Preliminary Results:

Authors: Jane Acierno, Yuri Oh, Camden Parker, Jonathan Phillips, Gokul Srinivasan.

Terminology:

***Beck Score*:** The beck score is the sum of an individual's Beck's Depression Inventory (BDI) and Beck's Anxiety Inventory (BAI) scores:

$$Beck = BDI + BAI.$$

We are using this score to represent the joint influence of both depression and anxiety. For reference, individuals with lower beck scores tend to be less afflicted with anxiety and depression, whereas the converse is also true.

Sentiment Score: A sentiment model in NLP is a model that, when presented a sentence as input, outputs the predicted sentiment score between zero and one. Zero corresponds to very negative sentiment, whereas one corresponds to very positive sentiment.

Experimental Overview:

Two hundred participants (**n = 200**) were presented a series of ten vignettes, within which they were asked to generate six contextually relevant possibilities per vignette. For example, participants were provided the following vignette:

> "You are leaving the mall on a hot summer day. In the parking lot, you notice a dog in the back of a car without any of its windows open. The dog is panting heavily and looks tired."

Once they completed reading the vignette, they were asked: "In this situation, what are some things you could do?"

Thus, each respondent generated 60 possibilities throughout their experiment. And, in total, 12,000 possibilities were generated.

Global Average Sentiment vs. Beck:

We hypothesized that individuals who scored high on depression and anxiety-related measures (e.g., beck) would demonstrate lower global average sentiment scores, where the global average sentiment was defined as:

$$Global\ Average\ Sentiment = \frac{\sum_{g \in generations} predict\_sentiment(g)}{|generations|}$$

This hypothesis was inspired by the symptomatology of patients suffering from severe depression. These patients often report feelings of intense helplessness and hopelessness. One explanation for these feelings may involve the core faculty involved with modal cognition. The intuition was this: hopelessness might alternately be described as an inability to generate sufficiently compelling possibilities. If this is on the right track, we would predict to see some signs of this process within the possibilities generated. One such signature might manifest in the sentiment of these possibilities.

However, as the data shows, there were no significant correlations between the global average sentiment and beck score.

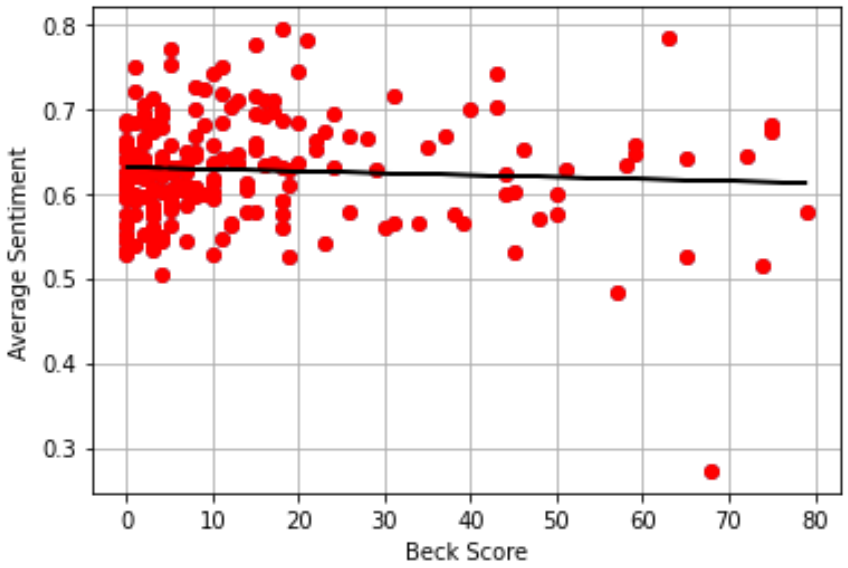

Figure 1: Average Sentiment vs. Beck Score

R-value: -0.065245
P-value: 0.362347

Average Sentiment vs. Possibility (Stratified by Beck):

Another suite of experiments consisted of determining the average sentiment per possibility number. Recall that participants are asked to generate six possibilities per vignette. Per Phillips, Morris & Cushman 2019, we would expect the earlier possibilities generated to be "better" (at the moment, this notion is ambiguous, but it will shortly be rectified) than those later ones. Here, we take "better" to mean something like the sentiment of the earlier possibility generations ought to be more positive than those later generations.

