First Validation Task – Corpus Extension

What is the Corpus Extension?

It's an API developed by KARV Data Consulting Ltd. That linguistically extrapolates a limited corpus of text, in order to increase the size of dataset, and in turn a more comprehensive input to our NLP-AI models, and get around the feature-sparsity limitation present in many non-rich data problems.

How are we applying it to PsychAI?

The Corpus Extension loops through our corpus, performs Named Entity Recognition, and Part of Speech Tagging, while applying a medical lexicon known as SNOMED CT to discern the important terms & phrases to keep intact, then web scrapes concise meanings, and synonyms of all relatively replaceable words, and extractively summarizes them with an unsupervised NLP model, and accommodates them cohesively within the corpus, consequently generating many more augmented corpuses.

What would we like JHU to do?

We have run the corpus extension again 300 sample psychiatric clinical cases (along with random individual sentences within them). We have generated ~20 augmented copies per case & ~8 augmented copies per sentence. We would like you to validate the viability of using each of these artificial copies as input data for our NLP-AI, by judging how closely similar they are (in terms of meaning) to the original, and translating that to a quantified score from 1 to 4. Please use the following criteria while scoring:

Score 1: Augmented case/sentence preserves very little to no relation to the original clinical meaning, and the interpretation of text is fundamentally contrasting.

Score 2: Augmented case/sentence resembles some congruence to the original (even if limited) while some phrases have no clinical meaning, and the overall interpretation is disorganized.

Score 3: Augmented case/sentence is functionally equivalent to the original, with well-founded differences that are valid, and help contribute to the same meaning.

Score 4: Augmented case/sentence is more or less the same as the original with very little grammatical and content changes – effectively synonymous.

How will this help us?

This validation exercise will allow us to significantly boost the precision of our Corpus Extension API's outputs. We will use your scores to train a supervised AI model to recognize "bad" augmentations (as reference, to filter out those cases/sentences which perpetuate the characteristics of cases/sentences deemed of score 3) that are being produced, in real time, and limit them. This will allow us to have an authentically more unambiguous training set, hence removing bias from our diagnostics, and risk-factor predictive NLP-AI models. This would accordingly lead to a considerable stimulation of our application's fidelity, and improve patient experience.