The Mathematics of Clinical Diagnosis:
Cognitively-Inspired Computational Psychiatry

by

Matthew Cook

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
In partial fulfillment degree of

MASTER OF ARTS

Department of Psychology
University of Manitoba
Winnipeg

Abstract

Medicine has been advanced by diagnostic technologies such as artificial neural networks that can diagnose cancer from radiological images. However, these diagnostic technologies have mainly been applied to the diagnosis of physical illness, even though mental illness is as relevant as a diagnostic problem in our everyday lives as physical illness. Diagnostic technologies have not been applied to the diagnosis of mental illness for good reason. Psychological diagnosis does not depend on an analysis of physical symptoms. Rather, psychological diagnosis depends on interpreting and understanding the language people use to describe their thoughts and emotions. However, language is a complex and imprecise presentation of mental health. To solve these problems, I evaluated established models of distributed semantics and machine learning classification models to build a computational system that can diagnose people's mental health from their written language. The system was trained and tested on database of essays written by 1016 participants who also completed five standard measures of mental health. The work joins a growing effort to translate basic cognitive psychology and computational psychology research into the design of cognitive technologies capable of solving complex real-world problems.

Keywords: categorization, clinical diagnosis, LSA, machine learning, computational psychiatry

The Mathematics of Clinical Diagnosis: Cognitively-Inspired Computational Psychiatry

Categorization (or classification) is a basic psychological process where individual exemplars are mentally organized according to shared properties (Murphy, 2002). This process encodes a psychological structure of the natural and artificial worlds, and in the context of language, eases communication. Rather than having to think about, remember, and discuss an endless number of unique exemplars, exemplars can be conceptualized and discussed according to their similarity on relevant properties or dimensions.

Classification has been studied extensively by cognitive psychologists. Instance theories (e.g., Brooks, 1978, 1987; Hintzman, 1984; Medin & Schaffer, 1978; Nosofky, 1986) were developed in the 1970s and 80s as a description of classification behaviour. At the time, there were two dominant theories of classification. The classic view (beginning with Aristotle, trans. 2001) proposed category membership is determined by the presence versus absence of defining features. Alternatively, prototype theory proposed category membership is determined by similarity to a category prototype or average category member (Rosch, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). In contrast, instance theory demonstrated that categorization emerges from a process of storing instances (or exemplars) to memory with categorization determined by similarity to specific known category members.

In the 1990s, psychologists studied and examined clinical diagnosis via instance-based similarity. To diagnose disease, a diagnostician must translate groups of symptoms (or features) into diagnosed categories. The task is complicated for many reasons: Symptoms can present themselves differently in different people and a single symptom can point to multiple diseases. Although instance theory was developed in relation to artificial categories in the 1970s, it was not until the work in the 1990s that researchers began to show that the instance perspective

described the applied problem of diagnostic decision making. Work by Brooks (e.g., Norman, Young, & Brooks, 2007) demonstrated that the reasoning methods (and success) of novice versus expert diagnosticians differs. Whereas novice diagnosticians use analytic reasoning methods consistent with Aristotle's classic view of defining features, expert diagnosticians rely on experience-based, non-analytic reasoning methods to diagnose physical and mental disorders. Rather than making diagnostic decisions based solely on a list of diagnostic criteria, expert diagnosticians also classify novel cases by comparing them against a history of relevant experience. Contrary to initial expectations, this classification process has shown to be fast, accurate, and inaccessible to introspection.

Today, we have technologies to augment many of our cognitive processes. We have prescription lenses to augment visual perception and sticky notes to augment memory (Clark & Chalmers, 1998). Importantly, in the domain of medicine, we have diagnostic technologies such as artificial neural networks to diagnose cancer from radiological images that augment our classification processes (e.g., Lee et al. 2017). These technologies are important because they reduce the need of a highly trained medical professionals, thus allowing for automated and more accurate medical diagnosis. Furthermore, these systems can also detect subtle patterns that may go unnoticed by the human eye, even when that eye has decades of experience.

Though medicine has benefited from these diagnostic technologies, they have mainly been used for detecting physical illness. However, mental illness is as relevant a diagnostic problem in our everyday lives as physical illness. I surmise that the reason why these technologies have not been applied to the detection of mental illness is because psychological diagnosis does not depend on an analysis of physical symptoms. Rather, psychological diagnosis

depends on interpreting and understanding the language people use to describe their thoughts and feelings. Our words are the window into our mental lives and our mental health.

Since the beginning of psychology, psychologists have sought to understand the relationship between the language people use and the workings and content of their mental lives. Freud (1901) described slips of the tongue (dubbed *Freudian slips*) where a person's latent emotions, thoughts, and intentions would expose themselves in mistakes in their use of language. Rorschach (1921) developed his famous inkblot test where people reveal the contents of their mental life through their verbal description of ambiguous paint blots. McClelland (1979) developed the Thematic Apperception Test (TAT) where people tell stories about ambiguous drawings or photos of people, thus revealing their motives and understanding of social interactions. Most generally, psychologists and psychiatrists, regardless of their theoretical orientation (e.g., cognitive-behavioural, psychodynamic, humanistic) depend on the clinical interview to elicit language from their patients.

However, language is one of the most complex processes in our cognitive toolbox. Due to the generative and imprecise nature of language, thoughts can be expressed in a great number of ways. Consequently, it is a challenging job for clinicians to map the particular expression of language people use to the truth of their mental lives. For these same reasons, in the domain of developing automated diagnostic technologies, it is a challenging job to build machines that can classify peoples' mental health from their language. Unlike an x-ray image that can be pixelated into an array of numeric pixel intensities, it is not immediately clear how to numerically represent language that people use in a way that captures meaning. And yet, to build automated methods for clinical diagnosis, we must represent the language people use in a numeric and machine-readable fashion.

The goal of developing automated methods for psychological diagnosis can be traced back to the 1950s with Paul Meehl (1954) who wrote about automated methods of clinical diagnosis. Though computers were not readily accessible at the time, Meehl's work showed how his "clerk" could enter easily obtained values into a multiple regression formula and produce as, or more accurate, diagnoses than trained clinicians. Meehl's methods for automated diagnostic presented a formal and quantitative method of predicting people's mental health. However, his methods did not analyze the language people used to arrive at a diagnostic decision.

In the 1980s, Walter Weintraub (1981, 1989) analyzed the relationship between written text and mental health. Weintraub hand counted words and parts of speech (i.e., first-person singular pronouns) from text that people wrote and found that these counts were reliably correlated with peoples' levels of depression. Recognizing the importance of Weintraub's work, but seeking to automate the counting process, Pennebaker (e.g., Tausczik & Pennebaker, 2010) developed software for counting individual words and word classes. Pennebaker's software LIWC (Linguistic Inquiry and Word Count) automatically compares each word in text to a dictionary where it counts and summarizes counts in more than 80 linguistic categories. As with Weintraub's work, text analysis with LIWC shows certain linguistic categories are reliably correlated with people's mental health, especially first-person singular pronouns such as *I, me,* and *my* (Tausczik & Pennebaker, 2010; Willits, Rubin, Jones, Minor, & Lysaker, in press).

LIWC presented a big leap forward in automated methods of clinical diagnosis from written text. However, the method has several shortcomings. Whereas many of the LIWC categories are unambiguous and objective, such as the category for function words like *the, it,* and *to,* other categories are more subjective, such as self-reflective thinking. These subjective categories require a great amount of human expertise and effort to construct, given that each

word in the dictionary must be agreed upon by researchers. More seriously however, other than a crude notion of semantics based on pre-defined categories, LIWC fails to encode meaning. LIWC has no has no way to appreciate that words like *spaceship* and *galaxy*, or *worthless* and *depressed* are related unless these relationships are encoded by building word categories that include both words.

For the last 60 years, researchers have worked to derive a formal theory to represent word meaning. In 1952, Osgood had people rate words on varying dimensions, such as *good-evil* or *valuable-worthless*. Each word then occupied a point in a space with the axes representing the dimensions on which participants rated the words. This work, known as the semantic differential, resulted in a geometric space of word meaning, where semantically similar words occupied similar regions of space. The work provided researchers with a formal mathematics of meaning. The problem was scalability: Many participants and a great deal of time were required to build the semantic space of many hundreds or thousands of words. What researchers yearned for were automated methods of building a geometric vector-representation of semantics.

Recent advances in models of semantics developed by cognitive psychologists provide a formal method to solve the problem of word meaning. Starting in the 1990's researchers used machine learning methods grounded and informed by psychological theory to build a numeric representation of word meaning. These models read a large body of text and learn from patterns present in language to derive semantic vector representations of words. The models operate by relying on the distributional hypothesis, the notion that words that are used in similar contexts have similar meanings. As Firth (1957) said, "You shall know a word by the company it keeps". Some of the most well-known models of distributional semantics are Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), Bound Encoding of the Aggregate Language Environment

(BEAGLE; Jones & Mewhort, 2007), Hyperspace Analogue to Language (HAL; Burgess & Lund, 1997), and Google's word2vec (Mikolov, Chen, Corrado, & Dean, 2013). Though these models all differ in their theoretical foundations and implementations, they all derive a numerical representation of words by reading a large body of natural language. Of all these models, the most widely known is Latent Semantic Analysis (LSA; Landauer & Dumais, 1997).

Latent Semantic Analysis (LSA)

LSA is a statistical method for modelling the meaning of language. LSA works by first "reading" a large body of text, called a corpus. The corpus can be any body of text, such as newspaper articles, textbooks, blog posts, or webpages; the corpus provides the knowledge base of LSA in the same way that peoples' experience inform their knowledge base. Formally, the corpus is rewritten as a word-by-document matrix, where each row of the matrix codes each unique word in the corpus and each column codes each document in the corpus. A document in the context of LSA can be any unit of meaning, such as a sentence, paragraph, book, or web page. Thus, cells of the matrix record the frequency of each word in each document. Because most words do not occur in most documents, most of the cell entries in the matrix will be 0, in what is called a sparse matrix.

Of course, some words (e.g., function words like *the*, *and*, *to*) appear in every document and thus provide very little or no information about the meaning of a document. In contrast, other words provide useful clues about the meaning of a document (e.g., content words like *statistics*, *worthless*, *spaceship*). It is for that reason that the cell frequencies in the word-by-document matrix are transformed to their information values. In particular, the inventors of LSA (Landauer & Dumais, 1997) use entropy weighting (described in Martin & Berry, 2005) of the matrix frequencies based on Shannon's (1949) measure of information (i.e., entropy),

$$\text{entropy} = 1 + \frac{\sum p_{ij} log_2(p_{ij})}{log_2 n}$$

where p_{ij} is the number of times word *i* occurs in document *j* divided by the total number of times word *i* appears in the corpus, and *n* is the number of documents in the corpus. This weighting is applied to all non-zero entries of the word-by-document matrix. This procedure has the result of assigning a small entropy to common words that appear in every document and therefore do not reveal much information about the meaning of the document while also assigning a large entropy to words that appear in fewer documents and therefore discriminate between documents.

After the word-by-document matrix is formed and a weighting scheme is applied, a technique from linear algebra called Singular Value Decomposition (SVD), is applied to the matrix (e.g., Strang, 1998). SVD is analogous to factor analysis and is identical to Principal Component's Analysis when the column vectors of the word-by-document matrix are mean centered. In the same way that a composite number, such as 30, can be decomposed into fundamental prime factors (e.g., 30 = 2 x 3 x 5), a matrix can be decomposed using SVD into three fundamental matrices:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathrm{V}^{\mathrm{T}},$$

where **A** is the original matrix, **U** is a matrix of the eigenvectors of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, **Σ** is a diagonal matrix containing the square root of the eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ (termed singular values), and **V** is a matrix of the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ (where superscript T indicates the matrix is transposed). The eigenvalues contained in **Σ** are ordered from largest to smallest, and the eigenvectors of **U** and **V** are arranged to match the order of their corresponding eigenvalues.

The purpose of applying SVD to the matrix is to reduce the noise in the matrix by leveraging the statistical regularities in language that manifest themselves in the matrix. This is accomplished by producing the least squares best approximation of the original matrix by using a

reduced number of dimensions to reconstruct the matrix. The number of dimensions to retain depends on the task the researcher is faced with and is usually determined by selecting the solution that provides the best fit to the data. SVD and related techniques are called dimension reduction techniques.

Dimension reduction is accomplished by reconstructing the original matrix with a smaller number of dimensions, usually denoted *r* (for rank):

$$\mathbf{X} = \mathbf{U}_r\mathbf{\Sigma}_r ,$$

where **X** is the least squares best approximation of the original matrix. $\mathbf{U}_r$ is a matrix including the first *r* columns of the **U** matrix, and $\mathbf{\Sigma}_r$ is a matrix including the first *r* rows and *r* columns from the $\mathbf{\Sigma}$ matrix.

The number of dimensions used to reconstruct the matrix determines the proportion of variance retained in the least squares best approximation of the original matrix. When the matrix is reconstructed with the full number of dimensions, 100% of the variance in retained. The fewer the number of dimensions that are used to reconstruct the matrix, the smaller the proportion of variance that is retained in the reconstructed matrix.

Due to the regularities in how people use language, a large proportion of variance will be contained in a relatively small number of dimensions (e.g., 300) even if the matrix originally consisted of tens of thousands of columns (i.e., documents). Statistical regularities occur in the word-by-document matrix because language is largely redundant and therefore predictable (e.g., Shannon & Weaver, 1949). Our knowledge that a word like *spaceship* is more likely to occur in a document that contains the word *galaxy* is intuitive, but also statistical. SVD exploits and leverages these statistical regularities.

The result of applying SVD to the word-by-document matrix is a matrix with the same number of rows as the original word-by-document matrix (each row still represents a word), but a smaller number of columns (by tradition around 300 dimensions, which was found to optimize performance of the model in a synonym task; Landauer & Dumais, 1997). Each row of the matrix is called a semantic word vector, because the row vectors of the matrix contain meaning. The effect of SVD is that the word vectors for semantically similar words will be made more similar, whereas semantically dissimilar words are made less similar, relative to the corresponding vectors before SVD is applied (Appendix B presents a toy example to demonstrate the relationship between words before and after applying SVD).

LSA belongs to a larger family of methods called *vector space models* (for a review see Jones, Willits, & Dennis, 2015; Turney & Pantel, 2010). The name comes from the fact that geometrically, these words occupy an *n* dimensional space, where *n* is the number of dimensions retained using SVD. In the space, semantically similar words (e.g., *dog* and *wolf*) are close together, whereas semantically dissimilar words (e.g., *dog* and *toolbox*) are farther apart. Figure 1 shows a visual example of a semantic space in two dimensions.

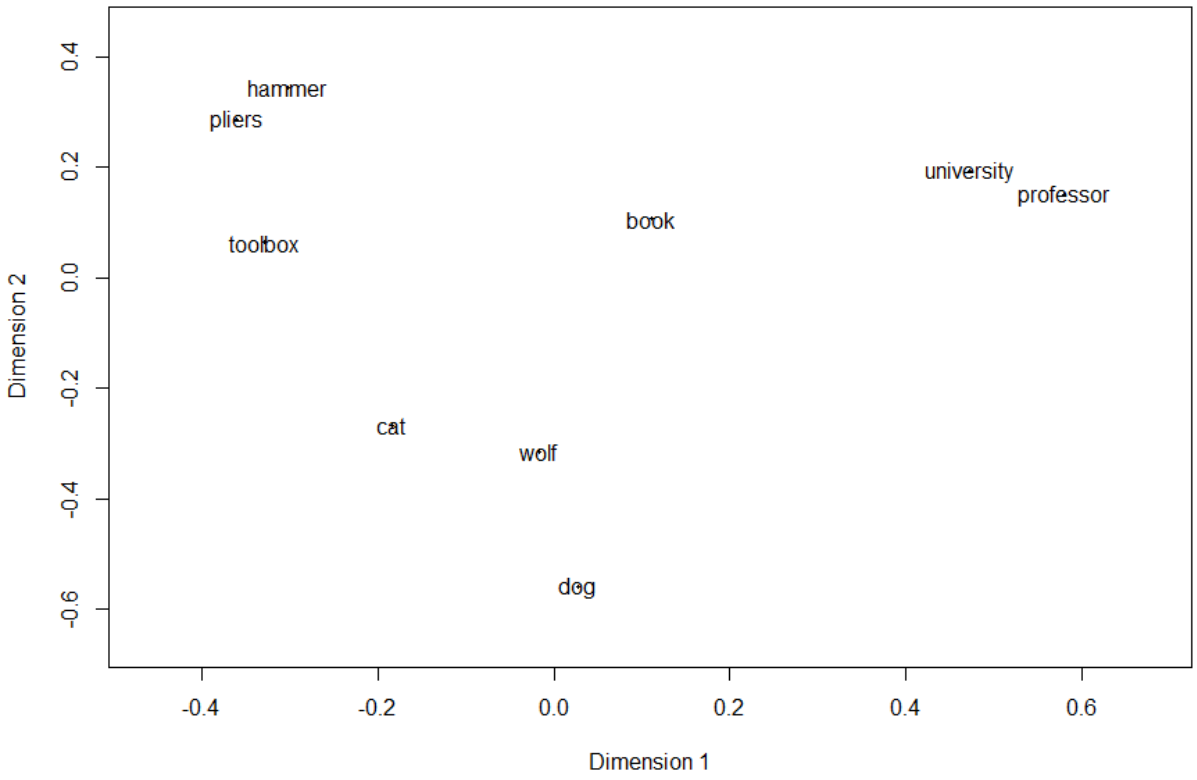

Figure 1. Multidimensional scaling solution of nine words from LSA derived vector space plotted in two (as opposed to 300) dimensions. In the space, semantically similar words (e.g., dog and wolf) are close together, whereas semantically dissimilar words (e.g., dog and toolbox) are farther apart.

The similarity between two words can be calculated by any number of standard methods. One commonly used method is Euclidean distance which measures the straight-line distance between any two vectors (points in an *n* dimensional space),

$$d = \sqrt{\sum_{j=1}^{n} (\mathbf{a}_j - \mathbf{b}_j)^2},$$

where *d* is the Euclidean distance, and **a** and **b** are vectors of the same dimensionality (*n*). Vectors that occupy the exact same location will have a Euclidean distance of 0. The further two points are in space, the greater their Euclidean distance.

But, because Euclidean distance measures the distance between vectors, and not their similarity per se, researchers will often use cosine similarity, which only considers the direction of the vectors and not their magnitude.

Cosine Similarity (or normalized dot-product) measures the angle between vectors **a** and **b** ignoring the length of the vectors,

$$cosine = \frac{\sum_{j=1}^{n} \mathbf{a}_j \mathbf{b}_j}{\sqrt{\sum_{j=1}^{n} \mathbf{a}_j^2} \sqrt{\sum_{j=1}^{n} \mathbf{b}_j^2}},$$

where **a** and **b** are vectors of the same dimensionality *n*. Cosine similarity can range from -1 to 1 representing dissimilarity and similarity, respectively, with a cosine similarity of 0 indicating orthogonality between vectors.

Vector space models such as LSA have been enormously successful in a range of applications. These successes range from modelling children's semantic memory (Denhière, Lemaire, Bellissens, & Jhean, 2008), modelling basic memory processes (Howard, Addis, Jing, Kahana, 2007), assessing personality (Kwantes, Derbentseva, Lam, Vartanian, & Marmurek, 2016), assessing reading skills (Magliano, & Millis, 2003), assessing and improving text comprehension (Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 2007), automatic essay grading (Landauer, Laham, & Foltz, 2003), and building semantic search engines of academic journals (Aujla, Jamieson, & Cook, in press).

Recently, researchers have used semantic vectors to predict peoples' mental health. For example, Johns et al. (in press) predicted semantic decline in mild cognitive impairment using BEAGLE word representations. The researchers had participants complete a verbal fluency test, in which the participants' goal was to name items from a category. By analyzing the path of responses, they could predict which participants would go on to develop mild cognitive

impairment. Bedi et al. (2015) used a similar method, but instead used LSA word vectors to predict which participants would develop schizophrenia.

As previously argued, classification models have played an important role in medical diagnosis. But, these technologies have mainly been applied to the diagnosis of physical illness. To extend methods of automated diagnosis to mental illness, we need to build smart machines that can interpret and understand language we use to describe our thoughts and emotions. In the remainder of this thesis, I will investigate and evaluate the ability of models of distributed semantics like LSA to classify people's mental health from written language.

Current Project

To carry out the project, I have used established psychological vector space models and machine learning classification methods. In particular, I collected data on peoples' mental health and used classification algorithms to predict their psychological distress, depression, anxiety, and positive affect from free-form written reports. The task of classifying text for this project had four major aspects, including: (a) building a semantic representation of words, (b) building a semantic representation of essays, (c) choosing and training the classification models, and (d) evaluating the performance of the models. But before I begin, let me describe the data I collected for the project.

Method

Collection of Essay Data

To conduct this project, I required both written reports that describe how people self-describe their mental health and an objective standard measurement of their mental health. 1016 University of Manitoba undergraduates recruited form the undergraduate psychology participant pool completed the study. Each participant was asked to write a minimum of 200 words to four

free-form questions meant to assess mental health: (1) Please describe your mental health in as much detail as possible, (2) Please describe the role of depression in your day-to-day life (i.e., do you think you experience depression, why or why not?), (3) Please describe the role of anxiety in your day-to-day life (i.e., do you think you experience anxiety, why or why not?), and (4) Please describe the positive emotions you experience in your day-to-day life.

After writing the essays, participants completed a brief mental health questionnaire: the Kessler K10 Psychological Distress Scale (described in detail in Kessler et al., 2003; Kessler 2010) that assesses psychological distress (depression and anxiety). The K10 scale is a reliable and valid scale that corresponds with clinical diagnoses based on diagnostic criteria from the Diagnostic and Statistical Manual of Mental Disorder (DSM-IV; Andrews & Slade, 2001). The K10 scale contains 10 items (see Appendix A) aimed at measuring psychological distress characterized by depression and anxiety. The 10 questions in the K10 measure the frequency of experiencing psychologically distressing symptoms (e.g., worthlessness, anxiety) with a 5-point Likert-type scale where 1 = *None of the time,* 2 = *A little of the time,* 3 = *Some of the time,* 4 = *Most of the time,* and 5 = *All of the time*. The test gives a single summative score ranging from 10 to 50, with higher scores indicating greater psychological distress. The scale includes *cut scores* that categorize summative scores into four groups of psychological distress (Kessler et al., 2010; Andrews & Slade, 2001). Scores < 20 are considered “Well”, scores between 20 and 24 are considered “Mildly Distressed”, scores between 25 and 29 are considered “Moderately Distressed”, and scores >= 30 are considered “Severely Distressed”. Based on participants responses to the ten questions, they are categorized into one of the four ordinal levels of psychological distress.

Data were collected online using Qualtrics, a website for survey development, management, and deployment. Each of the four free-form written questions was presented on an individual page, with a textbox for writing essays and a word counter to present the number of words participants had typed. After writing the essays, participants responded to the K10 and four other mental health measures (described later). Each of these five scales was presented on individual pages.

Modelling

Word and essay representation. The essays that participants wrote about their mental health needed to be re-expressed in a manner that captures their meaning and that can be used with machine learning algorithms (i.e., a vector). For these reasons, I made use of published Latent Semantic Analysis 300-dimensional word vectors trained on the TASA (Touchstone Applied Science Association) database of over 37,651 documents that encode the meaning of 92,393 unique words (Günther, Dudschig, & Kaup, 2015).

To generate one essay representation for each participant, I combined the four essays each participant wrote into one essay. Each participant's essay was represented as the arithmetic mean of the word vectors contained in the essay. Psychologically, the average word could be conceptualized as the semantic gist of the essay. The result of this procedure is a 300-dimensional semantic essay space with each essay occupying a point in the space. As with the semantic word vectors of LSA, essays expressing similar sentiments will be closer in the space (or pointing in the same direction) compared to essays expressing dissimilar sentiments. As with word vectors, the similarity of essay vectors can be measured by their Euclidean distance or cosine similarity.

Classification algorithms. An overwhelming number of algorithms are available to classify exemplars from high-dimensional vector spaces (e.g., Nosofsky, 1986; Rumelhart & McClelland, 1986; for a broad overview see Kuhn & Johnson, 2016). But, classification models fall into several broad categories. For this project, I started by using one model from each of the several larger families of models: similarity-based models, decision boundary models, and probabilistic models.

Classification algorithms like k Nearest Neighbors, Centroid models, and the Generalized Context Model (Nosofsky, 1986), operate by categorizing based on the class of the most similar exemplars to the novel case. I used the k Nearest Neighbors (kNN) model from this family of models. The kNN model predicts novel cases by a two-step process. First, the algorithm computes the Euclidean distance between the unknown case and its neighbors. Then, the algorithm looks at some number (*k*) of nearest neighbors (i.e., items closest to the item in the space) and uses the majority class of these nearest neighbors as a vote for its prediction. When *k* is small, the system behaves according to the principles of instance theory. I tested all odd values of *k* between 1 and 100. Odd values were used to avoid the problem of ties in the vote.

Models like Artificial Neural Networks, Support Vector Machines, and Linear/ Non-Linear Classifiers operate by deriving a decision boundary that divide the vectors of one class from another (Abdi, Valentin, & Edelman, 1999). I used the Support Vector Machine (SVM; Vapnik, 2010) from this family of models. SVMs are similar to an artificial neural network in that they attempt to derive a decision boundary that separates the exemplars of one class from another. Unlike artificial neural networks, the SVM produces a decision boundary that produces the widest margin between the exemplars of different classes.

Probabilistic classification models are another family of machine learning models. I used a Naïve Bayes Classifier from this family of models (Kubat, 2015; James, Witten, Hastie, & Tibshirani, 2013). The Naïve Bayes classifier relies of Bayesian probability theory to classify novel exemplars. For each exemplar, the Naïve Bayes classifier computes the conditional probability of the exemplar belonging to a given class given the values of the exemplars' features.

Each family of models operates according to different principles. Trying several models offers the best chance (in the applied sense) of finding a model that can correctly classify novel essays.

Evaluating model performance. To be useful, a classification model should not merely learn patterns in data, but learn patterns in data that can be generalized for making novel classifications with new data. In this project, the classification models must be able to classify novel essays. To test the models initially, I used only the two most extreme categories of mental health – essays written by *well* versus *severely distressed* participants according to the K10 cut scores. This provides the models with the best initial chance of diagnosing mental health. To ensure an established and well-defined chance model, I randomly sampled the data so that there was an equal number of *well* and *severely distressed* essays (I address this potential issue in the results section).

To test the system's ability to classify novel essays, I randomly split the data into a cross-validation set that contained 80 percent of the essays I collected and a test set that contained 20 percent of the remaining essays. The cross-validation set was used to train the models and establish which combination of parameters performed best. To train the models, I used a cross-validation procedure (e.g., Howell, 2013) where the cross-validation set is repeatedly randomly

sampled into the 80/20 a training and validation sets to train the models and determine the best fit. The cross-validation set was used as an indication of how the models would perform on novel data. However, since it is possible that the models may overfit the cross-validation data and bias the estimate of accuracy upwards (Howell, 2013), the test set was used to test the final chosen models on novel data. If the models fail to learn the generalizable structure of the cross-validation data, the models should not classify the test data appreciably better than would be expected by chance. If the models learn the generalizable structure of the cross-validation data, the models should classify the test data appreciably better than would be expected by chance.

Method Summary

In summary, I had participants write four essays describing their mental health. The participants then completed the K10 psychological distress questionnaire (and four other standard questionnaires). I represented the essays as the average of the words participants used to describe their mental life using LSA word vectors in a high-dimensional essay space. I tested three standard classification models to classify the essays in the space with a kNN, SVM, and Bayesian classifier. I evaluated the performance of the models, not on their ability to classify known cases, but unknown cases. My practical goal was to find a model that correctly classifies peoples' mental health from their written reports at a rate greater than would be expected by chance. My theoretic goal was to show that cognitive psychology has theories and models that offer a solution to the development of technologies to solve practical real-world problems.

Results

Figure 2 (Panel A) shows the classification accuracy using the word averaging representation with LSA vectors. The y axis represents the classification accuracy expressed as a proportion. The results are displayed for the cross-validation (CV) and test data for each

classification model (kNN, SVM, and Bayes). The dashed horizontal line represents a chance model of 50 percent accuracy that a model would be expected to achieve if it was guessing.

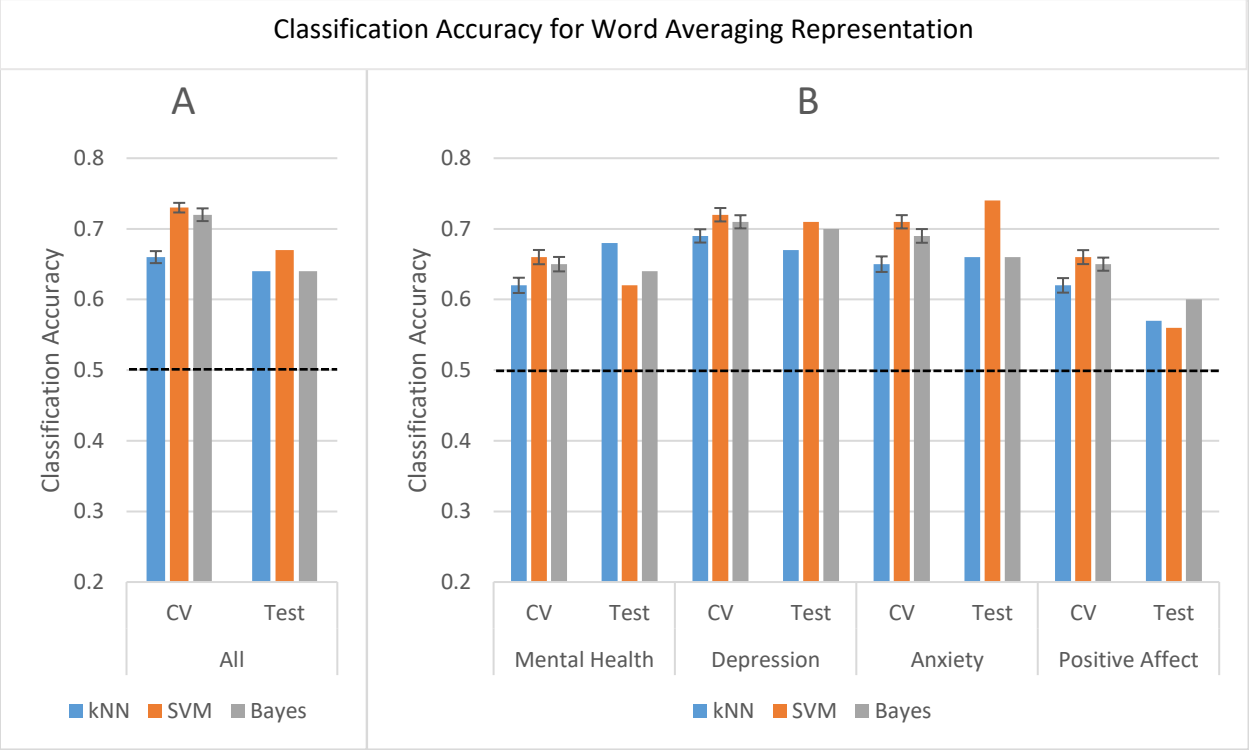

Figure 2. Results with LSA word vector averaging representation. Panel A shows the results with all four essays are combined, and panel B shows the results for each essay. Each panel shows classification accuracy (expressed as a proportion) for the cross-validation and test data as a function of classification model (kNN, SVM, and Bayes). The horizontal dashed line represents a chance model of 50% accuracy. Error bars for the cross-validation accuracy represent the standard error of the mean.

There are several important results of note. First, the three classification models (indicated by bar colour) perform substantially better than a 50 percent chance model on both the cross-validation data and (more importantly) on the test data. Second, there is a strong correspondence between the accuracy on the cross-validation data and the test set, indicating that the models are not over or under fitting the cross-validation data, but rather learning the patterns

in the data in a way that generalizes to new cases. Third, the SVM model tends to be the best performing model, though this advantage is negligible.

Figure 2 (Panel B) show the same results as Panel A, but decomposed by each of the four essays participants wrote (questions assessing mental health, depression, anxiety, and positive affect). The results with all essays combined (Panel A) show the same pattern of results as for each of the four essays when analyzed separately (Panel B).

These results demonstrate that essays written by well and unwell participants can be classified using a standard method of language representation and established classification models. The correspondence between results for individual essays and the aggregate of these essays demonstrates that the models perform as well when classifying one of the four essays participants' wrote as when classifying all essays aggregated. Furthermore, the results demonstrate that representing essays of mental health as the average of its word provides a valid method of representing essays for classification.

Though these results show that the models can classify peoples' mental health when the number of essays from *well* and *severely distressed* participants are equal, the results do not depend on this distribution. The models perform at a comparable (although higher) rate when the numbers of essays from the *well* and *severely distressed* categories are left unaltered (roughly 60% *well* and 40% *severely distressed*).

Furthermore, though these results show that the models can classify peoples' mental health when only the *well* and *severely distressed* groups are use, the results do not depend on this grouping. Figure 3 shows the results when a median split is used (i.e., the *well* and *mildly distressed* groups are combined into one group and the *moderately distressed* and *severely*

distressed groups are combined into a second group). Throughout the rest of the thesis I will continue to use the *well* and *severely distressed* groups.

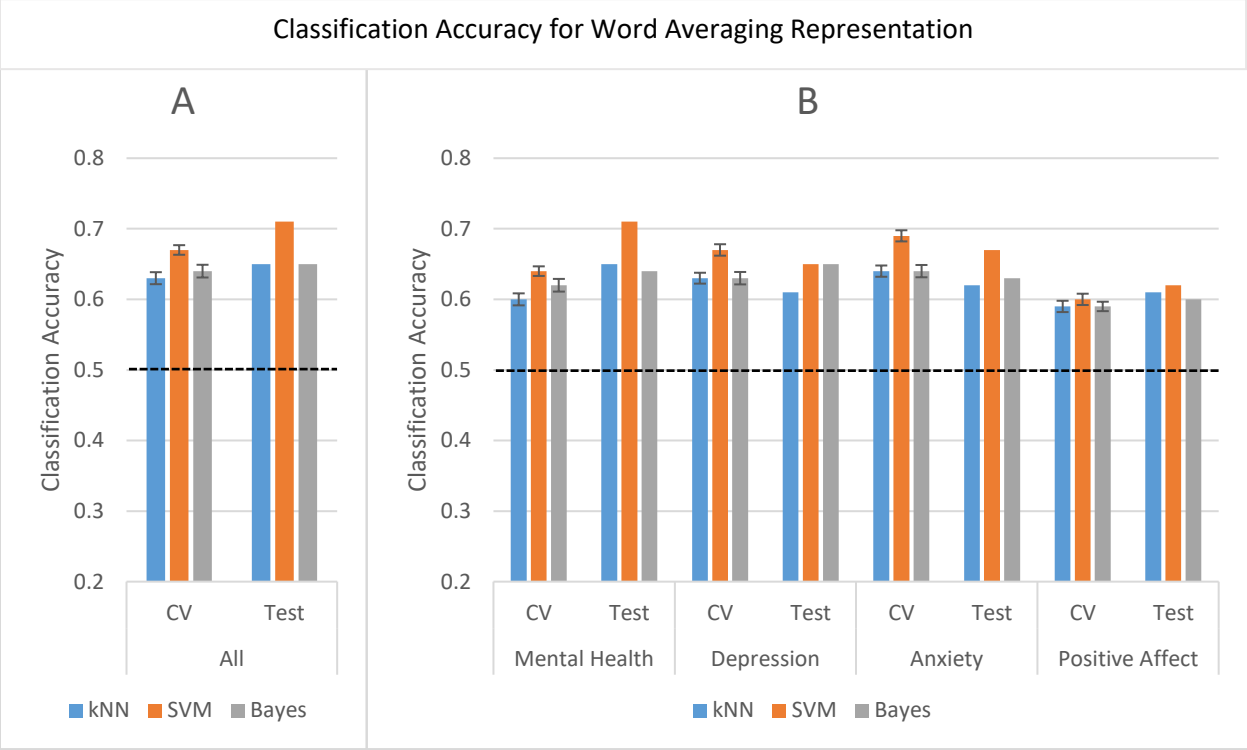

Figure 3. Results with LSA word vector averaging representation when the *well* and *mildly destressed* groups are combined into one group and the *moderately distressed* and *severely distressed* groups are combined into a second group (i.e., median split). Panel A shows the results with all four essays are combined, and panel B shows the results for each essay. Each panel shows classification accuracy (expressed as a proportion) for the cross-validation and test data as a function of classification model (kNN, SVM, and Bayes). The horizontal dashed line represents a chance model of 50% accuracy. Error bars for the cross-validation accuracy represent the standard error of the mean.

Replication with Additional Mental Health Measures

Although the K10 is the primary scale I used to validate the performance of the classification models, I wanted to be confident that the results were not specific to any particular measure. Therefore, I also tested the model's ability to classify people's mental health using four other self-report measures that all participants responded to. In addition to completing the K10

psychological distress scale, all participants completed four other self-report measures of mental health. Specifically, the scales were the Center for Epidemiologic Studies Depression Scale (CES-D), PROMIS (Patient-Reported Outcomes Measurement Information System) Short Form v1.0 - Depression 8b, PROMIS Short Form v1.0 - Anxiety 8a, and the Neuro-QoL (Quality of Life) Short Form v1.0 - Positive Affect and Well-Being scale (see Appendix A). Whereas the K10 scale measures psychological distress (broadly defined), the first two scales are specific measures of depression, and the last two scales are specific measures of anxiety and positive affect, respectively. These scales assess the extent to which emotions are experienced, measured with a 5-point Likert-type scale where 1 = *Never,* 2 = *Rarely,* 3 = *Sometimes*, 4 = *Often,* and 5 = *Always*. Like the K10, these tests give a single summative score. Higher scores on these scales indicate greater depression, anxiety, and positive affect. Unlike the K10, these measures do not have cut scores that categorize scores. For these scales, I created my own cut scores that I based on percentile ranking analogous to the *well* and *severely distressed* groups from the K10. For each scale, I retained the 30 percent of participants with the lowest scores and the 30 percent of participants with the highest scores to form two groups. I repeated the exact same analysis as previously reported with the K10 with each of the four other measures.

Table 1 shows the classification accuracy for each essay question and for each classification model. For fair comparison, the cross-validation and test data used with these other four measures are identical to the previous study.

Table 1

Replication of K10 classification results with alternate measures

	Mental Health		Depression		Anxiety		Positive Affect		All Essays	
	CV	**Test**	**CV**	**Test**	**CV**	**Test**	**CV**	**Test**	**CV**	**Test**
Center for Epidemiologic Studies Depression Scale (CES-D), NIMH										
kNN	0.60	0.71	0.63	0.65	0.65	0.60	0.58	0.62	0.61	0.67
SVM	0.66	0.71	0.67	0.69	0.66	0.68	0.62	0.68	0.69	0.75
Bayes	0.63	0.70	0.65	0.69	0.64	0.59	0.61	0.67	0.63	0.67
PROMIS Short Form v1.0 - Depression 8b										
kNN	0.64	0.57	0.69	0.62	0.64	0.59	0.63	0.62	0.67	0.63
SVM	0.67	0.60	0.73	0.71	0.68	0.65	0.64	0.67	0.73	0.69
Bayes	0.65	0.60	0.72	0.65	0.63	0.63	0.62	0.65	0.69	0.68
PROMIS Short Form v1.0 - Anxiety 8a										
kNN	0.59	0.59	0.67	0.66	0.67	0.73	0.55	0.57	0.66	0.66
SVM	0.63	0.73	0.68	0.66	0.72	0.72	0.60	0.64	0.70	0.72
Bayes	0.62	0.69	0.67	0.60	0.67	0.68	0.59	0.61	0.66	0.70
Neuro-QoL Short Form v1.0 - Positive Affect and Well-Being										
kNN	0.64	0.52	0.66	0.66	0.60	0.61	0.65	0.67	0.65	0.66
SVM	0.65	0.57	0.69	0.63	0.64	0.61	0.68	0.70	0.68	0.68
Bayes	0.64	0.59	0.65	0.68	0.62	0.61	0.66	0.70	0.65	0.63

Note: The standard error of the mean for the CV data are all between 0.0075 and 0.0102.

As demonstrated in Table 1, the main results from the previous analysis are replicated using these four scales. Importantly, across every condition the models perform substantially better than a 50 percent chance model. The classification accuracy with both the cross-validation and the test data are in strong agreement, indicating that the models are not over or under fitting the cross-validation data, but rather learning the patterns in the data in a way that generalizes to new cases. Having shown the results are not specific to any specific measure, I will continue to use the K10 as the category variable for the rest of results reported, due to its pre-defined cut scores.

Repeating the Analysis using Different Semantic Word Vectors

The results reported thus far have used LSA word vectors to construct the representation of essays. This demonstrates that LSA succeeds as a representation scheme for constructing the essay vectors using the average of all words participants used in their essays. However, LSA is first-generation model of distributional semantics, and whereas the model successfully learns that *dog* and *cat* are semantically related words because they often appear in sentences such as *the dog chased the cat*, the model fails to account for differences in the order that words are used. For example, the LSA representation of the sentence *the dog chased the cat* is indistinguishable from the sentence *the cat chased the dog*. It is possible that a more sophisticated model of semantics, one that encodes word order rather than ignores it, will allow the classification models to perform substantially better.

Second-generation models, notably Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007), were invented to encode word order information in addition to contextual information. BEAGLE represents a words' context and order information in a composite vector. Formally, each word in the corpus is initially represented by a random vector drawn from a Gaussian distribution with the parameters $\mu = 0$ and $\sigma = 1/\sqrt{n}$, where n is the dimensionality of the vector. Vector dimensionality is typically set at 1024 (Jones & Mewhort, 2007). BEAGLE represents word meaning as LSA does, but accomplishes the task by summing neighboring word vectors to a target word's vector representation. This has the effect of making the word vectors that appear in the same contexts more similar.

BEAGLE represents word order information using a mathematical operation called circular convolution (Plate, 1995). BEAGLE convolves a target word's vector with the

neighboring words in the same sentence. Formally, circular convolution is an operation that collapses the outer product matrix to a vector of dimensionality *n*,

$$\mathbf{z} = \sum_{j=0}^{n-1} \mathbf{a}_{j\ mod_n} \cdot \mathbf{b}_{(i-j)\ mod_n} \qquad \{\text{for } i = 0 \text{ to } n - 1\}$$

where **z** is a vector of the convolution of vectors **a** and **b** ($i.e., \mathbf{z} = \mathbf{a} \circledast \mathbf{b}$). Vectors **a**, **b**, and **z** all have dimensionality *n*. The subscript mod_n refers to the modulo operator, as the indices of the vectors are modulated by *n*. Circular convolution allows for a distributed and holographic (as opposed to local) representation of word order information.

A word's order vector is computed by summing the convolutions of *n*-gram chunks of the neighboring words. Formally, the order vector of word *i* is computed by

$$\mathbf{o} = \sum_{j=1}^{p\lambda-(p^2-p)-1} bind_{ij},$$

where **o** is the order vector for the *ith* word in the corpus, *p* is the position of the word in the sentence, λ is a parameter that defines the maximum number of neighbors a word can be convolved with, and $bind_{ij}$ is the convolution of word *i* with word *j*. Traditionally λ is set to 7 (Jones & Mewhort, 2007) consistent with Miller's (1956) famous number 7 ± 2.

For example, computing the order information of the word *fox* in the sentence *the fox ran quickly* involves binding together each *n*-gram chunk in the sentence that include the word *fox*. All the *n*-grams that contain the word *fox* in the sentence *the fox ran quickly* includes: bigrams (*the fox, fox ran),* trigrams (*the fox ran*, *fox ran quickly*), and quadgrams (*the fox ran quickly*). Each *n*-gram that is bound using circular convolution (*the* ⊛ Φ, Φ ⊛ *ran*, *the* ⊛ Φ ⊛ *ran*, etc.) is summed to form the word's order vector, where Φ is a placeholder vector for the word of interest (e.g., *fox*). This procedure results in an order vector for each word. The effect of this

procedure makes words with similar functions in language (e.g., nouns versus verbs) more alike, in much the same way that words with similar meaning become more alike. The word meaning and word order vectors are normalized to unit length and summed into a composite vector representation that contains both meaning and order information.

I will now use BEAGLE word vectors to build a representation of participants' essays to test the effect of including using word vectors that also contain order information using 1024 dimensional composite vectors.[1] For this experiment, I followed the same procedure to build the essay vectors as with the LSA vectors where I built each participant's essay vector by taking the average of all words vectors that composed a participant's essay. The BEAGLE word vectors were derived from the same TASA corpus of text as the LSA word vectors to offer a fair comparison of results for the different semantic representation schemes.

Figure 4 (Panel A) shows the classification accuracy using the word averaging representation using BEAGLE vectors. The y axis represents the classification accuracy expressed as a proportion. The results are displayed for the cross-validation and test data for each classification model (kNN, SVM, and Bayes). The dashed horizontal line represents a chance model of 50 percent accuracy.

[1] I would like to thank Dr. Brendan Johns for providing me with BEAGLE word vectors.

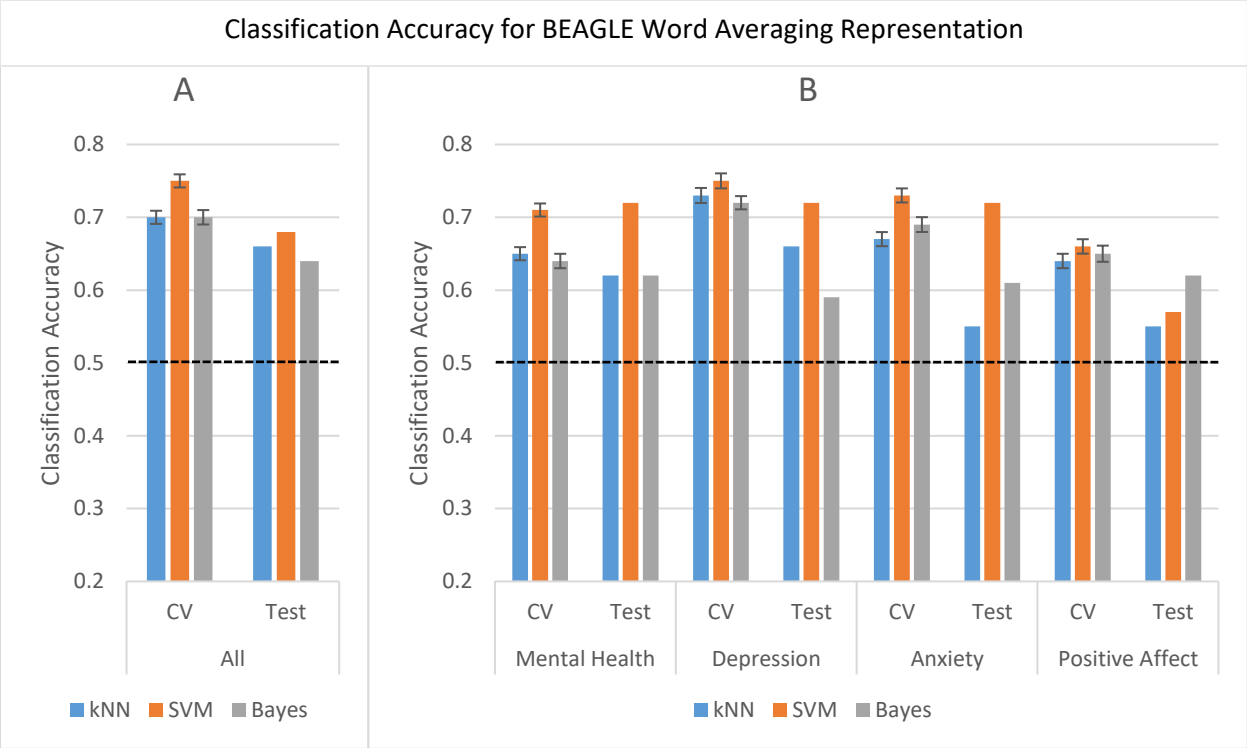

Figure 4. Results with BEAGLE word vector averaging representation. Panel A shows the results with all four essays are combined, and panel B shows the results for each essay. Each panel shows classification accuracy (expressed as a proportion) for cross-validation and test data as a function of classification model (kNN, SVM, and Bayes). The horizontal dashed line represents a chance model of 50% accuracy. Error bars for the cross-validation accuracy represent the standard error of the mean.

The main results with LSA are replicated using BEAGLE word vectors and do not appear to dramatically improve classification accuracy compared to the LSA vectors. Importantly however, across every condition the models perform substantially better than a 50 percent chance model. The classification accuracy with both the cross-validation and the test data are in strong agreement, indicating that the models are not over or under fitting the cross-validation data, but rather learning the generalization patterns in the data. Figure 4 (Panel B) show the same results as Panel A, but decomposed by each of the four essays. The results with all essays combined

(Panel A) show the same pattern of results as for each of the four essays when analyzed separately (Panel B). Going forward, I will continue to use LSA word vectors.

Repeating Simulations using another Representation

So far, I have represented each essay as the average of all the words of which the essay is composed. However, the word averaging representation suffers from a known problem called the *bag of words* problem. With the word averaging representation I have used, because the representation does not encode sequential information, an essay with the words randomly permuted would result in the same representation as a coherent essay where one thought logically leads to the next. Though it is unlikely any participant submitted a permuted essay (i.e., an essay without syntax or grammar), it is possible that there are differences in the levels of coherence in the writing between *well* and *severely distressed* participants. To examine the issue, I will use LSA to build a representation of coherence in participants' essays to evaluate how well this representation allows the classification models to classify novel essays.

The word averaging method used in the previous simulations encodes words of an essay without regard for the coherence of syntax or the thoughts expressed in the essays. To represent differences in how people use words and connect one idea to the next (and not just which words were used), I conducted another computational test where I represented the words as a summary of the similarity between neighboring sentences. This representation is inspired by Bedi et al. (2015) use of a similar method that had excellent applied success detecting schizophrenia from written language.

To build a representation for coherence, I first formed a sentence vector that was the sum of the word vectors in a sentence. This resulted in one sentence vector for each sentence in a participants' essay. Then, I computed the cosine similarity between each neighboring pair of

sentences. This resulted in $s - 1$ cosines, where s is the number of sentences in an essay. Because participants used varying number of sentences in their essays, there is no direct way to compare vectors of varying lengths; yet classification models require fixed length vectors. To solve this problem, I computed six descriptive statistics to summarize the cosines. Each participants' essay was represented as a six-dimensional vector that encoded the minimum cosine, 25^{th} percentile cosine, mean cosine, median cosine, 75^{th} percentile cosine, and maximum cosine. Large cosines mean that neighboring sentences are semantically related to one another. Small or negative cosines would indicate that participants were more disordered in their use of language from one sentence to the next. If participants vary in the coherence in expressing their language on their mental health, this coherence representation should classify essays at a rate better than chance.

Coherence Representation Results

Figure 5 (Panel A) shows the classification accuracy using the coherence representation. The y axis represents the classification accuracy expressed as a proportion. The results are displayed for the cross-validation and test data for each classification model (kNN, SVM, and Bayes). The dashed horizontal line represents a chance model of 50 percent accuracy.

The kNN, SVM, and Bayesian models performed better than a 50% chance model on the cross-validation data, whereas only the SVM and Bayesian performed better than a 50% chance model on the test data. The coherence representation does not appear to perform as well as the word averaging representation.

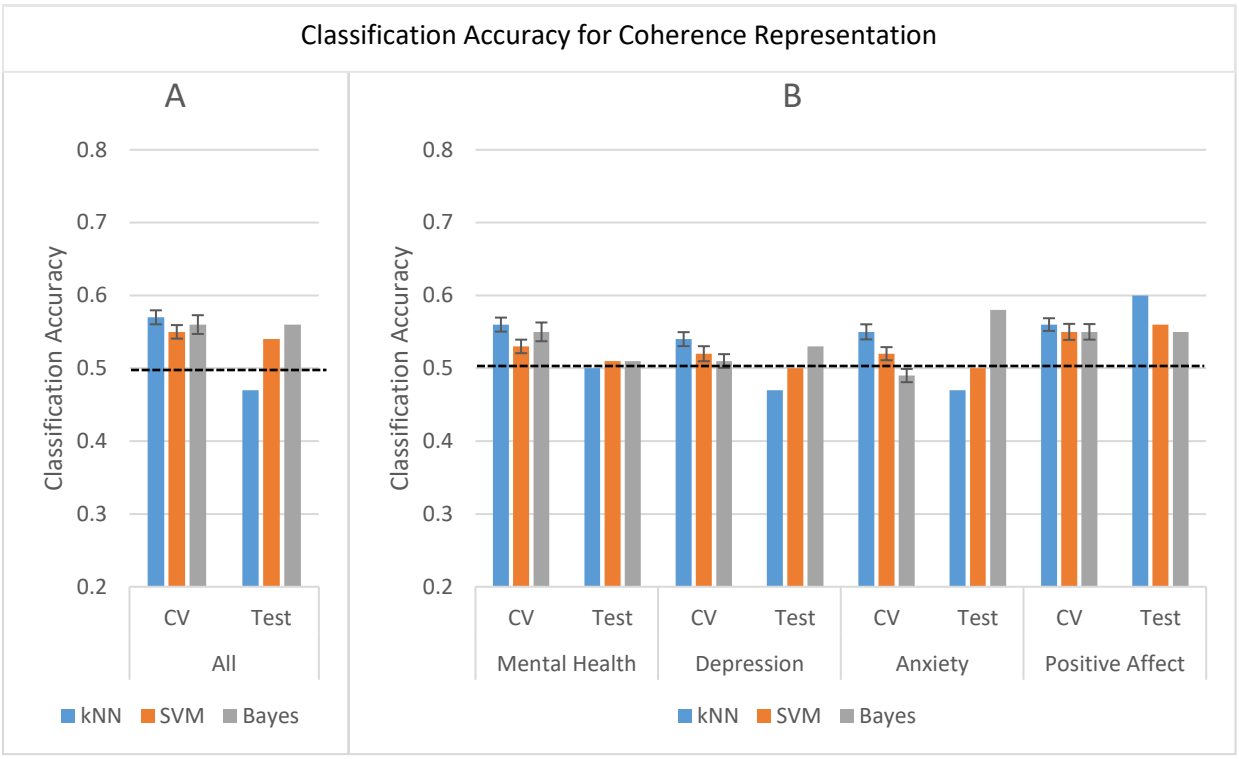

Figure 5. Results with coherence representation using LSA word vectors. Panel A shows the results with all four essays are combined, and panel B shows the results for each essay. Each panel shows classification accuracy (expressed as a proportion) for cross-validation and test data as a function of classification model (kNN, SVM, and Bayes). The horizontal dashed line represents a chance model of 50% accuracy. Error bars for the cross-validation accuracy represent the standard error of the mean.

To get a closer look, Figure 5 (Panel B) shows the same results as Panel A, but separated for each of the four essays that the participants wrote. As shown, the models do not appear to consistently perform better than chance. Furthermore, it is difficult to detect any consistent advantage of either of the three classification models. I conclude that the coherence representation does not adequately capture information in the essays.

Bedi et al. (2015) reported incredible success using a coherence representation of language to diagnose schizophrenia. A possible reason why the coherence representation did not work in this study is because our mental health categories of depression and anxiety are not

characterised by thought disorder and deviant verbalizations the way schizophrenia is characterized (Levy et al., 2010).

Psychologically-Inspired Classification Models

The classification models I have tested show that models based on very different principles (similarity-based models, decision boundary models, probabilistic models) can classify novel essays significantly better than chance, at least when using the word averaging representation. However, these models are the product of engineered solutions to classification.

In contrast, modern cognitive psychology has psychologically-inspired models of classification that not only seek to fit data, but to fit data in a way that is psychologically plausible. Specifically, Nosofsky's General Context Model (GCM; 1986) is a classic instance-based psychologically-inspired classification model that has been shown to succeed in a wide range of classification tasks. Though the model was originally developed for perceptual classification, I wanted to determine if it is possible that a psychologically-inspired classification model will perform as well as engineered solutions for text classification such as the k Nearest Neighbors, Support Vector Machine, and Bayesian approach.