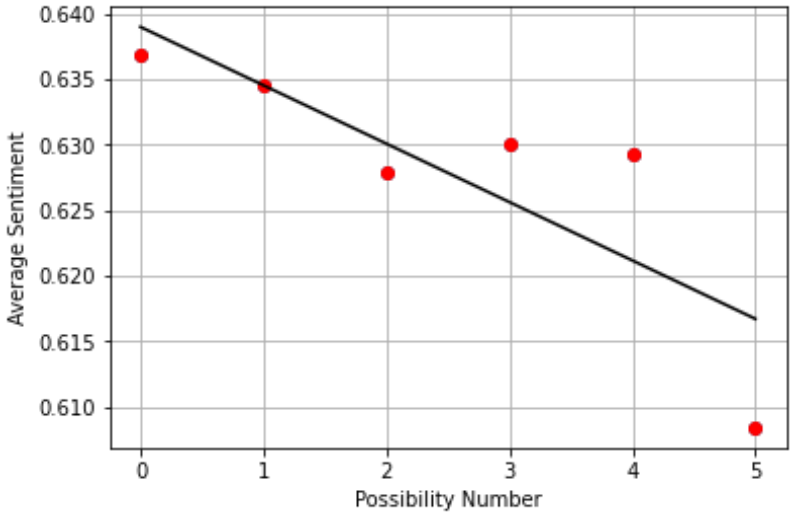

Figure 2: Average Sentiment vs. Possibility Number

R-value: -0.824304
P-value: 0.043592

The results shown in *Figure 2* seem to support the view advanced by Phillips et al., 2019. The first possibilities generated (and these possibilities are likely the first that came to mind) possess higher sentiment than those that follow. This data is agnostic towards the specific mechanisms described in Phillips et al., 2019 (on-line processing vs. candidate set generation), yet still supports the general theory in a separate arena: sentiment.

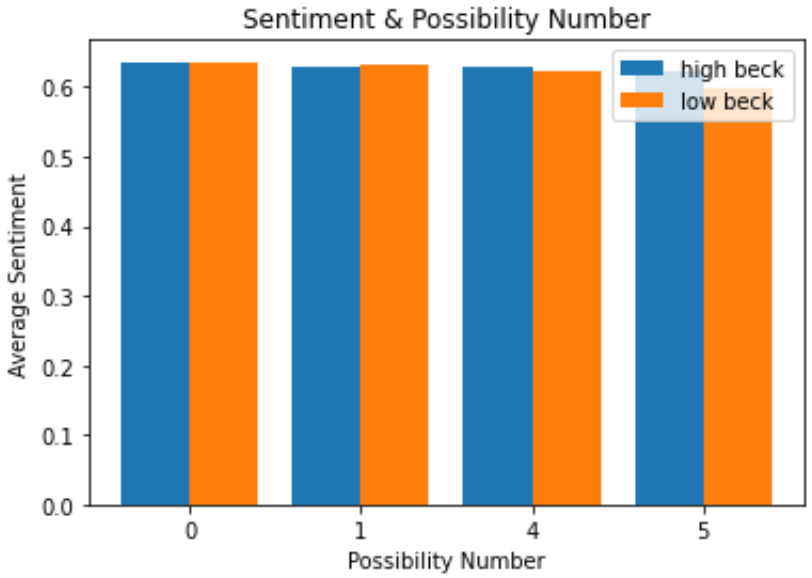

Figure 3: Average Sentiment vs. Possibility Number

High Beck	*0.63472*	*0. 628452*	*0. 628026*	*0. 622213*
Low Beck	0. 63604	0. 631753	0. 621997	0. 598542

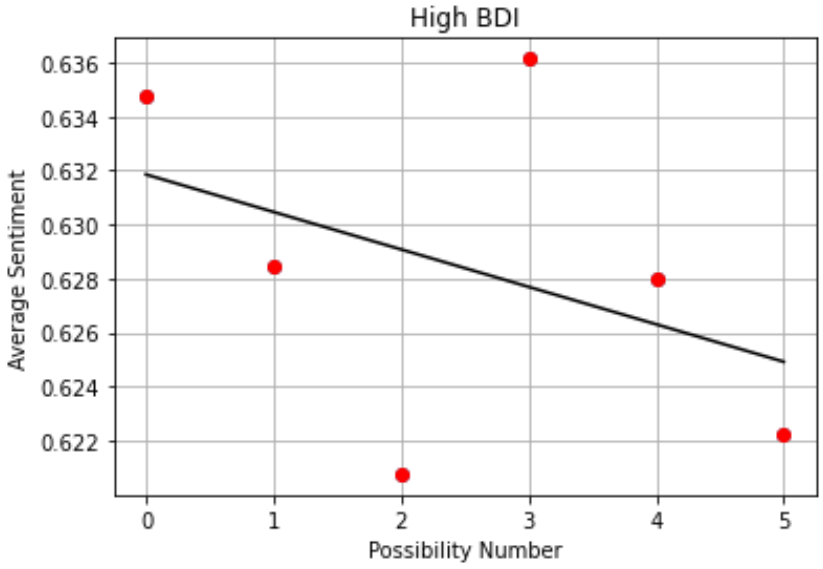

Figure 4: Average Sentiment vs. Possibility Number (High BDI)

R-value: -0.413531
P-value: 0.415062

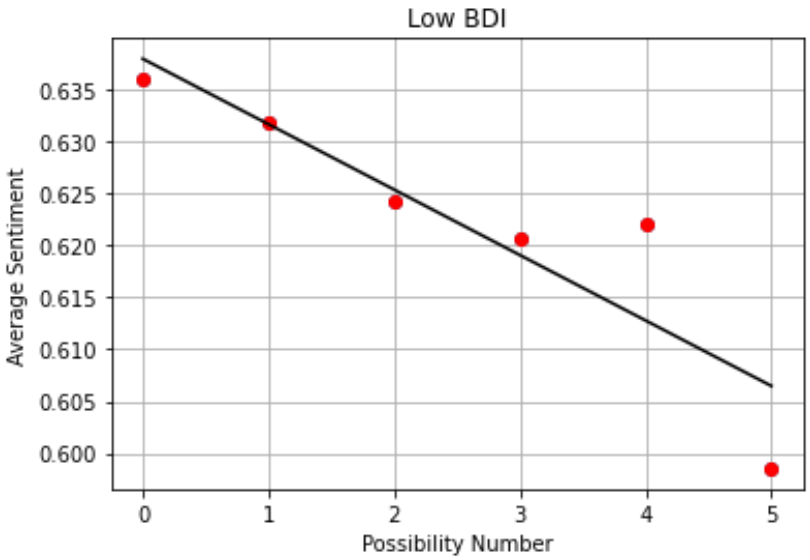

Figure 5: Average Sentiment vs. Possibility Number (Low BDI)

R-value: -0.903959
P-value: 0.013393

We also wondered whether individual differences would impact the general effect observed in *Figure 2.* To this end, we hypothesized that individuals with depression might invert the possibility generation pattern described by Phillips et al., 2019. If this were true, it would lend immense credibility to a host of modern cognitive-behavioral techniques. For example, if individuals with depression tended to generate worse possibilities first, and better possibilities later, therapeutic approaches that emphasize systematically generating more possibilities would generally have positive outcomes. Thus, in *Figure 3,* the average sentiment of each possibility was plotted for two groups: those deemed to have a "high" beck score and those with a "low" beck score. [1] As shown in *Figure 3,* those with higher beck scores demonstrated a unique sentiment pattering compared to those with lower beck scores. In the high beck score group, average sentiment does not decrease as much through successive possibilities as those in the low beck score group. On one reading of this data, it could be said that individuals with more severe depression tend not to generate the best possibilities first and the worst possibilities later. Instead, they seem to do the opposite (as seen in *Figure 6*), in line with our hypothesis. On the other hand, individuals with low beck scores exhibit the sharp decline in "goodness" that we expected through successive possibility generations.

Sentiment Gradient vs. Beck:

[1] The thresholds here were arbitrarily selected. We deemed high beck to be any score greater than 20 and low beck to be any score lower than 10.

