SHapley Additive exPlanations Model Interpretability Analysis for Predictions based on Top2Vec Natural Language Processing

Anthony Robnett
Applied and Computational Mathematics Program
Johns Hopkins University
Baltimore, MD
arobnet1@jhu.edu

Dr. Cleon Davis
Electrical and Computer Engineering Program
Johns Hopkins University
Baltimore, MD
cleon.davis@jhu.edu

***Abstract* — Reddit is particularly an advantageous source of data for surveying and sampling purposes. Unlike most other prominent data sources where users usually interact, communicate, and share data from identifiable profiles, Reddit is most commonly used from an anonymous profile, allowing users a relatively comfortable privacy to discuss and share more vulnerable thoughts and ask more personal questions. This can be particularly useful those interested in investigating sensitive topics such as mental health. Organizations may want to investigate further the discussions of these topics and the relation these topics may have with other topics as a potential way to forecast the frequency of communication regarding a particular topic to better guide policies and actions. The purpose of this project is intended to apply the relatively new Topic to Vector (Top2Vec) topic modeling algorithm on a large data set of web scraped texts sampled from the comment sections of the United States Air Force subreddit page as an example. The example topic clusters most related to "Permanent Change of Station (PCS)" and "Depression" are to be analyzed using Correlation and Regression methods to investigate and compare between potential types of relations between comments from the designated "PCS" cluster to comments from the designated "Depression" cluster. Predictions from the regression model on the monthly occurrences of these cluster comments will then be analyzed for model interpretability using a SHapley Additive exPlanations Model (SHAP) to see which words stand out as the most contributing features to the prediction model from the scraped Reddit text sample data.**

***Index Terms*—Top2Vec, Topic Modeling, Correlation, Linear Regression, SHapley Additive exPlanations Model (SHAP), Cosine Similarity, Euclidean, Distance Metric, United States Air Force, Mental Health, Depression, Permanent Change of Station**

I Introduction

Despite the resiliency that service members are commonly known to have, being a subsection of the general population, military members evidently suffer from mental health issues no less than the general population. Especially in these years following the COVID pandemic, service members are dealing with many of the same, if not more, mental health struggles that many citizens have been dealing with.

As commonly known, in addition to being vulnerable to common general population mental health issues, they're particularly burdened by a unique set of struggles whether from deployment, high stress environments, and/or combat theaters. One particular event that tends to weigh on service members are Permanent Change of Stations (PCS). This is when a service member is required to (often involuntarily) uproot and move their lives to a new base of operations and living. This brings unique challenges that come with moving and adjusting to a new environment, that they're unfortunately not always able to leave from neither.

Every branch of the military employs and refers to psychological research to guide the specific policies applicable with their respective branches. To collect data for studies, wellness surveys are frequently administered from various offices within each branch. However, in order to contribute to the improvements of their mental health programs, the United States Department of Air Force (DAF) has been seeking out accessible and novel ways of exploratory analysis that may come from open-source social media outlets such as those like Reddit in order to potentially translate findings to actionable insights. This is where statistical and machine learning methods play a significant role. [27] [28] [29]

This paper will proceed to review the previously related works, proposed methods, and then will investigate results and analysis on if it is better to predict the monthly frequency of comments from one particular cluster such as "depression" by using the monthly frequency of comments from another particular cluster such as "PCS" or if it is better to use the monthly comment texts from a "PCS" cluster as the features themselves. The potential implications of this analysis will be explored in the conclusion sections of this paper. All referenced sources and Python code in Colab are cited at the end.

II RELATED WORK

Topic modeling on web scraped social media is nothing new considering that it is one of it's modern primary purposes. Topic modeling has also been commonly used for mental health based research such as to investigate topics discussed in a mental health internet support group or to discover online mental health-related communities. Topic modeling is even commonly used for mental health related text analysis on Reddit as well. [32] [33] [34]

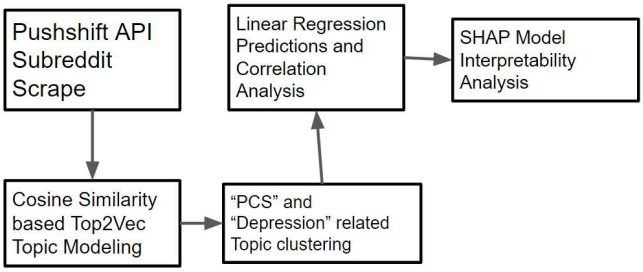
As far as topic modeling goes, Top2Vec is a relatively new topic modeling algorithm dating only back to 2020. Known for it's ease of implementation, it's quicly become a more prominent topic modeling approach. [35]

Because of the rise of prominence of artificial inelligence (AI) technology over the years becoming an integral part of our society, there has been a demand for explainable AI (XAI) in recent years so that users can have more transparency and understandability in the use of machine learning predictions. A prominent XAI is the SHapley Additive exPlanations (SHAP) algorithm. Rooted in game theory, SHAP uses machine learning to quantitatively evaluate the most contributing feature variables to a model's prediction. [36] [37]

An influential paper to this project that was released this year came from researchers in China who had performed a SHAP analysis on various regression models that made predictions based on various features from various text data sources in regards to online psychological health services. Some of those prediction features were clustered using a Neural Embedding-Based Topic Model. [31]

III PROPOSED METHODS

- Tools: Python, Pandas, Colab, Excel.



III.A Data Sourcing and Management: [40]

The data in this study is web scraped from comment conversations in the subreddit called "r/AirForce". This subreddit is used members of the United States Air Force (USAF) and for Reddit users who are interested the USAF. Reddit users here can post about mutual interests, experiences, and questions among the USAF community as well as express grievances regarding any current events and trends. While branches of the military have analogous subreddit pages, the USAF was chosen for this research project due to my general familiarity from having worked with the Deparment of Air Force (DAF). [1]

1.a How the data was obtained/collected: [47]

The subreddit data was scraped from over an approximately 10 year span in order to be used for correlation, regression, and model interpretability analyses between the association of "Permanent Change of Station (PCS)" topic related comments and "depression" topic related comments. Two separate, non-overlapping data-set files with about 100,000 comments each were scraped for analysis using Pushshift Reddit API. [2] [47]

1.b Data Cleaning and Preprocessing: [41]

A natural language processing (NLP) pipeline is built for cleaning and preprocessing the text data. Data is cleaned by stripping newlines, tabs, HTML tags, links, whitespaces, accented characters, special characters, stopwords. Data is preprocessed by converting upper to lower case characters, reducing repeated characters and punctuations, expanding contractions, correcting mis-spelled words, lemmatizing the words, and stemming the words. [3]

III.B Topic Modeling

Topic modeling is the process of extracting underlying structures within a collection of texts in the form of statistical language models. This involves representing a text in a topic space rather than in its feature space in a process called dimensionality reduction. By building clusters of words in the form of topics that make up a text with each topic cluster having a particular weight using an Unsupervised Learning process, the information within abstract topic clusters can be best represented from the document collection using a tagging process. [4]

B.1 Top2Vec Algorithm

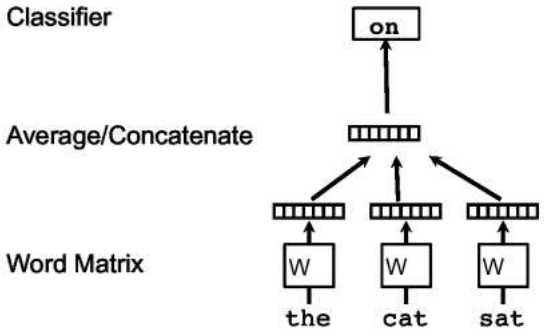
Top2Vec expresses the semantic similarity from the distance between document and word vectors that the topic vectors are jointly embedded with by using joint document and word semantic embedding to find topic vector clusters. Top2Vec can automatically find the number of topics without the use of stop-word lists, stemming, or lemmatization. The Top2Vec algorithm comprises the following steps: [5]

1.a Create the jointly embedded document and word vectors using Doc2Vec.

In this step, the algorithm generates the embedding where the distance between document vectors and word vectors represents semantic association. A Doc2Vec model is used to learn document embeddings to estimate a distributed representation of documents using a model architecture called a Paragraph Vector Distributed Bag of Words (PV-DBOW). As a common fixed-length vector representation for texts within machine learning algorithms such as k-means or logistic regression, Bag-of-Words (BOW) can be used for clustering and text classification functions that utilizes applications such as web searching, spam filtering, and document retrieval.

Paragraph Vectors come from the unsupervised learning of fixed-length feature representations of variable-length pieces of texts including sentences, paragraphs, and documents. The unsupervised learning algorithm is trained to predict words in the document in order to represent each document as a dense vector. Paragraph vector learning comes from word vector learning methods such as Word2Vec. In the example figure here, the context from the three words "the," "cat," and "sat" is used to predict the

fourth word “on”, which is mapped to the matrix W columns for the output word prediction. [5] [6]



[6]

In the Word2Vec learning framework, every word is mapped to a unique vector in the form of a column within matrix *W*. By indexing the column by the word position within the vocabulary, the concatenation takes the sum of the vectors to use as features to predict the next word in a sentence. The word vector mode is supposed to maximize the average log probability given a sequence of training words.

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p\left(w_t | w_{t-k}, \ldots, w_{t+k}\right) \quad [6]$$

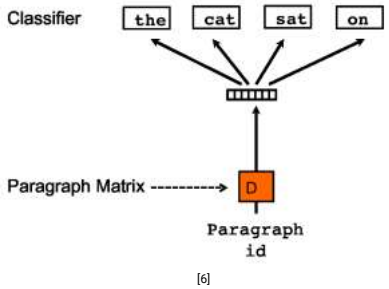
A multi-class classifier, such as Softmax is used for the prediction task.

$$p\left(w_t | w_{t-k}, \ldots, w_{t+k}\right) = \frac{e^{y_{w_i}}}{\sum_i e^{y_i}} \quad [6]$$

The value of each y_i represents an un-normalized log-probability for each output word i, where U and b are the softmax parameters and h is generated from the average of word vectors extracted from W.

$$y = b + Uh\left(w_{t-k}, \ldots, w_{t+k}; W\right) \quad [6]$$

PV-DBOW trains paragraph vectors to predict words in a small window, ignoring the word context from the input to predict randomly sampled words from the paragraph output. The classifier samples a random word from a randomly sampled text window at each stochastic gradient descent iteration given to the paragraph vector.



[6]

The surrounding word in a context window is predicted by DBOW swapping the context word with the document vector. This similarity allows for joint embedding, which occurs when the simultaneous training of document and word vectors become interleaved. A dense area of documents within a semantic space can be interpreted as an area of highly similar documents indicative of an underlying common topic. The topics of the documents represented by the document vectors can have the average of those vectors calculated as the centroid, which is most representative of the dense area of documents it’s calculated from. The topic vector gives the best semantic description of the words closest to it, with each number of dense clusters representing a number of prominent topics. [6]

1.b Create lower dimensional embedding of document vectors using UMAP.

The *Uniform Manifold Approximation and Projection* (UMAP) algorithm is a manifold learning technique for dimension reduction that approximates the manifold that the data is assumed to lie on. The geodesic distance on the manifold can be approximated by letting the input data be $X = \{X_1,..., X_N\}$, meaning that a ball centered at X_i with exactly the k-nearest-neighbors of X_i should have approximately fixed volume regardless of the choice of $X_i \in X$. The validity of assumption of uniform distribution on the manifold can be ensured with a custom distance for each X_i. The functors between the relevant categories to convert from metric spaces to fuzzy topological representations provide a means to merge the incompatible local views of the data.

Letting $Y = \{Y_1,..., Y_N\}, \subseteq \mathrm{R}^d$ be a low dimensional ($d << n$) representation of X such that Y_i represents the source data point X_i, a target manifold for Y can be chosen from the former, so that the manifold and manifold metric from the former can be known to directly compute the fuzzy topological representation. To incorporate the distance to the nearest neighbor for the local connectivity requirement, a parameter can be supplied that defines the expected distance between nearest neighbors in the embedded space. [8]

For the Top2Vec algorithm, the number of nearest neighbors is set to 15 to give more emphasis on local structure. The distance metric to measure the distance between points in the high dimensional space is the cosine similarity to measure the similarity of documents irrespective of their size. [5]

1.c Find dense areas of documents using HDBSCAN.

Hierarchical Density-based spatial clustering of applications with noise (HDBSCAN) assigns a label for each dense cluster of document vectors and assigns a noise label to all document vectors that are not in a dense cluster so that the dense areas of identified document vectors can be used to calculate the topic vectors. Documents that are classified as noise are considered not to be descriptive of a prominent topic. With a minimum cluster size of 15, larger values have a higher chance of merging unrelated document clusters for better results. [5]

HDBSCAN main steps:

1. Compute the core distance for all data objects.
2. Compute the mutual reach-ability distance graph.

3. Add for each vertex a “self edge” with the core distance of the corresponding object as a weight.
4. Extract the HDBSCAN hierarchy as a dendrogram diagram representing a tree:
    - 4.1 The tree assigns all objects the same label for the root of the single cluster.
    - 4.2 Iteratively remove all edges in decreasing order of weights: [8]
        - 4.2.1 Set the dendrogram scale value of the current hierarchical level as the weight of the edges to be removed before each removal.
        - 4.2.2 Assign labels to the connected components that contain the end vertices of the removed edges, to obtain the next hierarchical level and assign a new cluster label to a component if it still has at least one edge, or assign it a null noise label after each removal.

1.d Calculate the centroid of document vectors in the original dimension for each dense area to find the topic vector.

The dense document clusters and noise documents identified by HDBSCAN in the UMAP reduced dimension, correspond to locations in the original semantic embedding space so that each document in the embedding space can be labeled with either a noise label or a dense cluster label. The topic vector can be calculated from the document vectors similarly to how the labels for each dense cluster of documents can be calculated. The centroid can be calculated by finding the arithmetic mean of all the document vectors in the same dense cluster. [5]

1.e Find n-closest word vectors to the resulting topic vector.

Every point in the semantic space represents a topic best described semantically by its nearest word vectors that are closest to a topic vector most representative of semantically where the distance between each word vector to the topic vector indicates how semantically similar the word is to the topic. The words closest to the topic vector can be considered as the words most similar to all documents in that dense area, with the topic vector being the centroid of that area to summarize the common topic of the documents in the dense area. Common words appearing in most documents are often in a region of the semantic space equally distant from all documents so the words closest to a topic vector should rarely be stop-words, so there shouldn’t be a need for stop-word removal. [5]

B.2 Metric Distance Modified Top2Vec Model

Because by default, Top2Vec uses Cosine Similarity as its distance metric in order to evaluate cluster density and semantic similarity between the word and document vectors, for comparative purposes, this study will briefly explore the topic clustering results for when the internal UMAP argument used for this function uses the Euclidean distance instead. [11] [12]

III.C Correlation Analysis

C.1 Monthly PCS Comment Frequency to correlate with Monthly Depression Comment Frequency:

1.a Autocorrelation:

When taking the autocorrelation of both the PCS and depression monthly comment frequency variables, the data within both of the column vectors can be compared, for example, with the lag = 1 to lag = 2, meaning scaling the data by it's regular time proportions (data(t)) and comparing it to itself with lagging time proportions (data(t-1)).

Because of this, it is generally said that the total number of observations (T) should be less than or equal to 50. The greatest lag value (k) should be less than or equal to T/k.

