

ABSTRACT

The World Health Organization (WHO) in 2023 reported that more than 700.000 people die due to suicide every year. The process of suicide starts with suicidal ideation and then matures into a suicidal attempt. Nowadays many people are active on several popular social media networks such as Reddit to share their daily life and experiences, good or bad. Suicide can be prevented by doing early identification and detection and the phenomena recently been used as a domain for research related to the detection of suicidal ideation thought in social media. However, finding and comprehending patterns of suicidal ideation represents a challenging task and at the same time needs to overcome the natural language problem in social media. Therefore, this study aims to improve suicidal ideation detection using Machine Learning on Reddit using C-SSRS Reddit Suicide Dataset and Other non-mental-health subreddits by utilizing various feature enriching methods such as Content-Based, Linguistic Inquiry and Word Count (LIWC) to capture the pattern and Feature Expansion to overcome the natural language problem in social media then employed a baseline derived from previous research[10]. The best result from this study achieved 85% Accuracy and 82% Recall by combining all of the feature enrichment methods because each method successfully provides the model with additional information such as characteristics from Content-Based Feature, psychological dimension of message content from LIWC and also handle the vocabulary mismatch happen in social media by using Feature Expansion.

Keywords: Suicidal Ideation Detection, Reddit, Machine Learning, Feature Enriching, Content-Based, Linguistic Inquiry and Word Count (LIWC), Feature Expansion