**Text Mining and Readability Test of Diagnostic and Statistical Manual of Mental Disorder**

SDS 400: NLP Special Studies Final Report

Zhu Shen

Professor. Jordan Crouser

**Abstract**

The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) were written by 13 separate groups of psychiatrists in extremely technical language that may be intimidating and abstruse to primary social workers who need to make diagnosis using this reference book. However, limited understanding of the document causes misconception of the diagnostic features and criteria, and when people assume that they fully understand the text, they would never go back and double check whether they are right. Consequently, misconception leads to misdiagnosis, an alarmingly prevalent phenomenon all the time. According to a 2012 study by Bipolar UK and Bipolar Scotland, “people with bipolar disorder wait for an average of 13.2 years before they are diagnosed, and often spend years receiving treatment for other conditions.”[1] In the special study, I used Python NLTK library and Smog Index [2] for text readability to test the assumption: the more difficult the description for the disorder is, the disorder is more likely to be misdiagnosed.

**Data**

The Beautiful Soup, a Python package for scraping web data, was used to fetch diagnostic features and associated features of all disorders from DSM-5 online version and the text was saved as plain text file. Only diagnostic features and associated features were used for analysis because sentences in diagnostic criteria are commonly too short to do effective readability test and diagnostic/associated features are often the supporting evidence for socials workers to make diagnosis.

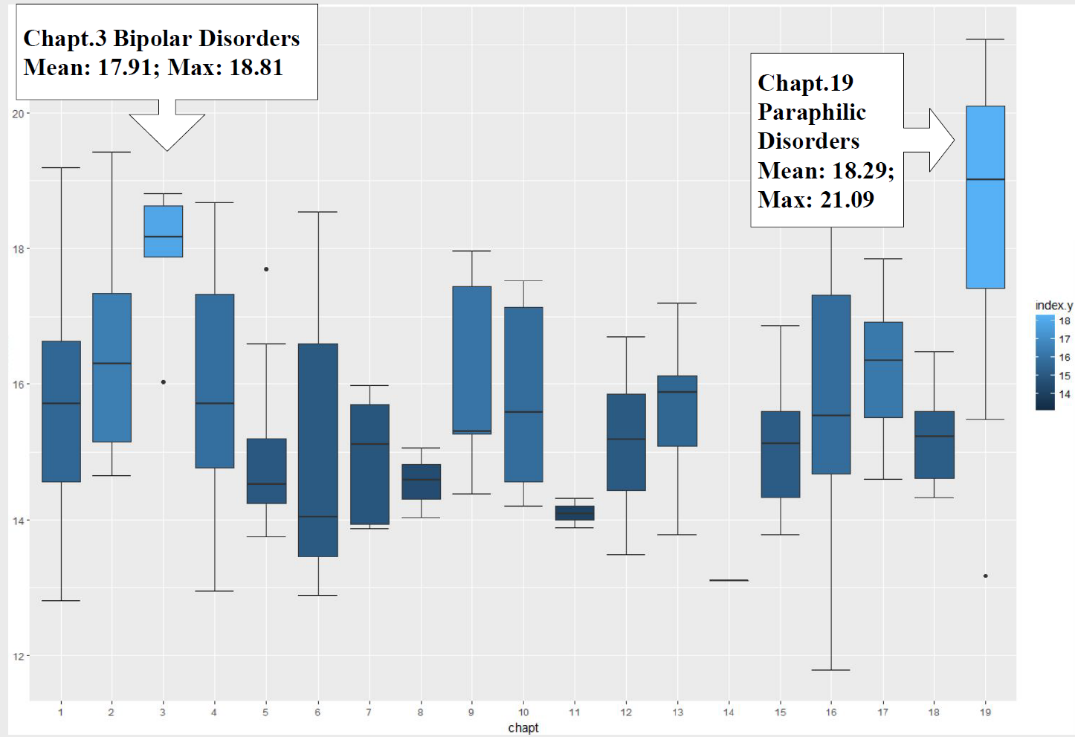
**Readability Test**

Before computing Smog Index for any sentence or paragraph, the text was cleaned by eliminating names of disorder, reference and non-character notations (all non-letter elements except for numbers and periods). In order to check whether a disorder is hard because one or two sentences in the text is extremely hard or a disorder is hard in general, I calculated the difficulty ratio (the number of polysyllabic words over the number of words in the sentence) per sentence for each of 158 disorders to detect any sentence with extremely high difficulty level and is recognized as an outlier. Then unusually difficult sentences for each diagnosis were analyzed and removed if they are irrelevant to making diagnosis. Here are some sentences that I will consider eliminating (part of sentence):

* About coding notes: “If the substance is known, it should be reflected in the name of the disorder upon coding.” (Chapter 16 Diagnostic Feature of Other (or Unknown) Substance Intoxication)
* Brief Introduction of disorders: “Intellectual disability is a heterogeneous condition with multiple causes.” (Chapter 1 Associated Feature of Intellectual Disability (Intellectual Developmental Disorder))
* Refer to a diagnostic criteria: “The learning difficulties manifest as a range of observable, descriptive behaviors or symptoms (as listed in Criteria A1–A6).” (Chapter 1 Specific Learning Disorder) But here I will delete the parenthesis but not the complete sentence.

After cleaning and standardizing these difficult sentences, I parsed the text for each disorder into the Smog Index Calculator in Python and created the following boxplot.

**Figure 1. Boxplot Showing Smog Index for 19 Chapters in DSM-5**



**Result and Discussion**

The boxplot shows that Chapter 19 (Paraphilic Disorders) and Chapter 3 (Bipolar and Related Disorder) have significantly higher Smog index than other chapters and Chapter 14 (Gender Dysphoria) is the easiest chapter containing only one disorder with a Smog Index 13.1. Such result somehow explains the high misdiagnosis possibility of Bipolar Disorder, a disorder in Chapter 3. Of all 158 disorders, the mean Smog Index is 15.76, which means on average the text in the DSM-5 is difficult to understand even for a college graduate. However, some other commonly misdiagnosed disorders, such as Borderline Personality Disorder (14.33), Schizophrenia (14.65) and ADHD (14.83), have Smog indexes which are below the average. Such discrepancy suggests two possible reasons for misdiagnosis: (1) diagnostic feature is too difficult to understand; (2) diagnostic feature is relatively simple but fails to contain enough information to distinguish.

**Conclusion**

We can conclude that misdiagnosis might be related to the difficulty level of the text but it is not true for all cases. When text in diagnostic features is simple but obscure, not containing enough information to distinguish the disorder, misdiagnosis would also occur with higher possibility. However, we should also realize another possibility that there is actually no different between disorders; the classification is wrong essentially. A future direction for revision of DSM-5 is finding a balance; description of the disorder should have lower difficulty level in general but specialized enough for the key difference with other disorders.

**Acknowledge**

This project was completed in partial fulfillment of the requirements of SDS 400: Special Studies. This course was taught by R. Jordan Crouser during Fall 2017. I would also thank Syeda Zainab Aqdas Rizvi for her generous help for data cleaning.

**Reference**

1. The Dangers of Mental Health Misdiagnosis: Why Accuracy Matters. (2017, November 10). Retrieved December 11, 2017, from https://www.bridgestorecovery.com/blog/the-dangers-of-mental-health-misdiagnosis-why-accuracy-matters/

2. “A Readability Assessment of Online Parkinson’s Disease Information.” J. R Coll Physicians Edinb. 2010.