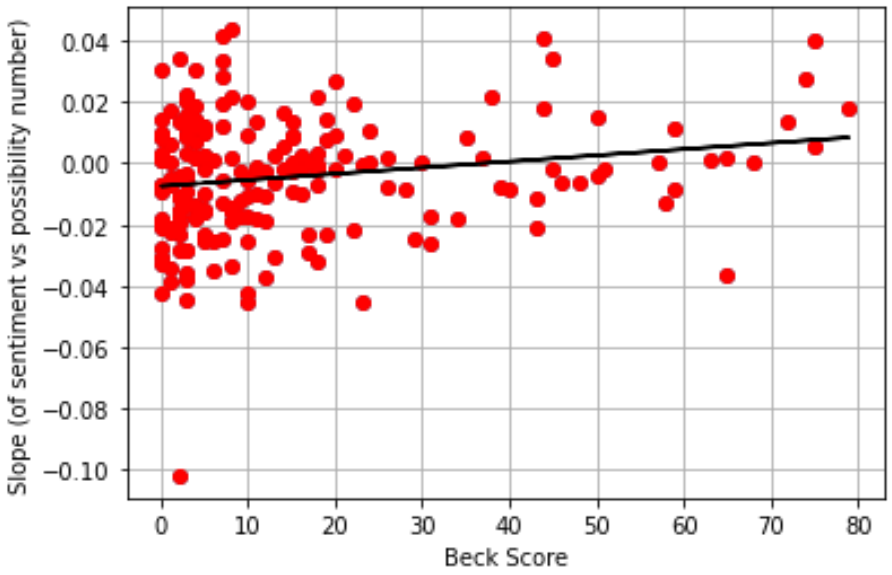

Figure 6: Sentiment Gradient vs. Beck Score

R-value: 0.188277
P-value: 0.008061

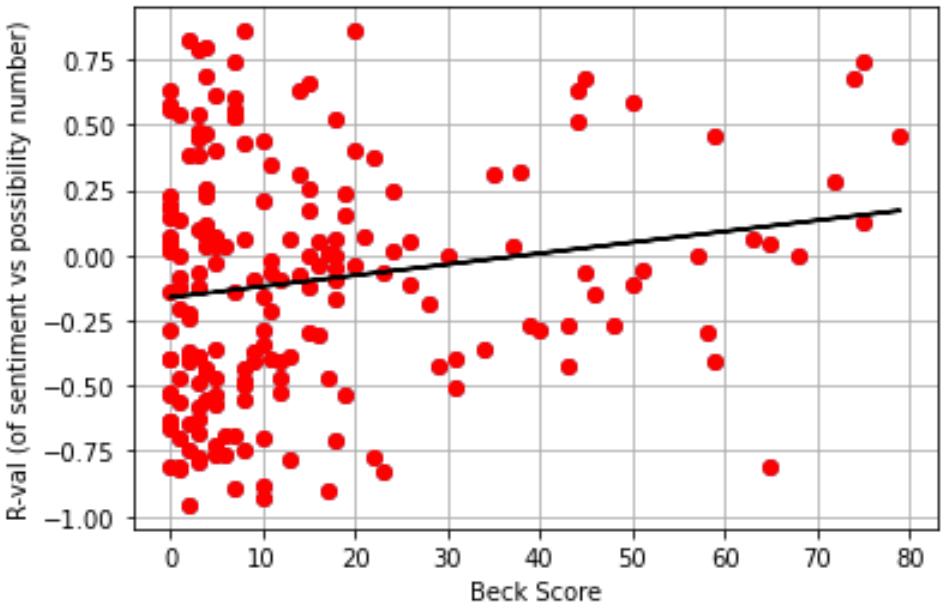

Figure 7: Sentiment Gradient Modified vs. Beck Score

R-value: 0.171563
P-value: 0.015927

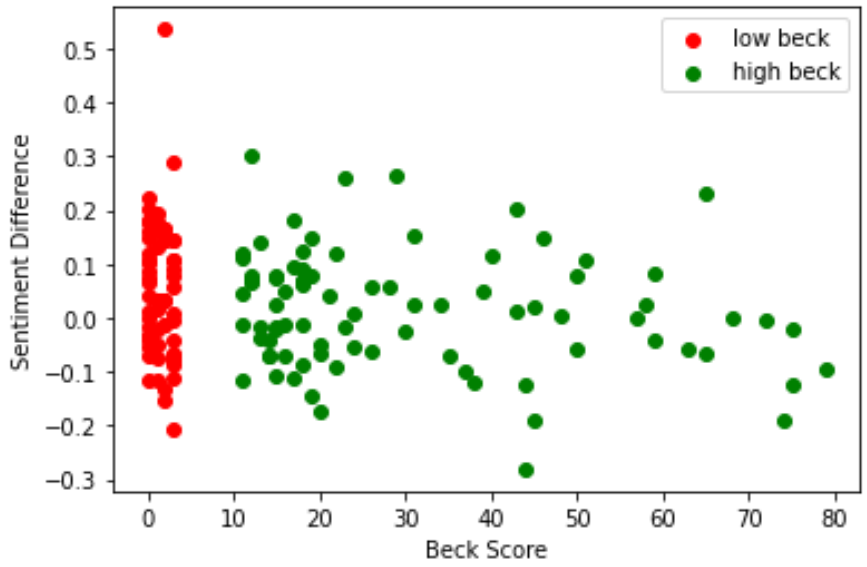

Figure 8: Sentiment Difference vs. Beck Score

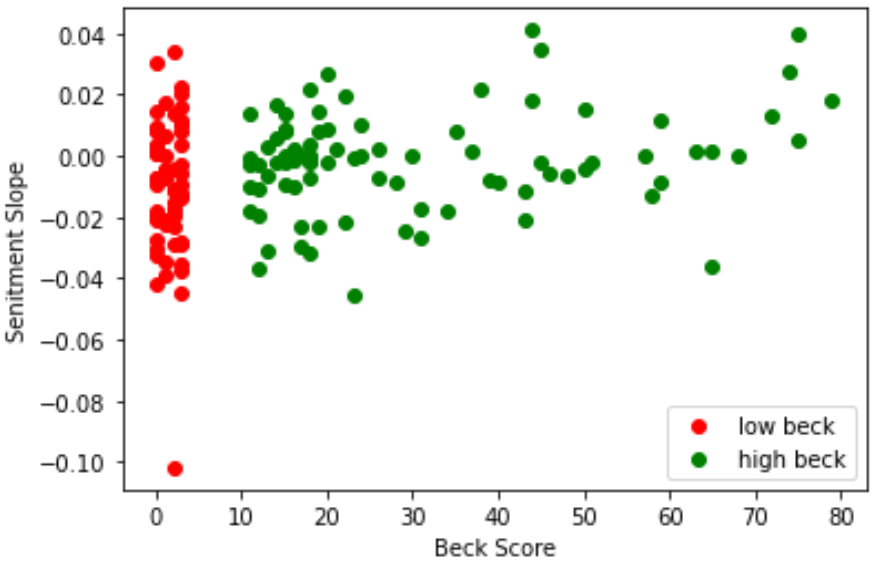

Figure 9: Sentiment Slope vs. Beck Score

Inspired by the results found in the previous section, we wondered whether there was some relationship between the sentiment progression through successive possibility generations (for simplicity, let us refer to this trend as the *sentiment gradient*) and individual differences like the beck score. What we found was that there was a significant relationship between the sentiment gradient and beck score (depicted in *Figure 6 and Figure 7)*; the exact direction of this effect corresponds with the hypothesis described