Student Poster Session

Women In Data Science (WiDS)
Central Massachusetts Conference 2022

March 7th, 2022

Poster Title: StudentSADD versus DepreST: Collecting Data During COVID-19 for Rapid Mental Illness Screening

Authors: **ML Tlachac (presenter)**, Miranda Reisch, Ricardo Flores, Elke Rundensteiner

Abstract: The COVID-19 pandemic has only further exacerbated the increasing rates of mental illness among the general population. Given the need for universal mental illness screening technology, we deployed apps to collect data from over 300 college students and over 400 crowdsourced workers, resulting in the StudentSADD and DepreST datasets. Both sets of participants shared voice recordings, which we successfully used with Audio Assisted BERT (AudiBERT) to screen for common mental illnesses. Further, we used Gated Recurrent Units (GRU) models to screen for depression with time series of DepreST call and text logs. The datasets are valuable resources for the mobile health research community.

Poster Title: Constructing lexicons to improve depression screening with texts

Authors: **Avantika Shrestha (presenter)**, ML Tlachac, Mahum Shah, Benjamin Litterer, and Elke A. Rundensteiner

Abstract: Given that depression is one of the most prevalent mental illnesses, developing effective and unobtrusive diagnosis tools is of great importance. Recent work that screens for depression with text messages leverage models relying on lexical category features. Given the colloquial nature of text messages, the performance of these models may be limited by formal lexicons. We thus propose a strategy to automatically construct alternative lexicons that contain more relevant and colloquial terms. Specifically, we generate 36 lexicons from fiction, forum,and news corpuses. We then compare the depression screening capabilities of these lexicons for text messages with machine learning models. Out of our 36 constructed lexicons, 17 achieved statistically significantly higher average F1 scores over the pre-existing formal lexicon and basic bag-of-words approach. In comparison to the pre-existing lexicon, our best performing lexicon increased the average F1 scores by 10%. We thus confirm our hypothesis that less formal lexicons can improve the performance of classification models that screen for depression with text messages. By providing our automatically constructed lexicons, we aid future machine learning research that leverages less formal text.

Poster Title: Ensemble Models for Depression Classification

Authors: **Saskia Senn (presenter)**, ML Tlachac, Ricardo Flores, Elke Rundensteiner

Abstract: Depression is among the most prevalent mental health disorders and increasing prevalence worldwide. While early detection is critical for the prognosis of depression treatment, detecting depression is challenging. Previous deep learning research has thus begun to detect depression with the transcripts of clinical interview questions. Since approaches using Bidirectional Encoder Representations from Transformers (BERT) have demonstrated particular promise, we hypothesize that ensembles of BERT variants will improve depression detection. Thus, in this research, we compare the depression classification abilities of three BERT variants and four ensembles of BERT variants on the transcripts of responses to 12 clinical interview questions. Specifically, we implement the ensembles with different ensemble strategies, number of model components, and architectural layer combinations. Our results demonstrate that ensembles increase mean F1 scores and robustness across clinical interview data.

Poster Title: K-means Clustering of Student Behavioral Patterns and Advanced Visualization Methods of Learning Technology Data

Authors: **Reilly Norum (presenter)**, Jieun Lee, Erin Ottmar, Lane Harrison

Abstract: This paper presents two studies where distinct student profiles (N = 760) emerged based on their behavioral patterns in an online algebraic learning game. We applied k-means clustering analysis to clickstream data collected in the game and then examined how studentsâ€™ behavioral patterns varied across the clusters using data visualization. The results identified four groups of students based on their in-game behaviors, showing that there was a large variation in their behavioral patterns for engaging with the game.

Poster Title: Extraction of Named Entities from Text Messages

Authors: **Katie Houskeeper (presenter)**, Matthew Dzwil, Dante Amicarella, ML Tlachac

Abstract: Named-Entity-Recognition (NER) extracts information from unstructured sources and categorizes specific entities within the body of text. Our goal for the project was to use machine learning algorithms to remove HIPAA-protected and other Personally Identifiable Information (PII) from text messages. Previous WPI teams collected these crowd-sourced texts for the purpose of depression screening, but their use is limited to IRB-approved researchers since they contain PII. Our team concentrated on four types of named entities to identify: names, locations, organizations, and miscellaneous entities. We manually labeled the text messages and then assessed the effectiveness of machine learning models on NER. Specifically, we fine-tuned bert-base-NER to extract the named entities from the text messages. Our model could be used to extract PPI for in text messages for other datasets within and outside of this domain.

Poster Title: Understanding The Relationship Between Low Birth Weight and Gestation Periods Using Regression Modeling

Authors: **Jaya Kolluri (presenter)**,

Abstract: The high prevalence of preterm deliveries (i.e., short gestation periods) and their impact on neonatal mortality rates have been a significant challenge for the medical community. It is also widely known that low birth weight is one of the leading causes of neonatal mortalities. According to previous studies, neonatal mortality rates account for 17 deaths per every 1000 total deliveries.1 This study focused on understanding the relationship between low birth weight (LWB) and short gestation periods using a data set consisting of patient records of over 900 pregnant women in King County. Preliminary data exploration identified a high correlation between low birth weight and preterm pregnancies. Using the R statistical package, a simple regression model was developed to validate the correlation between LBWs and gestation periods. By building a simple regression model, it was concluded that our model could precisely predict the weight of a preterm baby throughout the course of a pregnancy. Our model was both working functionally and accurately with a high confidence level leading us to believe we could use this same procedure on real-time data as our next steps. However, we also noticed that the R squared score was low, indicating a high variance in the model predictions. The high variance suggests that there is room for improving the model performance with the addition of external related variables such as patient demographics.