Erdos Data Science Bootcamp May 2024 Executive Summary

Mental Health NLP Team: Emelie Curl, Alejandra Dashe, Mitch Hamidi, Gabe Khan, & Eunbin Kim
Github: https://github.com/mitch-hamidi/may-2024-mental-health-nlp

Overview: Using classification and Natural Language Processing (NLP) to analyze web-scraped data from Reddit, can we (1) identify who is undergoing or interested in mental health treatment, and (2) predict preference for treatment?

Existing pain point: Chronic mental health diagnoses, like borderline personality disorder (BPD), have historically been viewed as more challenging to treat due to comorbidities, provider bias, and lack of a standardized treatment plan.

Stakeholders: Healthcare providers (therapists, psychologists, psychiatrists, social workers, and treatment hospitals), BPD patients, BPD advocacy groups, and pharmaceutical companies. Our project can provide insight into patient preferences, needs, and pain points regarding treatment.

Data Source: A Kaggle dataset of scraped Reddit posts and a team-created dataset of scraped comments from eight BPD-relevant subreddits using 89 keywords of interest.

Methodology: After preliminary trial models built using a Kaggle dataset, we built our classification and NLP models on our scraped Reddit dataset (known as "dataset" throughout this document). We scraped using PRAW: The Python Reddit API Wrapper. The scraped data was then combined, preprocessed, and cleaned by accounting for duplicates in comments, eliminating missing or deleted comments and authors.

Input variable: was natural language text from posts and comments.

Output variables: Relevance of Content, Treatment Status, and Recommendation of Treatment.

 We vectorized the comments into numerical values using "CountVectorizer," and the "TfidfVectorizer". The train-test split was performed on approximately 500 manually coded comments classified as relevant or not relevant to BPD treatment, as well as recommendation sentiment. Then, we used an NER technique that was created to detect medication names, and "SentimentIntensityAnalyzer" imported from "nltk.sentiment.vader" to determine the valence of users' treatment experiences.
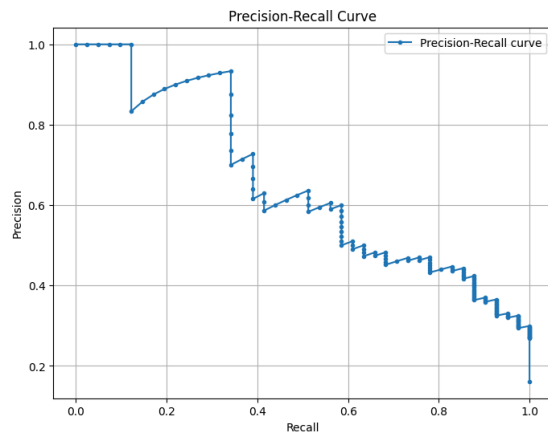
Finally, we tested our model using a Decision Tree classifier.

Discussion:

Results and outcomes:

- Trained on Kaggle data: among five model classes, a zero-inflated keyword+text model for relevancy classification was chosen, with area under precision recall score of 0.65

(compared to a baseline of .16 for random selection).

- Trained on scraped data: the Decision Tree model for relevancy classification was the best performer, with a recall of 0.7.

Limitations:

- The models and validation could be improved with more human-coded data.
- The scraped test set may not be representative of the actual distribution of relevant to non-relevant posts.
- The sentiment analysis model developed needs to be more sophisticated to capture some of the nuances of natural language when discussing various treatment plans. For example, one comment could express multiple sentiments while discussing aspects of the treatment plan.

Future Directions:

- Adapt models to evaluate relevance and sentiment for other mental health illnesses.
- Build a more sophisticated sentiment analysis model to capture nuance in language and multiple sentiments that could appear in a single post.
- Test preliminary models for relevancy classification on a larger data set.