Generalized Context Model (GCM)

To classify a novel exemplar, the GCM first computes the distance between the novel exemplar and the cross-validation data (i.e., the model's history of experience). The Euclidean distance formula provided earlier is a specific instance of the generalized Minkowski *r*-metric

$$\mathrm{d} = \left[\sum_{i=1}^{n} |\boldsymbol{a}_i - \boldsymbol{b}_i|^r\right]^{1/r}$$

where d is the distance between vectors **a** and **b** of dimensionality *n*. When *r* = 2, the formula is identical to the Euclidean distance, and when *r* = 1 the distance is the Manhattan (or city-block) metric.

However, distance does not correspond with psychological similarity (e.g., Brown, Neath, and Chater, 2007). To relate distance to psychological similarity, the distances between the novel exemplar and the training exemplars are transformed using one of two functions, the exponential decay function

$$\eta = e^{-d}$$

or a Gaussian function

$$\eta = e^{-d^2}$$

These functions are based on theoretical and empirical deliberations (e.g., Nosofsky, 1985b; Shepard, 1958a, 1958b).

Given similarity values, the model computes the conditional probability of an exemplar belonging to each of *c* categories, where the similarity is converted to a probability of category membership. The probability is computed as the similarity of the exemplar with all exemplars from a particular category relative to the similarity will all exemplars irrespective of category membership

$$P(C_i|S) = \frac{\eta_i}{\sum_{i=1}^{n} \eta_i}$$

where $P(C_j \mid S)$ is the probability of exemplar S belonging to category *j* where *n* is the number of categories. The category associated with the largest probability is the predicted category for the novel exemplar.

The GCM also modifies the Minkowski *r* metric by weighting each dimension in the space

$$\mathrm{d} = \left[\sum_{i=1}^{n} \mathbf{w}_i |\boldsymbol{a}_i - \boldsymbol{b}_i|^r\right]^{1/r}$$

where the vector **w** weights each element of vectors **a** and **b**. The sum of weights is equal to 1 and represents a set of attentional weights that stretch or shrink the psychological exemplar space. When each weight is 1/*n* the formula provides an unweighted solution, analogous to a person considering each feature of the exemplar equally. As the components of the **w** vector deviate from uniform, some features of the exemplar are attended to more, which necessitates other features being given less attention (because attention is a finite resource). Nosofky wrote, "it is assumed that subjects will distribute attention among the component dimensions so as to optimize performance in a given categorization paradigm". For this project, I will use information theory (Shannon & Weaver, 1949) to derive attentional parameters **w** for the model, reasoning that the attentional weights should be proportional to the amount of information a variable contains. Shannon's measure of information, or uncertainty, is defined as

$$H = \sum_{i=1}^{n} \mathbf{p}_i \log_2 \frac{1}{\mathbf{p}_i}$$

where **p** is a vector of (non-zero) *n* probabilities.

I computed the amount of information each of the 300 dimensions in the LSA solution (i.e., columns) contributes to knowledge about the categories. To do this, I computed the amount of information *H* for each variable irrespective of category. Then, I computed the average *H* for each variable within each category. The difference between these two measurements provides an index of the amount of information contained in any given feature. Formally,

$$H_{gain} = H - \frac{1}{n}\sum_{i=1}^{n} H_i$$

where H_{gain} is the amount of information a variable contributes about the category membership. H is Shannon's information of the variable irrespective of category membership. H_i is Shannon's information of the variable for category i, and n is the number of categories. This method is borrowed from the classification models called decision trees (e.g., Kubat, 2015).

To evaluate the GCM, I used the identical cross-validation and test sets as I used in my previous experiments to ensure a fair comparison against the first study with the kNN, SVM, and Bayesian approaches. I evaluated the model with an exponential decay and Gaussian function. I used the cross-validation set to estimate the weights **w** that act as attentional parameters in the model and these weights were used in the model to attempt to classify the novel test data.

GCM Classification Results

Figure 6 (Panel A) shows the classification accuracy using the GCM model with the word averaging representation. The y axis represents the classification accuracy expressed as a proportion. The results are displayed with the same test data as the previous results and are displayed by similarity function (exponential decay and Gaussian) as indicated by the coloured bars. The dashed horizontal line represents a chance model of 50 percent accuracy.

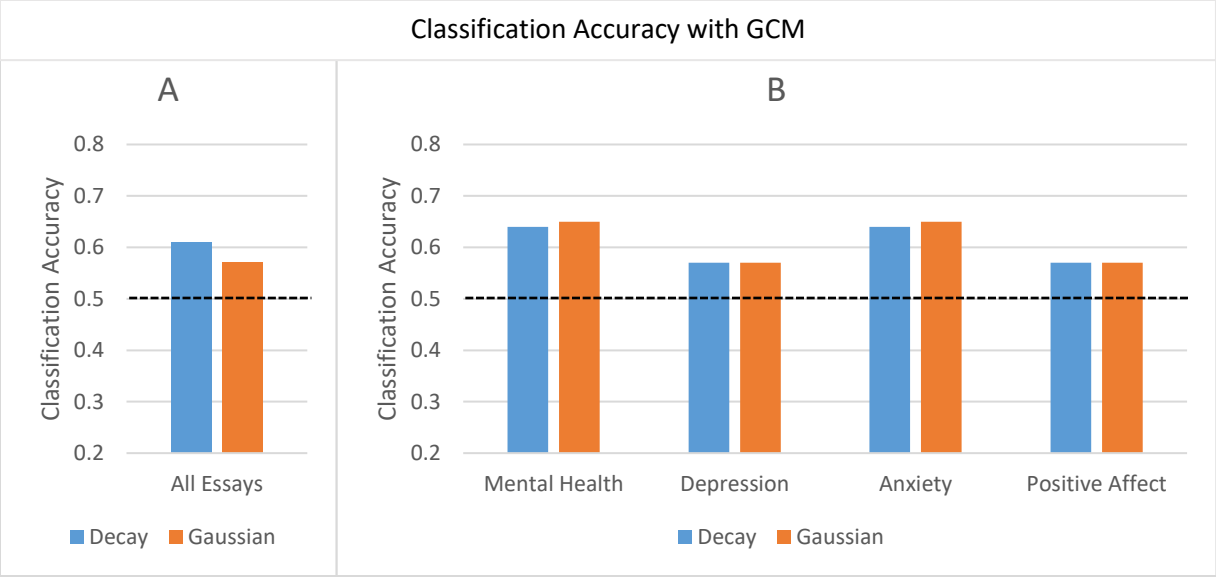

Figure 6. Results with Generalized Context Model (GCM) using word averaging representation with LSA word vectors. Panel A shows the results with all four essays are combined, and panel B shows the results for each essay. Each panel shows classification accuracy (expressed as a proportion) for the test data as a function of the similarity function (exponential decay and Gaussian). The horizontal dashed line represents a chance model of 50% accuracy.

There are several results to note. First, the classification accuracy irrespective the similarity function being used is above the 50 percent chance line. Second, the exponential decay function appears to perform better than the Gaussian function. Third, compared to the machine learning models (see Figure 2, Panel A), the GCM does quite well. The GCM with the exponential decay function only performed 3% lower than the kNN and Bayesian classification models and 6% lower than the SVM classification model.

Figure 6 (Panel B) show the same results as Panel A, but decomposed by each of the four essays. The results with all essays combined (Panel A) show the same pattern of results as for each of the four essays when analyzed separately (Panel B), but there are some other results to note. First, though the GCM doesn't outperform the machine learning classification models, with some essays (e.g., the anxiety essay) it performs as well as the kNN and Bayesian models on the

exact same test data. Second, there does not seem to be any clear advantage of the similarity function that is used – the performance of the exponential decay and Gaussian functions appear to perform roughly the same and one does not consistently outperform the other.

In summary, the GCM, a classic instance-based model of categorization originally developed for modelling perceptual categorization judgements can be successfully extended to the domain text of classification. By using a psychologically-inspired method of classification, in addition to models of distributed semantics like LSA and BEAGLE, the model offers a fully psychologically-inspired method for text classification. Though the model did not perform better than the engineered machine learning classification models, the method offers promise for a complete psychologically-inspired method of natural language text classification that deserves additional analysis and consideration.

Replication with Additional Databases

To give the reader some confidence in the generalizability of the results I have reported, I have also replicated the main results of this project with two other databases.

One year prior to the collection of data reported in this thesis, I collected data from nearly 500 University of Manitoba undergraduates. These data were a smaller version of the data I presented in this thesis: Participant's only wrote one essay with a minimum of 200 words asking them to describe their mental health, and were only asked to complete one mental health scale (the Kessler K10). The models successfully classified novel essays at a rate consistent with the results reported in this paper.

I have also been fortunate enough to apply these methods on a 30-year database of transcripts of clinical interviews using the Rorschach (inkblot) test with non-psychiatric controls, schizophrenic patients, bipolar patients, and family members of the schizophrenic and bipolar

patients (described in Morgan et al. 2017). Each of the over 700 transcripts were diagnosed by professional clinical psychologists. The same classification methods presented in this paper work with this database as well.

These past results, taken with the results in this thesis, present strong evidence that language and classification models from psychology can successfully diagnose peoples' mental health from both written and transcribed verbal language. These results apply to data collected online with university undergraduates as well as a psychiatric population. The methods work for identifying emotion-based disorders like depression and anxiety, as well as thought-based disorders like schizophrenia and bipolar disorders. The methods also work when the category of mental health is determined by self-report measure or by professional diagnosis.

Developing a Method for Diagnostic Language Identification

The models evaluated can correctly diagnose the mental health of people from novel examples of written language. But, because the models' classify entire essays, they provide no insight into the words or phrases of an essay that are particularly diagnostic. Rather, the models simply provide a global diagnostic decision with no clinically relevant reasons to support the decision. In this project, I have used models of distributed semantics like LSA and BEAGLE to show we can represent peoples' written language describing their mental health in a way that allows standard classification models to classify the mental health of a person who wrote the essay. I also wanted to explore methods of using these language models from cognitive psychology to analyze the words and phrases people use to describe their mental life. By doing so, we will not only be able to diagnose people's mental health from their written language, but to also identify diagnostic passages and phrases. Specifically, I wanted to use the language models I have considered to identify words that signal depression. My goal was to build a system

that can read in an essay and rewrite the essay as a language heatmap, with words signalling depression highlighted for a visual representation of written language.

To accomplish this, I compared every word in a participants' essay to a prototypical example of words associated with depression. Words or phrases that are similar to the prototype vector would signal psychologically depressing sentiment.

A depression prototype vector was built by summing the word vectors for 13 words related to depression (*depressed*, *sad*, *worthless*, *depressing*, *useless*, *unhappy*, *hopeless*, *shame*, *guilt*, *numb*, *empty*, *irritable*, and *lonely*). The purpose of building a prototype vector rather than using a single word vector for identifying psychologically distressing language is that by definition, a prototype is a good representation of category membership. Kwantes (2016) used a similar method to identify language revealing of the Big Five personality traits.

To build the language heatmap, I computed the cosine similarity between each word in the depression essay to the depression prototype. I then rewrote the essay as an HTML file, color coding each word based on its cosine similarity to the prototype. Words with a cosine similarity less than 0.2 are written in black, whereas words with a cosine similarity greater than 0.2 are colored ranging from yellow to dark red. Figure 7 shows four of the depression essays authored by severely distressed participants rewritten by the program as a depression-defined heatmap. Ideally, we would like to see depressive language highlighted and non-depressive language not highlighted.

