected from August 2015 to January 2017. It includes user ID, follower count, friend count, and favourite count, enabling the extraction of profile features for improved predictive modelling.

Big data analysis:

The datasets from different platforms have over 10,000 rows each, providing a diverse and sizable sample of posts, comments, tweets, and discussions related to mental disorders. Users from various backgrounds, demographics, and locations contribute valuable insights into mental health. The datasets are real-time with frequent updates, incorporating diverse language styles like informal language, slang, abbreviations, and emoticons. Preprocessing and analysis are crucial to extracting meaningful insights due to the noise present.

Initial processing of data:

Data preprocessing:

Preprocessing techniques such as tokenizing, stemming, stop word removal, part of speech, and spell-checking are used to remove noisy content, filter out unrelated hashtags or keywords, and handle user-generated noise like typos, slang, abbreviations, and emoticons.

Feature extraction:

Features including linguistic patterns (n-grams, TF-IDF), sentiment indicators (sentiment scores, emotion analysis), and contextual information (user demographics, time of posting) might be extracted by advanced techniques such as word embeddings or recurrent neural networks.

The statistical analysis measures tweets and subreddit posts, such as the number of posts and text length. In addition, it measures the ratio of suicidal posts to total posts per user. Sentiment analysis examines the type of emotion to compare scores between suicidal and non-suicidal users.

Predictive model:

The datasets will be labelled suicidal and non-suicidal. With the labelled and feature-extracted data, we create supervised classification predictive models to analyze linguistic patterns and sentiment indicators for accurately predicting suicide risk.

Deficiencies in the data:

Mental disorder datasets may not explicitly capture suicide ideation, as individuals may not openly discuss or disclose their thoughts.

Social media text can be challenging to interpret accurately due to informal language, sarcasm, irony, and cultural nuances.

Datasets may lack sufficient contextual information, such as demographic details, environmental factors, or personal circumstances, that can influence suicide ideation's presence and severity.

Solutions:

Developing datasets specifically focused on suicide ideation helps gather more direct and comprehensive information about suicidal thoughts.

Enriching textual datasets with external sources such as medical records, or psychological evaluations can provide a more comprehensive understanding of suicide ideation.

Using advanced NLP techniques, such as entity recognition and topic modelling extract contextual information from text.

Question refining: Can we identify young people's suicidal intentions from mental disorder-related social media posts?

Detecting depression or mental disorders is not necessary when there are different levels of mental issues to classify, so the project only focuses on examining whether a user has suicide ideation or not.

Backup questions and datasets:

Can we identify young people's suicidal intentions from suicide-related social media posts?

Identifying suicidal thoughts from mental-disorder-related textual data is not easy due to the limited sample size, but we can use the backup question and dataset to get a more accurate model. This dataset is a collection of 348,000 posts from "SuicideWatch" and "Depression" on the Reddit platform labelled as suicide or non-suicide, from December 2008 to January 2021.

How can machine learning algorithms be combined with different features to improve the accuracy of detection?

After successfully detecting suicide ideation from textual datasets, the next critical task is to compare and determine which machine-learning algorithms and features might create a hybrid machine-learning model with the highest accuracy.

Preliminary plans:

We utilize various machine learning algorithms, including logistic regression, random forests, support vector machines, and decision trees, to train different predictive models and evaluate them using appropriate evaluation metrics (accuracy, precision, recall, F1-score) and cross-validation techniques.

Reference list:

Adha, K 2022, *Mental Disorders Identification (Reddit)*, kaggle, 24 November, viewed 12 June 2023, https://www.kaggle.com/datasets/kamaruladha/mental-disorders-identification-reddit-nlp.

Chatterjee, M, Kumar, P, Samanta, P & Sarkar, D 2022, 'Suicide ideation detection from online social media: A multi-modal feature based technique', *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100103–.

Infamouscoder 2022, *Depression: Twitter Dataset + Feature Extraction*, kaggle, 6 June, viewed 12 June 2023, https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media.

Komati, N 2021, *Suicide and Depression Detection*, kaggle, 19 May, viewed 12 June 2023, https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch.