**CSSM502 Final Project Report**

My project for this class is a machine learning model aimed at analysing text data for detection of sources of meaning (SoMe) in life. There are 5 sources of meaning identified in the psychology literature. These sources are defined as the primary domains of meaning making, finding purpose or ways that enables the individual to connect with life. The 5 categories are identified as Horizontal Self Transcendence, Vertical Self Transcendence, Order, Self-Actualization, and Wellbeing/Relatedness. These five are the main domains of meaning making that were identified in the literature and validated multiple times. I use this categorization as my theoretical framework and the categorization criteria of my model.

My starting point and the main motivation for this project was that I think pyschological constructs can be studied with computational methods. More specifically open-ended text data can be analysed with natural language processing in a way that psychological inference can be made from these sources. Given that I could not collect my own data for this project, my immediate potential data sources were either qualitative psychology studies who have shared their interviews, or social media data. However the first is traditionally collected as a very case study-like approach where sample sizes thus the amount of text data was small. Additionally, these studies often do not share their data with respect to the privacy of the participants, and the studies I could find were scarce in the way that training an ML model with this data would require a combination of data from a multitude of studies. Combination was also not viable as the focus of these studies were of high diversity, hence not compatible with each other for a holistic study. The latter option, social media data, is in very large quantities and is definitely possible to train an ML model with. However, both twitter and reddit data, even coming from subreddits related to meaning making, was not in the depth required for a proper analysis of the psychological construct I wanted to study. As both of these options were shown to be not viable, I came up with an alternative which I think wasn't done before in Computational Social Sciences, which was using historical diaries.

A large quantity of historical figures had held extensive journals spanning for years if not decades. When segmented by paragraphs or sentences, the amount of data far exceeds the requirements of training an ML model. The text in these journals also possess high paychological

richness as they are intimate works of their authors. To a large extent, we can also justify the assumption that the content is authentic and free from observation bias, as these journals are often written without the purpose of publishing, and were often published after the death of their authors. I wanted to prevent any potential intervention from translation bias, so I searched for the largest journal collections that are written in English originally. Then I found Henry David Thoreau's journals which spanned across about 15 volumes, ranging from 1838 to 1861. Although he wrote on and off, his overall consistency of journaling across this timeline was very high, especially during the 1850s. He also dated his journal entries with the same consistency. This allowed for a unique N=1 computational case study that could follow later on with dynamic network analysis.

I first collected the entire collection of Thoreau's journals in pdf format. I used an OCR library to scrape the text content from the pdfs. I tried a few different versions of the journals and two different OCR libraries to find the ideal combination with the highest reading accuracy. The page structure created a problem. To fit such a large collection into as many little pages as possible the publisher fitted 4 actual journal pages into each page, and the reading order of the OCR. Thankfully the ordering errors followed a consistent pattern, so during reading I was able to reorder the pages. After I scraped the entire data successfully, I now had the raw text and I started data cleaning and segmenting.

First of all, I cleaned the page meta data as much as possible, like the page titles, page numbers, header and footer notes, and editorial additions given within the text between brackets. Then I had to segment the data into ML compatible sized chunks. I first detected the dates noted at the beginning of each entry, the format was fairly consistent (e.g.Jan 24), but there were exclusions, such as the month of March and April was never abbreviated, or the entries closer to Thoreau's death were more like to be written in full (e.g. January instead of Jan). Moreover, he sometimes also mentioned dates within the entries themselves (like "It had snowed a lot last Dec 18"). The dates were also only included as months and days, the years were not noted. I tried iterations that increment a year counter every time when the next date followed was a previous month in the year (e.g. Sept to March indicated a new year), but this resulted in counting errors when the author mentioned a date from a previous month. In the end, I used different regex

structures that aimed to only include actual entry dates (with identifiers like page breaks or dots) and infer the year semi-manually, using the page header data and human checking.

I wanted to use the SetFit method for this model, so I kept segmenting the data based on the requirements of the method. I set the minimum and maximum length limit of each data point as 7 and 250. I segmented the entries further within this range, while respecting sentence structure. I also cleaned the numbers appearing in the text, since the frequency of numbers appearing in the text was low and highly variable, and I thought that they would harm the training process. Most of the cleaning was done successfully at this stage, but I still went through the data manually to clean unintelligible rows.

I trained a SetFit model to avoid extensive data labeling while maintaining accuracy. SetFit learns by comparison of different categories, so while it learns which category includes what, it also learns what it should not include. This allows accurate models with limited amounts of data labeling. I created a gold standard test set and an initial training set of 50 data points. I labeled these data points and trained an initial model, which then labeled the rest of the data itself. Among the ones it labeled, I selected a second training set based on the data points with highest uncertainty,l. The model attributed a probability score to each category for each data point, and I took the 50 data points that are scores closest to 0.50 probability for their categories. I manually labeled this second set of data, and included it into the training set and trained another model. The idea was to repeat this process a few times until the latest model is as capable as possible. However, as I was manually labeling the data and checking the ones labeled by the model, I realized a problem. One of the categories, specifically horizontal self transcendence, was highly dominant in the dataset (probably around 70% of all data points were purely HST), as Thoreau mostly found subjective meaning in unison with nature and natural phenomena. The other four categories were very underrepresented, and one of them (Vertical Self Transcendence) did not even have an example in the initial training set, so I had to add a keyword search to my script to find potential data points related to VTS. In the end I decided that the model is not reliable as is within the context of my aim to build an ML model that can extract sources of meaning from text. The model could only perform well for one of the categories, and that probably involved a large amount of false positives. I also added a sentiment analysis column to

note emotional valence with an established model but I was not able to complete that part for now.

I learned a lot from this project. I have a solid idea for a novel useful ML model, I have learned a lot about data cleaning and preparation as I worked with the most noisy data I have ever worked with, and I learned the fundamentals for training a machine learning model for text analysis. I will continue this project beyond this course. The next steps I have planned is to train the initial model as balanced and as grounded as possible. SoMe has a 161-item scale that is consistently used in literature. I plan to use these items as the core of my model, and train the initial model based on them. Second, I will collect the journals from a variety of persons and not one with more representation for each category. I think 1 person per category would be good enough, but my target is 2 people per category to avoid overfitting to specific people. Based on the labeling with the initial model, I will take the strongest and weakest labeled data points and add them to the training set. I will continue this iterative approach to keep minimizing labeling costs as one person and maximize the accuracy. Even though one person probably will represent one category mainly, there will be data points from all of them in each category, which I hope will allow a cross pollination effect that leads me to a much more balanced and accurate model. Lastly, I will use this model on 1-2 other authors' journals that were not included in my training set as my ground truth. The ultimate aim is to create a universal ML model for SoMe classification with text data. I can then again use this model on timestamped journal data for N=1 case studies of meaning making.