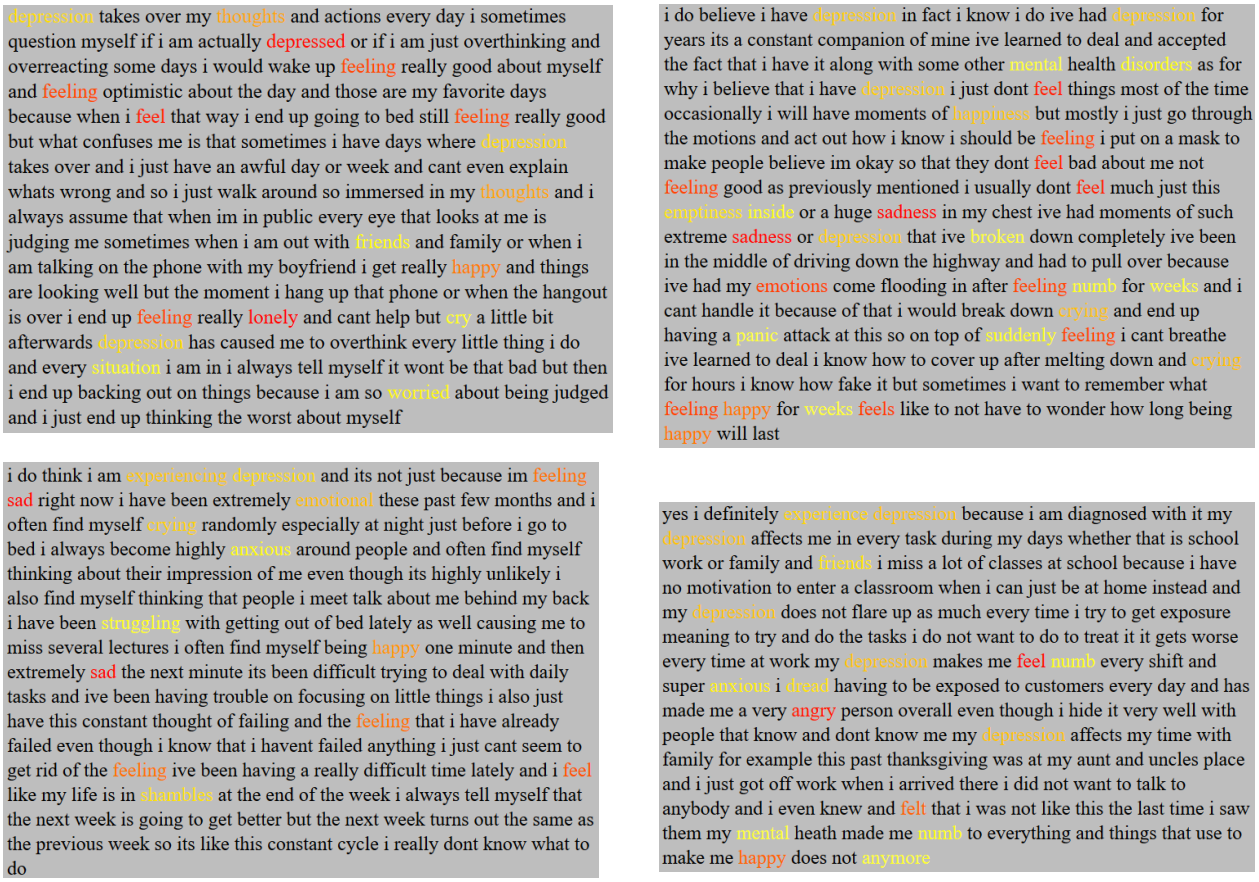

Figure 7. Essays about depression written by four severely distressed participants. The essays are rewritten by a program I developed to generate a language heatmap where darker red colors correspond to more depressive language. The extent of depressive language is determined by each words' similarity to a vector prototype.

There are a few results to note. First, the system correctly highlights many of the words that compose the prototype, such as *depression*, *sad*, and *numb*. This is a recognition-based word-classification results. Second, the system also identifies depression related words that were not used to compose the prototype such as *sadness*, *broken*, and *struggling*. The fact that the system identifies depressive language that were not explicitly included in the prototype representation points to the motivation and value for using semantic vectors to identify language that provides insight into a diagnosis. Third, whereas there are many correct identifications, there

are also misidentifications, such as when the system highlights words such as *happy* or *friends*. This occurs simply because these words are related the words used to construct the prototype. Lastly, there are also a few instances in these essays where the system fails to identify words that may be diagnostic such as *failing*.

Overall, the system seems to correctly identify language that signals depressive thoughts and emotions. The goal of this technology is to build a tool that can augment clinicians' wisdom and clinical experience. The system would allow clinicians to identify depressive language in a large body of written or transcribed text at a glance, augmenting and facilitating diagnostic decision making. Taken together with the classification results, the system I have developed can classify people's mental health and rewrite the text with a heatmap overlay, providing an automated aid for clinical diagnosis and a tool to identify potentially diagnostic language.

Discussion

I have used established models of distributed semantics based on state-of-the-art techniques from cognitive psychology to build a representation of language that people use in service of describing their mental health. Having built an essay space using these semantic vectors, I applied engineered machine learning based classification models as well as classification models developed within cognitive psychology. Several combinations of representation and classification model produced classification accuracies in the low seventy percent range on novel test data, with the best model classifying the test data with 74 percent accuracy.

Admittedly, classification accuracy of 74 percent is less than I had hoped when starting this project. However, given the ill-defined nature of mental health, the lack of precision in the questions participants responded to (e.g., describe your mental health), and the generative and

imprecise natural of human language in general, classification accuracies in the seventy percent range may very well be near the upper limit of what is achievable.

However, the results reported are also not so unlike the accuracy of professional clinicians. For example, in a meta-analysis of judgement accuracy for practicing clinical psychologists, the accuracy of novice and expert clinicians was only 47 and 53 percent, respectively (Spengler et al., 2009). Accuracy may be even worse in medical diagnosis. Meyer et al. (2013) conducted an experimental study in which nearly 120 physicians diagnosed only 53.3 percent of easy cases and 5.8 percent of difficult cases. Considering these results, classification accuracy of 74 percent is better than it first appears. However, it is important to note that both the psychologists and physicians in these reported studies had an overwhelming number of illnesses to select from, whereas our models had only a two-choice forced decision with a 50% chance model.

This project (and related work) has important implications for the study of language and mental health more generally. Our mental health is complicated to understand and study. These complexities are exaggerated by the fact that we must use language to describe our mental lives. And yet, diagnostic tools like the K10 psychological distress scale represent mental health on a one-dimensional integer scale. This representation scheme limits the question we can ask regarding the relationship between language and mental health, as well as the types of analyses we can apply to a database of mental health reports.

In contrast, by representing language describing peoples' mental health in a formal mathematical framework in a space of hundreds or thousands of dimensions, there are nuances represented that are washed out in a one-dimensional integer scale representation. By probing the structure of these semantic spaces with machine learning classification models, or studying the

spaces using other methods (i.e., clustering methods), we can begin to answer more complex (and interesting) questions than current clinical analytic methods allow.

I see this thesis as a direct continuation of the goals and research started by Meehl back in the 1950s. Meehl saw the value in automated methods for clinical diagnosis: the equivalent or improved accuracy over clinicians, the transparency in which the models make decisions, and the intellectual challenge of developing and investigating methods that researchers and clinicians believed were beyond the limits of quantitative methods. This thesis was conducted in the spirit of Meehls ambitions, while drawing on the theories, techniques, and models from modern psychology's distributed models of semantics and psychologically-inspired classification models.

Artificial Intelligence and Cognitively-Inspired Technology

As a particular applied aim, the goal of this research is to build a cognitively-inspired technology for automated mental health diagnosis. The system could be used as a pre-screening tool to reduce the burden on mental health care diagnosticians and providers, especially in remote areas where mental health care providers are a scare resource. Building a system of this sort has several important advantages. The vector space models of distributed semantics that I have used in this project allow for peoples' thoughts and emotions to be communicated through natural language. The goal of communicating with machines using our natural language is not only the stuff of science fiction, but is also visible throughout the history of computer science as programming evolved from binary instructions on punch-cards to high-level programming languages that resemble natural language. Additionally, we know there are therapeutic benefits to merely writing about ones' thoughts and feelings (Pennebaker & Beall, 1986).

In order to take the work presented in this thesis that were developed and tested in a well-controlled and artificial setting to the development of an automated diagnostic tool that is used in the real world, several further problems would need to be addressed.

First, the results reported in this thesis used an equal number of essays from *well* and *severely distressed* participants to ensure an established and consistent chance model across experiments. Yet in the real world, the division of *well* vs *severely distressed* people is far from 50-50. The models I have tested in this project would need to be tested with the imbalances in groups found in the real world. One approach would be to further analyze the sensitivity (true positive rate) and specificity (true negative rate) of the models rather than only analyzing their overall accuracy with an established chance model. Another approach would be to change the decision mechanism of the kNN and GCM models (i.e., change the threshold from 0.5 so that the models do not merely base their prediction from a majority vote).

Second, the results reported in this thesis used a university sample and may not be generalizable to a non-university sample. However, my work applying the same methods reported in this thesis on a collection of transcripts from a psychiatric population may offer some reassurance. That analysis shows that the results reported in this thesis are not limited to the contrived and controlled analysis, but apply more broadly to the real-world diagnostic problem of the clinical interview.

The Role of Psychologists in Artificial Intelligence

Modern established companies like Google, Facebook, and IBM are also interested in developing solutions for automated clinical diagnosis from written language. However, their approach is to develop an engineered solution to clinical diagnosis. The approach that I explored with distributed theories of semantics and the Generalized Context Model classification approach

in this project is fundamentally different. Rather than trying to invent artificial intelligence from scratch, I have leveraged the empirical and theoretical contributions made by cognitive psychology over the past six decades. Cognitive psychology has prudently collected data and developed theories and models that describe the processes by which humans perceive, learn, think, decide, remember, and know. By studying and understanding these processes, we may be able to leverage what we know about human knowledge representation and classification processes to build technologies and methods that align with human intuition and judgement. Because psychologists study the process by which natural intelligence emerges, they can use their knowledge of natural intelligence to serve as a productive and empirically informed model for developing artificial intelligence. I believe psychologists have a unique role to play in the development of artificial intelligence because they are not forced to approach the problem of developing artificial intelligence from scratch. Rather, they approach problems with data, theories, models, and techniques that have been tested against natural behaviour.

References

Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks* (Quantitative applications in the social sciences; 07-0124). Thousand Oaks, Calif.: Sage Publications.

Andrews, G., & Slade, T. (2001). Interpreting scores on the Kessler Psychological Distress Scale (k10). *Australian and New Zealand Journal of Public Health, 25*, 494-497.

Aujla, H., Jamieson, R. K., & Cook, M. T. (in press). A psychologically inspired search engine. *Lecture Notes in Computer Science: High Performance Computing Systems and Applications.*

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1, 15030.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale: Lawrence Erlbaum Associations, Inc.

Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 141–174). Cambridge: Cambridge University Press.

Brooks, L. R., LeBlanc, V. R., & Norman, G. R. (2000). On the difficulty of noticing obvious features in patient appearance. *Psychological Science, 11*, 112-117.