With T = 74 observations in the current dataframe, at k = 8 lags, T/k = 74/8 = 9.25 > 8. So the first 8 values could be considered. In this instance, lags at 4 and 8 shifts are used for autocorrelation. [17] [18] [19]

C.2 Numerically Vectorized Monthly PCS Comment Texts to correlate with Monthly Depression Comment Frequency:

For finding the correlation of the monthly PCS comments in the form of numerically vectorized texts to be use each word as a feature variable predictor of the monthly frequency for depression comments, in the case of a multiple regression, the coefficient of multiple correlation is always positive and is simply the square root of R-squared. [50]

III.D Regression Analysis

D.1 Using Linear Regression of Monthly PCS Comment Frequency to predict Monthly Depression Comment Frequency:

A linear regression model will be used to train the Monthly PCS Comment Frequency along with it’s 4 and 8 month lag shift as the independent variables to predict the Monthly Depression Comment Frequency as the dependent variable. The data will be split into 80% train and 20% test sets. After creating a regression object, the linear regression model will be trained and fitted to generate regression coefficients, variance scores, and to plot for residual errors.

D.2 Using Multiple Linear Regression of Numerically Vectorized Monthly PCS Comment Texts to predict Monthly Depression Comment Frequency:

2.a Feature Extraction:

The scikit-learn function CountVectorizer can be used to convert a collection of text documents to a matrix of token counts. [51] In order to perform the text analysis necessary within some machine learning algorithms, scikit-learn provides utilities for common ways to extract numerical features from text content because the raw data as a sequence of symbols cannot be fed directly to the algorithms themselves. Most algorithms expect numerical feature vectors with a fixed size instead of the raw text documents with varying lengths.

A corpus of documents can be represented with a matrix with one row per document and one column per token in the corpus. Vectorization is the process of turning a collection of text documents into numerical feature vectors using

tokenization, counting, and normalization as the Bag of Words representation. Documents can be described by word occurrences while ignoring the relative position information of the document words. [52]

2.b Multiple Linear Regression:

For this linear regression that was intended for SHAP analysis, the data frame used throughout analysis up until this point needed to be preprocessed again by not only separating the Monthly PCS comment topic counts from the Monthly depression topic comment counts, but the texts their selves needed to be separated as well in order to be input into a Count Vectorizer function. This is so the text data can be converted into numerically representative fixed length vectors to be processed by regression and shap analysis algorithms. In order to prevent poor shap results that would display excessive zeros, missing data for months with no comments had to have zeros replaced by "Nan" until after monthly aggregation counts, at which point zero string data types could be input back into the missing data points in order to be processed by a vectorizer, regression, and then SHAP function. [19] [48]

III.E Model Interpretability Analysis

E.1 SHAP

SHapley Additive exPlanations (SHAP) uses game theory to interpret the output of a machine learning model by connecting optimal credit allocation with local explanations by using Shapley values, which are commonly used for cooperative game theory. Shapley value based explanations of machine learning models uses these fair allocation results to allocate the contribution value for a model's output among its input features. Therefore, function $f(x)$ must match a model's input features with players in a game and match the model function with the rules of the game. As done in game theory, a feature can be defined to have joined a model when the value of that feature is known, and that it hasn't joined a model when the value of that feature is not known. An existing model can be evaluated when only a subset of features are part of the model by integrating out the other features with a conditional expected value formulation.

$$E[f(X) \mid X_S = x_S]$$

Shapley values will always sum up to the difference between the game outcome with all players present and the game outcome when no players are present. Therefore machine learning models will have SHAP values of all the input features sum up to the difference between the expected model output and the current model output to explain the prediction and this can be explained with a waterfall plot.

Explainers can be used for the permutation definition of SHAP values with the non-interventional conditional expectation form having $f(x) = \beta x + b$ where β is a row vector and b is a scalar.

$$\phi_i = \frac{1}{M!}\sum_{R} E\left[f(x)|x_{S_i^R \cup i}\right] - E\left[f(x)|x_{S_i^R}\right]$$

This can be used to derive the Shapley equation.

$$\phi = \beta T x.$$

The transform matrix T can be computed by iterating many random permutations of R averaging the results in order to explain a number of samples with the use of matrix multiplication. A linear model can be made for the SHAP value of feature i for the prediction $f(x)$.

$$\phi_i = \beta_i (x_i - E[x_i]).$$

The units of the SHAP values can be explained with a logistic regression model in the log-odds space. [19] [20] [21] [22] [23] [24] [25] [26]

IV RESULTS AND ANALYSIS

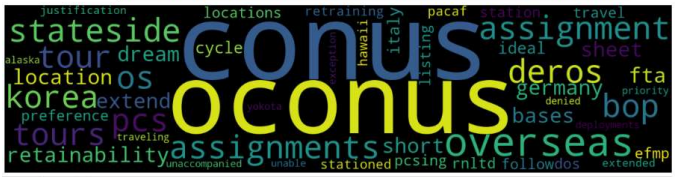
IV.A Topic Model [42] [49]

To build the initial Top2Vec Model, it was found that most of the default parameters worked well to general qualitatively recognizable topic cluster word clouds, with the speed of the modeling parameter set to "deep-learn" to optimize topic vector quality despite slower processing speeds. [9] [10]

To ensure that the Top2Vec model was clustering jointly embedded word and document vectors in a sensible manner, a qualitative assessment step was utilized that would use search keywords. "PCS" and "depression" were the chosen keywords given that they were to be used for time series, correlation, regression, and model interpret ability analysis in the later steps. With the recognizable associated terms in the word clusters, it was assumed that the comment documents making up the topic clusters with the highest cosine similarities to "PCS" and "depression" were adequate to proceed with further data analysis.