in the previous section. Individuals with higher beck scores tended to have more positive sentiment gradients (that is, their later generations possessed more positive sentiment than their earlier generations) than their lower beck score

counterparts. Conversely, individuals with lower beck scores tended to have negative sentiment gradients – confirming the hypothesis sketched above.

Variation vs. Beck:

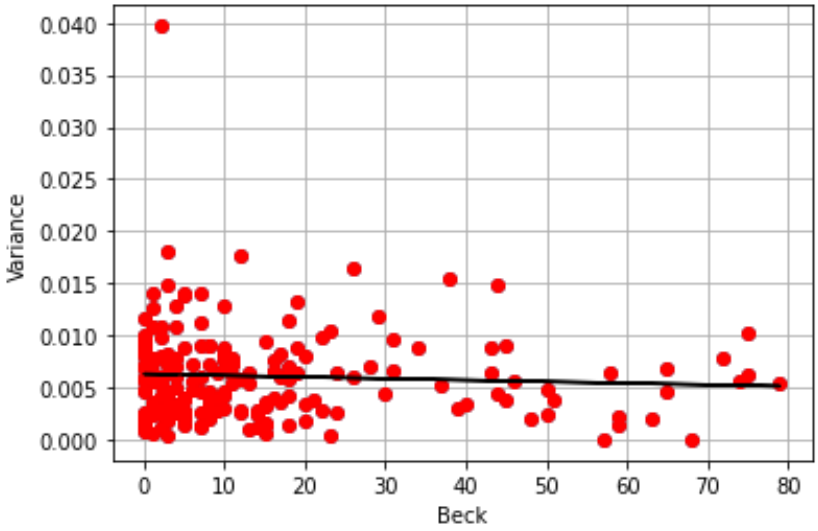

Figure 10

R-value: -0.061798
P-value: 0.389535

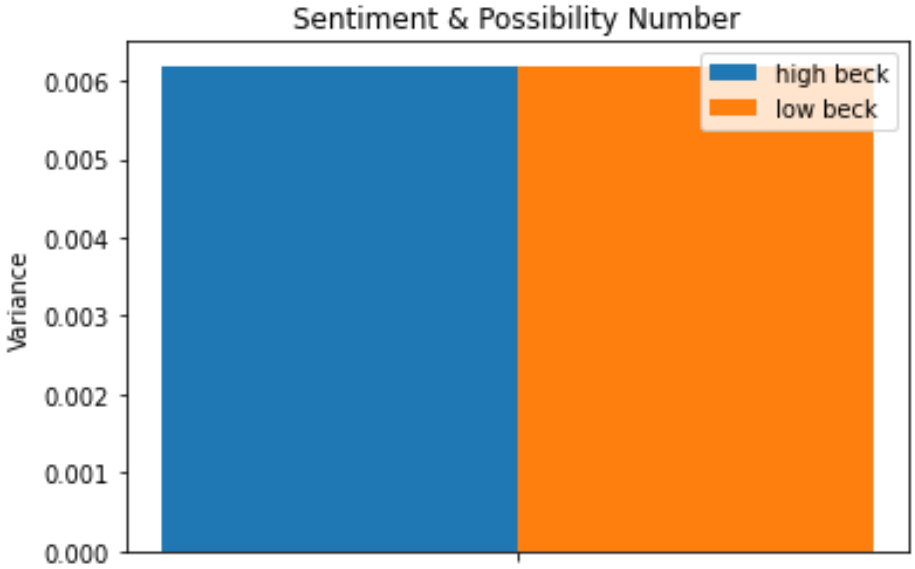

Figure 11

Variance	*Group Size*
0.00619	57
0.00617	117

A natural question that arose when examining *Figure 4* was whether more significant sentiment variations would be observed in groups with higher beck than those with lower beck. Here, we

take variation to mean the *variance* of the average sentiment per possibility number. Regretfully, there did not seem to be any significant correspondence.