Clark, A, & Chalmer, D. (1998). The Extended Mind. *Analysis, 58,* 7-19.

Denhière, G., Lemaire, B., Bellissens, C., & Jhean, S. (2008). A semantic space for modeling children's semantic memory. In T. K. Landauer, D. S. McNamara, S. Dennis, and W.

Kintsch (Eds.), *Handbook of latent semantic analysis.* 143-165. Mahwah, NJ: Laurence Erlbaum and Associates.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In Philological Society (Great Britain) (Ed.), *Studies in linguistic analysis*. Oxford, England: Blackwell.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1, 939-944.

Freud, S. (1901). *Psychopathology of everyday life*. New York: Basic Books.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun - An R package for computations based on Latent Semantic Analysis. Behavior Research Methods, 47, 930-944.

Hintzman, D. L. (1984). MINERVA-2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers, 16*, 96–101.

Howard, M. W., Addis, K. A., Jing, B., & Kahana, M. J. (2007). Semantic structure and episodic memory. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 121–141). Mahwah, NJ: Laurence Erlbaum and Associates.

Howell, D. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Thomson Wadsworth.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103, Springer Texts in Statistics). New York, NY: Springer New York.

Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., Unverzagt, F. W., & Jones, M. N. (2017). Cognitive Modeling as an Interface between Brain and

Behavior: Measuring the Semantic Decline in Mild Cognitive Impairment. *Canadian Journal of Experimental Psychology*.

Jones, M., & Mewhort, D. (2007). Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review*, *114*(1), 1-37.

Kessler, R. C., Green, J. G., Gruber, M. J., Sampson, N. A., Bromet, E., Cuitan, M., . . . Zaslavsky, A. M. (2010). Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) survey initiative. *International Journal of Methods in Psychiatric Research, 19*(1), 4-22.

Kubat, M. (2015). *An Introduction to Machine Learning*. Cham: Springer International Publishing.

Kuhn, M., & Johnson, K. (2016). Applied Predictive Modeling. New York: Springer.

Kwantes, P.J. (2005). Using Context to Build Semantics. *Psychonomic Bulletin & Review: A Journal of the Psychonomic Society, Inc., 12*(4), 703-710.

Kwantes, P. Derbentseva, N. Lam, Q. Vartanian, O. & Marmurek, H. (2016). Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences, 102*, 229-233.

Landauer, T. K., Laham, R. D. & Foltz, P. W. (2003). Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In M. Shermis & J. Bernstein, (Eds.), *Automated Essay Scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Landauer, Thomas K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review, 104*(2), 211-40.

Lee, J.G., Jun, S., Cho, Y.W., Lee, H. Guk, B., Seo, J.B., Kim, N. (2017). Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology*, *18*(4), 570–584.

Levy, D.L., Coleman, M.J., Sung, H., Ji, F., Matthysse, S., Mendell, N.R., & Titone, D. (2010). The genetic basis of thought disorder and language and communication disturbances in schizophrenia. *Journal of Neurolinguistics*, *23*(3), 176-192.

Lund, K., & Burgess, C. (1996). Hyperspace analogue to language (HAL): A general model semantic representation. *Brain and Cognition*, *30*(3), 5.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction, 21*, 251–284.

Martin, D. I., & Berry, M. W. (2011). Mathematical Foundations Behind Latent Semantic Analysis. In T. K. Landuaer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis,* (pp. 35-55). Mahwah, NJ: Laurence Erlbaum and Associates.

McClelland, D. C. (1979). Inhibited power motivation and high blood pressure in men. *Journal of Abnormal Psychology, 88*, 182-190.

McKeon, R. (2001). *The basic works of Aristotle* (Modern Library Pbk. ed., Modern Library classics). New York: Modern Library.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN, US: University of Minnesota Press.

Meyer, A. N., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA internal medicine*, 173, 1952-1958.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63, 81.

Millis, K., Magliano, J., Wiemer-Hastings, K. Todaro, S., & McNamara, D. S. (2007). Assessing and Improving Comprehension with Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 121–141). Mahwah, NJ: Laurence Erlbaum and Associates.

Morgan, C., Coleman, M., Ulgen, A., Boling, L., Cole, J., Johnson, F., . . . Levy, D. (2017). Thought Disorder in Schizophrenia and Bipolar Disorder Probands, Their Relatives, and Nonpsychiatric Controls. *Schizophrenia Bulletin*, *43*(3), 523-535.

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education*, *41*, 1140-1145.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*, 104-114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Review*, 49, 197–237.

Pennebaker, J. W., & Beall, S. K. (1986). Confronting a traumatic event: Toward an understanding of inhibition and disease. *Journal of Abnormal Psychology, 95*(3), 274-281.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks, 6*, 623–641.

Rorschach, H. (1921). *Psychodiagnostik*. Leipzig, Germany: Ernst Bircher Verlag.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*(4), 532-547.

Rosch, E. Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Rumelhart, D., McClelland, J. L, & University of California, San Diego. PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, Mass.: MIT Press.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Spengler, P., White, M., Ægisdóttir, S., Maugherman, A., Anderson, L., Cook, R., . . . Rush, J. (2009). The Meta-Analysis of Clinical Judgment Project. *The Counseling Psychologist, 37*(3), 350-399.

Strang, G. (1998). *Introduction to Linear Algebra*. Wellesley, MA: Wellesley Cambridge Press.

Tausczik, Y., & Pennebaker, J. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

Turney, P., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research, 37*, 141-188.

Vapnik V. (2010). *The Nature of Statistical Learning Theory*. New York, NY: Springer New York.

Weintraub, W. (1981). *Verbal behavior: Adaptation and psychopathology*. New York: Springer.

Weintraub, W. (1989). *Verbal behavior in everyday life*. New York: Springer.

Willits, J. A., Rubin, T., Jones, M. N., Minor, K. S., & Lysaker, P. H. (in press). Evidence of disturbances of deep levels of semantic cohesion within personal narratives in schizophrenia. *Schizophrenia Research.*

Appendix A

K10 questionnaire

During the last 30 days:

1. about how often did you feel tired out for no good reason?
2. about how often did you feel nervous?
3. about how often did you feel so nervous that nothing could calm you down?
4. about how often did you feel hopeless?
5. about how often did you feel restless or fidgety?
6. about how often did you feel so restless you could not sit still?
7. about how often did you feel depressed?
8. about how often did you feel that everything was an effort?
9. about how often did you feel so sad that nothing could cheer you up?
10. about how often did you feel worthless?

Center for Epidemiologic Studies Depression Scale (CES-D), NIMH

During the past week:

1. I was bothered by things that usually don't bother me.
2. I did not feel like eating; my appetite was poor.
3. I felt that I could not shake off the blues even with help from my family or friends.
4. I felt I was just as good as other people.
5. I had trouble keeping my mind on what I was doing.
6. I felt depressed.
7. I felt that everything I did was an effort.
8. I felt hopeful about the future.
9. I thought my life had been a failure.
10. I felt fearful.
11. My sleep was restless.
12. I was happy.
13. I talked less than usual.
14. I felt lonely.
15. People were unfriendly.
16. I enjoyed life.
17. I had crying spells.
18. I felt sad.
19. I felt that people dislike me.
20. I could not get "going."

PROMIS Short Form v1.0 - Depression 8b

In the last 7 days:

1. I felt worthless
2. I felt that I had nothing to look forward to
3. I felt helpless
4. I felt sad
5. I felt like a failure
6. I felt depressed
7. I felt unhappy
8. I felt hopeless

PROMIS Short Form v1.0 - Anxiety 8a

In the last 7 days:

1. I felt fearful
2. I found it hard to focus on anything other than my anxiety
3. My worries overwhelmed me
4. I felt uneasy
5. I felt nervous
6. I felt like I needed help for my anxiety
7. I felt anxious
8. I felt tense

Neuro-QoL Short Form v1.0 - Positive Affect and Well-Being

Lately:

1. I had a sense of well-being
2. I felt hopeful
3. My life was satisfying
4. My life had purpose
5. My life had meaning
6. I felt cheerful
7. My life was worth living
8. I had a sense of balance in my life
9. Many areas of my life were interesting to me

Appendix B

Building word vectors using Singular Value Decomposition (SVD)

The first step of building word vectors is to build a word-by-document matrix. Each row in the matrix represents each unique word in the corpus and each column in the matrix represents each document in the corpus. The table below shows a toy example with only five words and nine documents (D1-D9). The cells represent the number of times each word occurs in each document. For example, the word *depressed* occurs once each in D1 and D4, and the word *happy* appears once in each D8 and D9.

	D1	D2	D3	D4	D5	D6	D7	D8	D9
depression	1	0	0	1	0	0	0	0	0
depressed	1	0	1	0	0	0	0	0	0
anxiety	0	0	0	0	0	1	1	1	0
anxious	0	0	0	0	0	0	1	1	1
happy	0	0	0	0	0	0	0	1	1

Singular Value Decomposition decomposes the word-by-document matrix (A) into three more fundamental matrices.

U is a matrix of the eigenvectors of $\mathbf{AA}^T$:

0.00	-0.71	0.00	-0.71	0.00
0.00	-0.71	0.00	0.71	0.00
-0.58	0.00	0.77	0.00	0.27
-0.67	0.00	-0.26	0.00	-0.70
-0.47	0.00	-0.58	0.00	0.66

Σ is a diagonal matrix containing the square root of the eigenvalues of $\mathbf{AA}^T$ (termed singular values):

2.47	0.00	0.00	0.00	0.00
0.00	1.73	0.00	0.00	0.00
0.00	0.00	1.25	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.56

V is a matrix of the eigenvectors of $\mathbf{A}^T\mathbf{A}$:

0.00	-0.82	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
0.00	-0.41	0.00	0.71	0.00

0.00	-0.41	0.00	-0.71	0.00
0.00	0.00	0.00	0.00	0.48
-0.23	0.00	0.61	0.00	0.48
-0.50	0.00	0.41	0.00	-0.76
-0.69	0.00	-0.06	0.00	0.43
-0.46	0.00	-0.67	0.00	-0.06

The original word-by-document matrix (A) can be reconstructed by multiplying these three matrices

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$$

Latent Semantic Analysis (LSA) uses SVD to reconstruct a least squares best approximation of the original matrix with a smaller number of dimensions by using the **U** and **Σ** matrices.

$$\mathbf{X} = \mathbf{U}_{\mathrm{r}}\mathbf{\Sigma}_{\mathrm{r}}$$

where *r* denotes the number of dimensions in the least-squares approximation.

For example, by multiplying the first three columns of **U** by the first three columns and rows of **Σ** produces a matrix with the same number of rows as the original word-by-document matrix, but a smaller number of columns. The table below shows the least squares best approximation of the original word-by-document matrix with only three dimensions to represent each word (i.e., *r* = 3).

	Dimension 1	Dimension 2	Dimension 3
depression	0.00	-1.22	0.00
depressed	0.00	-1.22	0.00
anxiety	-1.43	0.00	0.96
anxious	-1.66	0.00	-0.32
happy	-1.15	0.00	-0.73

The effect of the procedure is best illustrated by looking at the correlations between words before and after applying SVD. In the original word-by-document matrix, the correlation between the two very semantically related words *depression* and *depressed* is only 0.36. However, after SVD is applied the correlations between *depression* and *depressed* is 1.00. As another example, the antonyms *depression* and *happy* is -0.28 before applying SVD, whereas after SVD the correlation increases in magnitude to -0.93.