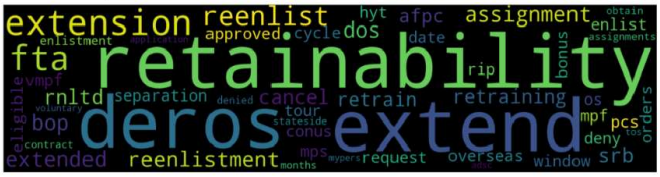
- Output: [38]

Topic 530 **"PCS" Topic Cluster Word Cloud from 1st Subreddit Scrape.**



Topic 192 **"Depression" Topic Cluster Word Cloud from 1st Subreddit Scrape.**

Topic 144 **"PCS" Topic Cluster Word Cloud from 2nd Subreddit Scrape.**



Topic 512 **"Depression" Topic Cluster Word Cloud from 2nd Subreddit Scrape.**

- The topic model from the 1st data set generated 902 topic clusters. The topic model from the 2nd data set generated 871 topic clusters.
- First data set clustering results:
  - For the cluster similar to the "depression" keyword of the first data set, the recognizable associated terms included 'meds', 'medication', 'diagnosed', 'doctor', 'pcm', 'anxiety', etc.
    - The top cosine similarity values here included 0.7172813, 0.70858824, 0.67225444, 0.6652465 , 0.6581058, etc.
  - For the cluster similar to the "PCS" keyword, recognizable associated terms included 'listing', 'os', 'ams', 'sheet', 'assignments', etc.
    - The top cosine similarity values here included 0.6059564 , 0.5679201 , 0.56763184, 0.5229083 , 0.51234925, etc.
- Second data set clustering results:
  - For the cluster similar to the "depression" keyword of the first data set, the recognizable associated terms included 'adhd', 'meds', 'medication', 'diagnosed', 'anxiety', etc.
    - The top cosine similarity values here included 0.6396215, 0.6251423 , 0.62161577, 0.56418055, 0.54438996, etc.
  - For the cluster similar to the "PCS" keyword, recognizable associated terms included 'retainability', 'bop', 'fta', 'deros', 'extend', etc.
    - The top cosine similarity values here included 0.6429174 , 0.60920167, 0.5860033 , 0.576714 , 0.5754652, etc.

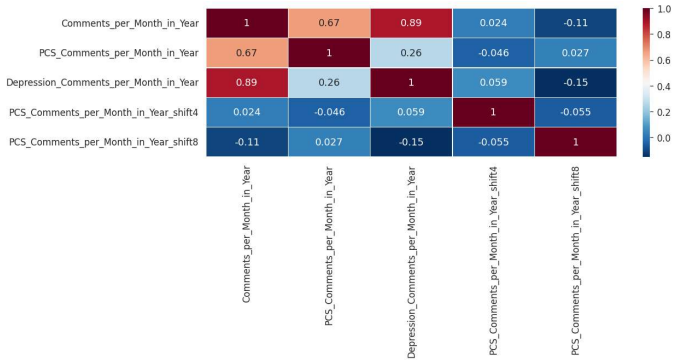
A count of the permanent change of station and depression related topics was generated using an Enumerate function. After the comment documents were aligned with the dates in UTC format, the data was input into a categorical data-frame to be used as a data file for analysis containing only "PCS" and "depression" related comments.

IV.B Analyzing Correlation [44]

- If X_t is the realization value from the run of a process at time t, the mean and variance at time t, for each t, holds the auto-correlation definition between t_1 and t_2 with the following formula:
  - $R_{XX}(t_1,t_2) = E\left[X_{t_1} \bar{X}_{t_2} \right]$ [55]

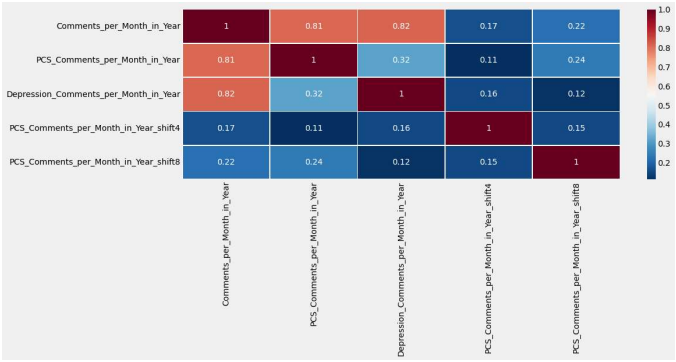
B.1 Autocorrelation Calculation analysis [44]

1.a 1st data set Correlations of Monthly Topic Comment Frequencies.



- In the figure above, both variables show each with a Pearson's correlation coefficient with a positive value at lag = 4 changing in tandem while the Pearson's correlation coefficient with a negative value at lag = 8 is changing inversely. However, because both variables show both lag value coeffiecients to be close to zero, there's little to no autocorrelation.
- Additionally, the heatmap indicates that the direct correlation between the monthly PCS comment frequency to the monthly Depression comment frequency is only 26%, which is also weak and an indication of being a weak predictor.

1.b 2nd data set Correlations of Monthly Topic Comment Frequencies.



- In the figure above, both variables show each with a Pearson's correlation coefficient with a positive value at all lags. However, both variables show both lag value coeffiecients to be close to zero here as well, indicating that there's little to no autocorrelation.
- Additionally, the heatmap indicates that the direct correlation between the monthly PCS comment frequency to the monthly Depression comment frequency here is only 32%, which is also weak and an indication of being a weak predictor.