Semantic Distance vs. Beck:

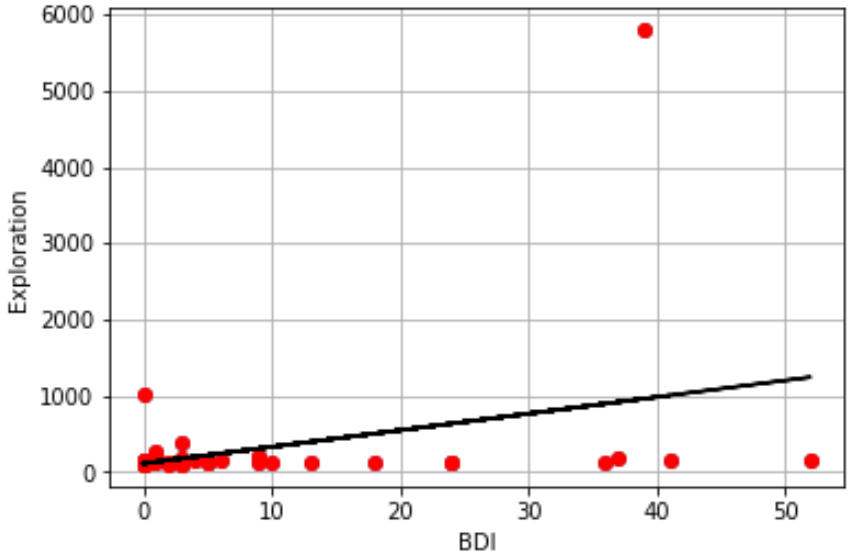

Figure 8: Semantic Distance vs Beck (N = 40)

R-value: 0.332222
P-value: 0.036222

At this point, we had an interest in yet another question: could the extent to which individuals explore semantic space vary according to individual differences in depression and anxiety. The intuition here is this: individuals who are more depressed may not only generate possibilities ranking lower in terms of sentiment, but they may also explore possibilities positioned closer to each other in semantic space. A more formal definition of "semantic distance" may be instructive here. Recall that each participant generates six possibilities per vignette and ten vignettes in total. Each of these possibility generations receives a corresponding word embedding vector. As such, semantic distance for each individual per vignette is defined as follows:

$$Semantic\ Distance_v = \sum_{i=1}^{6}\sum_{j=i}^{6} distance(embedding_i, embedding_j).$$

The notion of distance utilized here is standard Euclidian distance, or, more formally:

$$distance(embedding_i, embedding_j) = \sqrt[2]{\sum_{k=1}^{n}(embedding_{i,k} - embedding_{j,k})^2}.$$

Finally, to determine the vignette general semantic distance per participant, we simply average the semantic distance across all vignettes.

With the notion defined, *Figure 8* demonstrates the result of this experiment. Restricting the size of our sample to merely 40 participants out of the selection of nearly 200, we see statistically significant results. However, due to current computational limitations (generating large numbers

of word embeddings is computationally intensive), it is challenging to derive results for the entire sample. Though it should be flagged that by increasing the sample size to 80 individuals, the results change substantially.

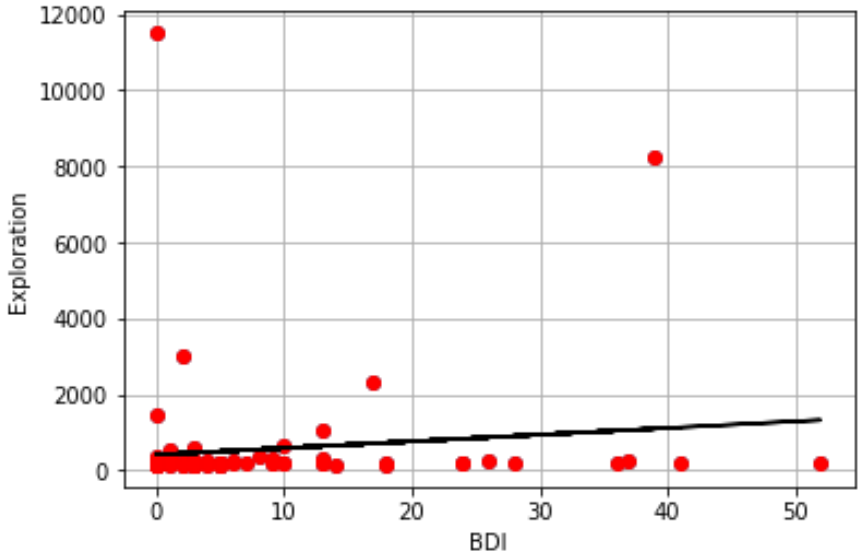

Figure 9: Semantic Distance vs Beck (N = 80)

R-value: 0.121328
P-value: 0.283686

Though by removing the point at y = 11,000, we see the results change once more:

