Personality Disorder Language Analysis

Dhrishti Hazari

Medical terminology evolves at a turtle-like pace over time as well as location. This gradual change is minute and often undetectable to those who are not well versed in the linguistic history of such niche topics, both inside and out of that profession. Studying and understanding the evolution of language commonly associated with personality disorders can assist with detecting biases and stigmas, detecting misinformation, and even scientific communication by identifying outdated terminology.

The goal of this project is to identify terminology commonly associated with various personality disorders throughout time, unlimited by any person's scope of knowledge. The final output is a classification model capable of categorizing new articles into thematic cluster and predicted year based on historical patterns. The results can then be interpreted in multiple ways, only a few listed above.

PROPOSED TASKS:

1) **Collecting appropriate data:** Fordham's online library of resources has some journals related to Personality Disorders. They will be downloaded and stored as either PDF or HTML files, separated by years.
2) **Create a DataFrame that will store information for each journal:** Journal name, Year, Title, file_path, Title Terms, Text, Word Frequency, Word Correlations, Classification
3) **Parse, clean, and save the data:** Using Python libraries, convert the files into text. Store the appropriate Journal name, Year, Title, file_path, and Text for each. Then clean all the

data, removing terminology that is considered unmeaningful, disregarding punctuation, and resolving errors caused by conversion. Store the text as a tokenized array.

4) **Sequential pattern mining:** Perform a sequential pattern search for each journal, using various support and confidence thresholds to determine the optimal output. This method allows us to find repeated phrases that may be otherwise disregarded if only searching for repeated terminology. Store this in the "Word Frequency".
5) **Association pattern mining:** Identify meaningful terminology in each title by **Parsing and cleaning** the titles, using the titles as a transaction and the word frequencies as itemset. Do this on a yearly basis. Store the words in the title that have the optimal lift and confidence for each title in "Title terms", and the entire rule generated in "word associations" along with its confidence score.
6) **Clustering** to detect similarities – use hierarchal clustering to group journals based on word embeddings. Separate based on thematic groupings.
7) **Visualize** the above two – use methods such as cosine similarity, converting words into vectors to display the similarities and differences of words over time, word clouds, hierarchal clustering, etc.
8) **SVM Classification:** Classify articles based on mined patterns. Use the "Word Frequencies", "Word Association", "Title Terms", and "Year". Create a new data frame, with each possible phrase found in the word frequencies column turning into its own column as well as each association rule from title terms becoming its own column. The word frequencies will contain binary values for "appears" or "does not appear" while the association columns will contain the confidence scores for each. The labels to be predicted will be the values found in "Title Terms" alongside "Year".

9) **Sentiment analysis:** Use logistic regression with a softmax classification to predict the probability of a term being used positively or negatively

EXAMPLE:

```
data = pd.DataFrame({
    "contains_borderline": [1, 0, 1, 0],  # If title contains "borderline"
    "contains_schizophrenia": [0, 1, 0, 0],  # If title contains "schizophrenia"
    "pattern_BPD_related": [0.8, 0.1, 0.9, 0.2],  # Confidence score of pattern matching BPD
    "pattern_Schizo_related": [0.2, 0.9, 0.3, 0.8],  # Confidence score for schizophrenia
    "label": [["BPD", "Schizophrenia", "BPD", "Schizophrenia"],[
```

TIMELINE:

- Week 1-2: Data collection and preprocessing.
- Week 3-4: Pattern mining and clustering.
- Week 5-6: Visualization and classification.
- Week 7: Final analysis and report generation.

RESOURCES:

- Personality Disorders: Theory, Research, and Treatment