IV.C Predicting with Regression [44] [45]

C.1 Simple Linear Regression of Monthly PCS Comment Frequency to predict Monthly Depression Comment Frequency: [44]

- The relationship between the true, but unobserved, underlying parameters α and β and the data points is called a linear regression model.
  - $y_i = \alpha + \beta x_i + \epsilon_i$ [56]
- 1st data set of simple linear regression for prediction.
  - Coefficient of Determination:
    - The coefficient of determination here is defined as a function of *u,* the residual sum of squares and *v,* the total sum of squares.
      - $R^2 = \left(1 - \frac{u}{v}\right)$ [57]
        - $= 0.1409176126596482.$
  - As expected from the correlation, the coefficient of determination score is weak at approximately 14% percent above displayed a particularly weak regression indicating no predictability of the Monthly Depression Comment Frequency from the Monthly PCS Comment Frequency or from it's 4 month and 8 month lag shifts. [44]
- 2nd data set of simple linear regression for prediction.
  - Coefficient of Determination:
    - $R^2 = \left(1 - \frac{u}{v}\right)$
      - $= 0.05065782314103351.$
  - As observed with the prior data set, the weak correlation indicated a weak regression prediction.

C.2 Multiple Linear Regression of Numerically Vectorized Monthly PCS Comment Texts to predict Monthly Depression Comment Frequenc

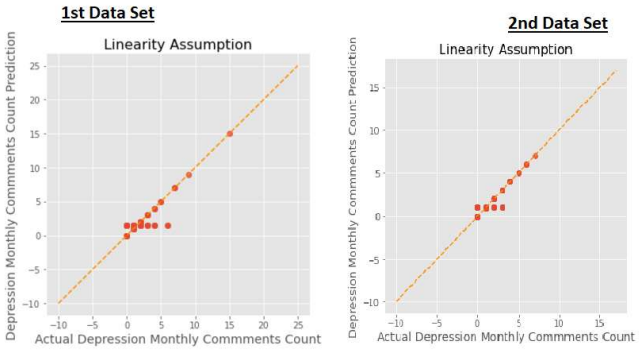
- For the Numerically Vectorized Text based model, it can be validated using residual analysis based on a test or assumption.
  - Linearity:
    - This assumes that there is a linear relationship between the independent variables (Numerically Vectorized Text based on Monthly PCS comments) and the dependent variable (Monthly frequency of Depression comments). The first five rows of the data set are displayed as an example.

	Monthly_PCS_Comments_Combined	Depression_Comments_per_Month_in_Year
0	'f[redacted] dead . went x ray , saw external had ...	1
1	0	0
2	`` op . lot shift replies thread . let please ...	1
3	' would take either want instructor , job woul...	3
4	' month thing seen firetruck called 335 cause ...	1

- In this case since, there are multiple independent variables in the form of each word feature of a comment. The first five rows of the data set are displayed as an example.

	0	1	2	3	4	5	6	7	8	9	...	1114	1115	1116	1117	1118	1119	1120	1121	1122	Depression_Comments_per_Month_in_Year
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	1
3	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

- This data, vectorized with each word serving as an individual predictor, a multiple linear regression model is to be used for prediction.
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$
- A scatter plot can be used to compare the predicted values versus the actual values.



- The scatter plot for both data set plots show a spread of the actual data points versus the prediction of the response variable reasonably evenly around the diagonal line, so it may be possibly assumed that there is linear relationship between the independent and dependent variables.

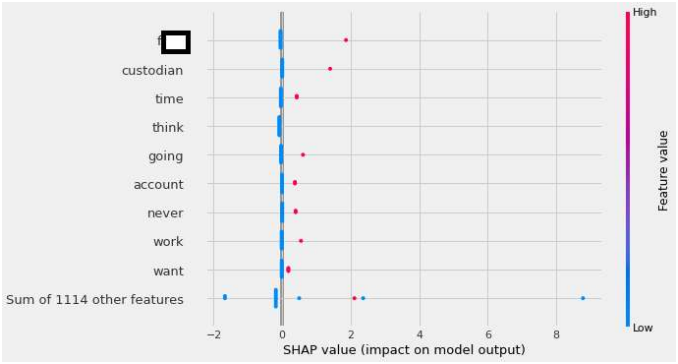
- Coefficient of Determination:
  - $R^2 = c^T R^{-1}_{xx} c$
    - 1st data set:
      - $= 0.8363137682933433.$
    - 2nd data set:
      - $= 0.8086457694982832$
- Coefficient of Multiple Correlation: [54]
  - $R_{xx} = \begin{pmatrix} r_{x_1 x_1} & r_{x_1 x_2} & \cdots & r_{x_1 x_N} \\ r_{x_2 x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ r_{x_N x_1} & \cdots & & r_{x_N x_N} \end{pmatrix}$
    - 1st data set:
      - $= 0.9145019236138016.$
    - 2nd data set:
      - $= 0.8992473349964865.$

- The corresponding Coefficient of Determination score at approximately 83% and Coefficient of Multiple Correlation score at approximately 91% indicate a substantially stronger regression coming from the numerically vectorized words as features to predict the Monthly Depression Comment Frequency. [45]