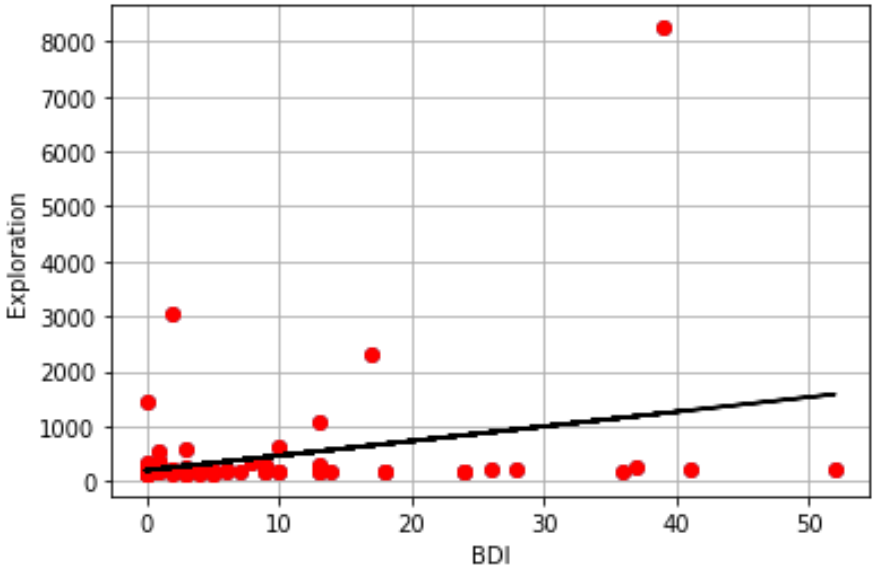

Figure 10: Semantic Distance vs Beck (N = 80; Outliers removed)

R-value: 0.294054
P-value: 0.008529

Now, though these results seem significant, caution should be exercised against taking them as such. The relationship described makes little sense. The line slope expressed in figures 8, 9, and 10 indicate that individuals who exhibit far stronger depressive tendencies explore semantic

space more – not less – than those who do not. These results run counter to our initial hypothesis, which we have strong reason to believe would be on the right track.

Trajectory Clustering, Reduced Dimensionality, and Prediction Analysis:

The bulk of the tests above sought to establish linear correlations between two variables in a manner legible to humans. For example, we investigated the relationship between features like semantic distance and beck, sentiment gradient and beck, and so on. In this section, the testing paradigm will take a different form: relationships between features legible to an algorithm, but not necessarily a human, will be explored.

Trajectory Clustering:

As mentioned above, each possibility generation corresponds to a particular word embedding that takes the form of a *[1-row x 768-columns]* vector. This implies that each vignette (and the possibility generations within it) takes the form of a *[1-row x 4608 -(6*768)-columns]* vector. This vector might be thought about as a trajectory through semantic space. A trajectory in lower dimensions can be conceived of as a collection of points, and correspondingly, a trajectory through higher dimensions can be conceived of as a list of features. For any vignette, there will thus be 196 vectors (and thus trajectories through semantic space) corresponding to each of the 196 participants. Given these trajectories, we wondered the following: if we were to run a clustering analysis to generate two distinct clusters, would the mean beck scores per cluster differ? This is an interesting question because these word embeddings are generated based on semantic content *independent* of the beck score. Unless there is some implicit relationship between the word embeddings (which act as a numeric representation of semantic properties) and the beck score – which is not, to my knowledge, the case -- we would not expect this to be the case. But if there is some relationship between trajectories through semantic space and beck scores, this would be highly intriguing!

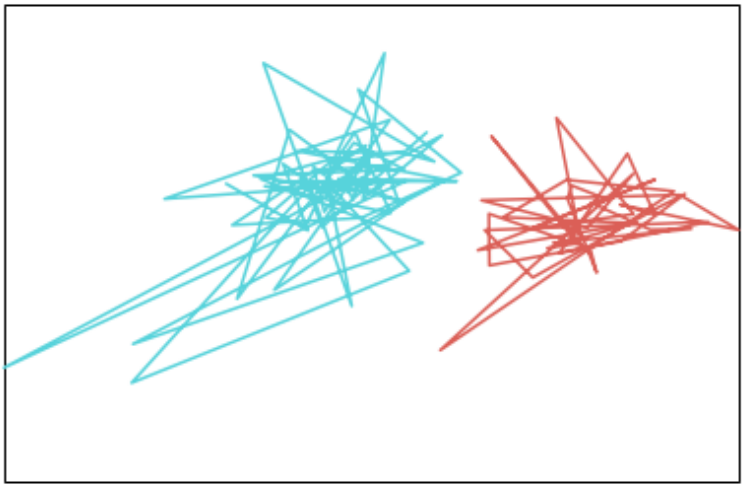

Figure 11: Clustering Analysis of Embeddings Represented in 2D[2]

[2] Notice how cleanly two clusters form when these embeddings are reduced to 2D. It is almost as if a line could be drawn through the figure to separate both clusters. This feature (a sharp divide between clusters) makes these data extremely receptive towards classification attempts using SVM.