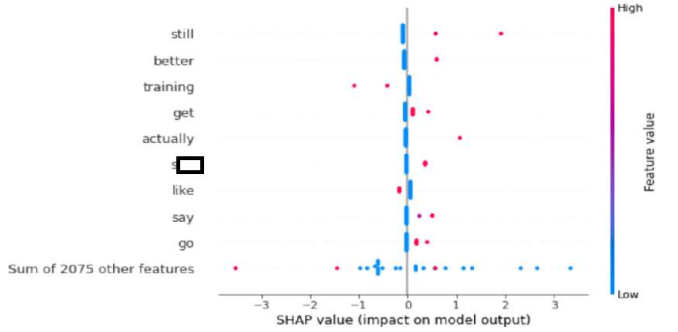
IV.D Interpreting Model with SHAP [45]

D.1 Beeswarm plot of Shap values. [45]

- 1st Data Set:

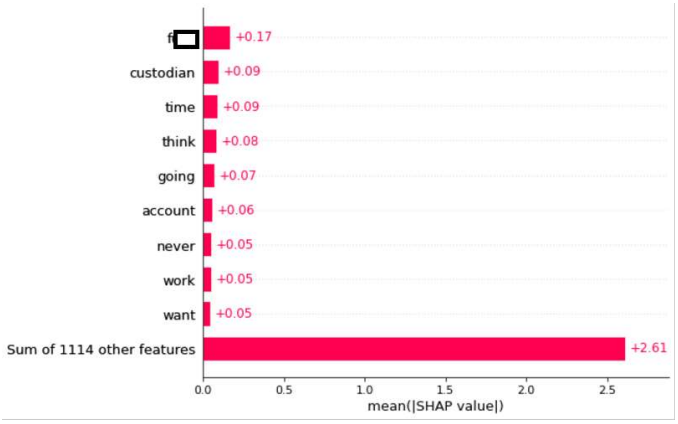


- 2nd Data Set:

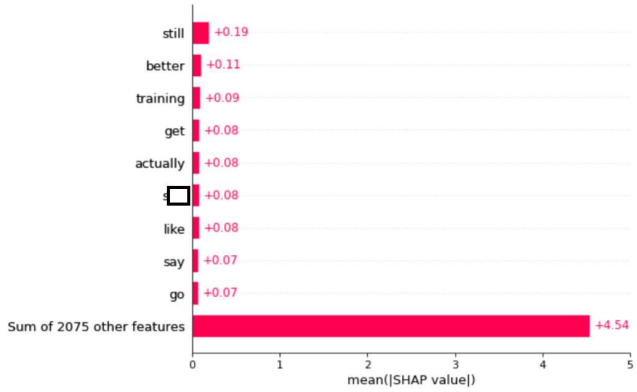


D.2 Bar plot of the mean of SHAP values.

- 1st Data Set:



- 2nd Data Set:



- The SHAP values plot the feature value importance by ranking words (vertically) color coding the months with the highest comments with brighter colors indicating a higher frequency of mentions within the month. The higher the absolute value of SHAP value scores (horizontally), the more significant of a contribution the word makes on impacting the prediction of the regression model.
- It's important to keep in mind here that the words from the prediction here are distinct from the word clouds in topic model due to Top2Vec clustering words based on Semantic Cosine Similarity, and the SHAP analysis ranking words are based upon contribution to the Linear Regresionn.
- Here, it looks like while words like "custodian" have the highest impact due to elevated frequency occurences within a particular month, the word "think" has a high rank with a negative shap absolute value due to it being the only negative value indicating that the lack there of the term in particular months contributed relatively significant to the prediction of the regression model. Between "PCS" and "Depression", the feature words of high ranking importance aren't particularly discernible of any factors of demonstrable impact from the prediction that an organization can immediately make an actionable decision from. However, from the use of substantially larger data sets based on predictor and target clusters to be topic modeled, scraped, and aggregated from multiple text data source types, a more thorough and targeted word text filter could be used to collect SHAP value information on more qualitatively recognizable words to make better sense of a prediction.

V CONCLUSION AND FUTURE WORKS

All in all, it was found between both of the data sets that rather than using the monthly frequency of a particular cluster to predict the monthly frequency of another cluster, it's a substantially better predictor to use the individual words of the monthly comments as feature variables to predict the monthly frequency of another cluster.

While the SHAP analysis of that same prediction rendered words that weren't particularly insightful. The SHAP model seemed to be working effectively, and could have much more utility if used with more data and a proper word filter.

In the future, it would be ideal to complete more iterations of experiment trials using methods such as conducting regression analyses on different time lag shifted variables, sourcing data from multiple data sources like from Reddit post titles or multiple different social media sites. Considering the complications with larger data-sets, it would've been useful to comparing the results from different topic models as well that may have been more equipped for that scale of data. Most of all, it would be ideal to complete a full set of different time series, correlation, and regression analyses on different distance metrics based topic models in order to be able to propose a novel distance metric equation. Additionally, it would be interesting to study the computational efforts of these models when used together potentially by bootstrapping a data sample and running an optimization model to find a balance between computational cost and algorithm accuracy. That being said, the most personally beneficial part of this research project is the extensive data Preprocessing pipeline that can now be used to more easily approach any of the above mentioned methods should the need arise.

ACKNOWLEDGMENT

I would like to give acknowledgment to my tutors Christian M. and Allen C. for assisting me in debugging my code and troubleshooting my algorithms.

REFERENCES

1 "R/Airforce." Reddit, https://www.reddit.com/r/AirForce/.