	Blue(n=)	Red(n=)
Beck	15.08	12.88

Figure 12: Beck by Cluster

Figure 11 shows the results of the analysis described above. The lines on the graph represent 2-dimensional depictions of each participants' trajectories through higher-dimensional semantic space (for a specific vignette). The dimensionality reduction was conducted using a PCA technique. Note that two distinct clusters emerge naturally on this graph (colored red and blue). Moreover, observing the average beck score for participants belonging to each group, we notice a substantial difference. The problem, however, is that it is unclear precisely what explains this difference. Clustering analysis on purely numerical feature embeddings, while interesting, leaves much to be desired by way of human inference. This analysis seems promising for further work.

Reduced Dimensionality:

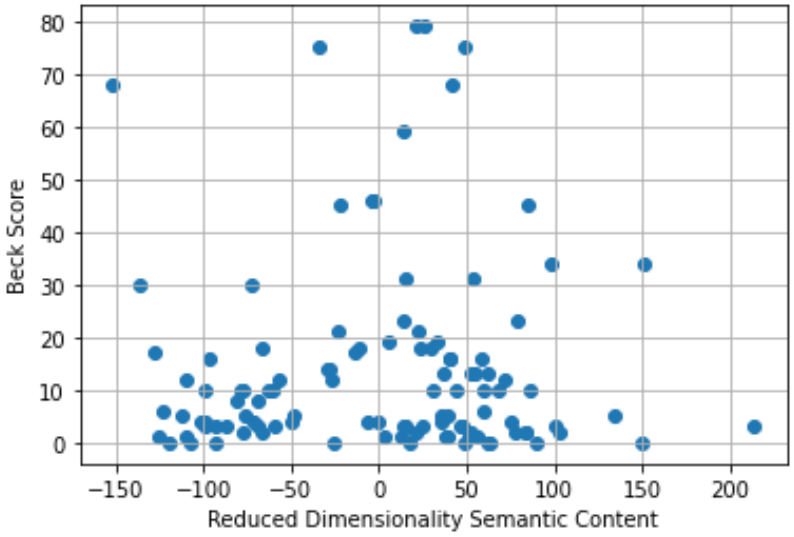

Figure 13: Beck vs. Reduced Dimensionality Semantic Trajectories

R-value: 0.022159
P-value: 0.819098

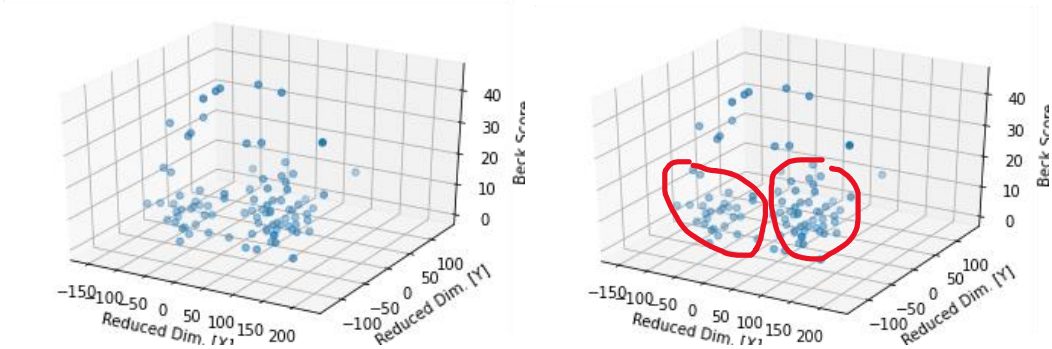

Figure 14: Beck vs. Reduced Dimensionality Semantic Trajectories

Like the trajectory clustering analysis, we wondered whether reducing the dimensionality of the trajectories and then plotting these data against individual beck scores would generate results amenable towards human understanding. Here, the results are mixed. *Figure 13* plots beck scores against trajectories reduced to 1 dimension, and the results don't seem particularly insightful. However, *Figure 14* plots a similar phenomenon with semantic trajectories reduced to 2D instead of 1D. Here, we can see the vague outline of two clusters forming, which further grounds the results seen in *Figure 11.* Note that the locations of these clusters vary along the 'Beck Score' axis.

Machine Learning:

Diverging from much of the above analysis, this section summarizes current attempts to incorporate machine learning into this project. Though these algorithms are varied, their aim is shared: given semantic trajectories for every participant as well as known beck scores, can a binary classifier be trained to determine whether a particular individual is 'depressed' (beck > upper threshold) or 'not depressed' (beck < lower threshold), where these thresholds may vary from case to case. Models that have been trained so far have used logistic regression and SVM. Both are report F1 scores around .8 through 5-fold cross-validation.