2 Podolak, Matt. "How to Scrape Large Amounts of Reddit Data." Medium, The Startup, 9 Apr. 2021, https://medium.com/swlh/how-to-scrape-large-amounts-of-reddit-data-using-pushshift-1d33bde9286.

3 Yadav, Kajal. "Cleaning & Preprocessing Text Data by Building NLP Pipeline." Medium, Towards Data Science, 8 Nov. 2022, https://towardsdatascience.com/cleaning-preprocessing-text-data-by-building-nlp-pipeline-853148add68a.

4 Kapadias. "Mediumposts/Introduction to Topic Modeling.ipynb at Master · Kapadias/Mediumposts." GitHub, 29 Dec. 2020, https://github.com/kapadias/mediumposts/blob/master/natural_language_processing/topic_modeling/notebooks/Introduction%20to%20Topic%20Modeling.ipynb.

5 Angelov, Dimo. "Top2vec: Distributed Representations of Topics." ArXiv.org, 19 Aug. 2020, https://arxiv.org/abs/2008.09470.

6 Le, Quoc V., and Tomas Mikolov. "Distributed Representations of Sentences and Documents." ArXiv.org, 22 May 2014, https://arxiv.org/abs/1405.4053.

7 McInnes, Leland, et al. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." ArXiv.org, 18 Sept. 2020, https://arxiv.org/abs/1802.03426.

8 Campello, Ricardo J. G. B., et al. "Density-Based Clustering Based on Hierarchical Density Estimates." SpringerLink, Springer Berlin Heidelberg, 1 Jan. 1970, https://link.springer.com/chapter/10.1007/978-3-642-37456-2_14.

9 Ddangelov. "DDANGELOV/Top2Vec: Top2vec Learns Jointly Embedded Topic, Document and Word Vectors." GitHub, https://github.com/ddangelov/Top2Vec#top2vec.

10 "Welcome to top2vec's Documentation!" Welcome to Top2Vec's Documentation! - Top2Vec 1.0.27 Documentation, https://top2vec.readthedocs.io/en/latest/index.html.

11 "UMAP API Guide." UMAP API Guide - Umap 0.5 Documentation, https://umap-learn.readthedocs.io/en/latest/api.html#umap.

12 "Sklearn.metrics.DistanceMetric." Scikit, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.DistanceMetric.html.

13 Anish, Athul. "Time Series Analysis." Medium, The Startup, 25 Nov. 2020, https://medium.com/swlh/time-series-analysis-7006ea1c3326.

14 6.4.4.2. Stationarity, https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm#:~:text=Stationarity%20can%20be%20defined%20in,no%20periodic%20fluctuations%20(seasonality).

15 Jacob_s. "Time Series Forecast : A Basic Introduction Using Python." Medium, Medium, 13 Nov. 2017, https://medium.com/@stallonejacob/time-series-forecast-a-basic-introduction-using-python-414fcb963000.

16 Ferus, Jacob. "The Powerful Feature Extraction Method You've Never Heard Of." Medium, Medium, 17 Oct. 2022, https://medium.com/@dreamferus/the-powerful-feature-extraction-method-youve-never-heard-of-1e960483e709.

17 "1.1. Linear Models." Scikit, https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

18 "Pandas.dataframe.corr#." Pandas.DataFrame.corr - Pandas 1.5.2 Documentation, https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html.

19 Dotis, Anais. "Autocorrelation in Time Series." Medium, Medium, 3 May 2019, https://dganais.medium.com/autocorrelation-in-time-series-c870e87e8a65.

20 "Linear Regression (Python Implementation)." GeeksforGeeks, 22 Aug. 2022, https://www.geeksforgeeks.org/linear-regression-python-implementation/.

21 Molnar, Christoph. "Interpretable Machine Learning." 9.5 Shapley Values, 14 Dec. 2022, https://christophm.github.io/interpretable-ml-book/shapley.html.

22 Molnar, Christoph. "Interpretable Machine Learning." 9.6 SHAP (SHapley Additive ExPlanations), 14 Dec. 2022, https://christophm.github.io/interpretable-ml-book/shap.html.

23 "Welcome to the Shap Documentation." Welcome to the SHAP Documentation - SHAP Latest Documentation, https://shap.readthedocs.io/en/latest/index.html.

24 "An Introduction to Explainable AI with Shapley Values." An Introduction to Explainable AI with Shapley Values - SHAP Latest Documentation, https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html.

25 "Math behind Linearexplainer with Correlation Feature Perturbation." Math behind LinearExplainer with Correlation Feature Perturbation - SHAP Latest Documentation, https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/linear_models/Math%20behind%20LinearExplainer%20with%20correlation%20feature%20perturbation.html.

26 "Sentiment Analysis with Logistic Regression." Sentiment Analysis with Logistic Regression - SHAP Latest Documentation, https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/linear_models/Sentiment%20Analysis%20with%20Logistic%20Regression.html.

27 RAND Corporation Provides Objective Research Services and Public Policy ... https://www.rand.org/content/dam/rand/pubs/research_reports/RR2300/RR2304/RAND_RR2304.pdf.

28 Higgins Neyland, M K, et al. "Permanent Change of Station Moves and Disordered-Eating Attitudes and Behaviors in Prevention-Seeking Adolescent Military-Dependents." Eating Behaviors, U.S. National Library of Medicine, Jan. 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7906934/.

29 Military.com | By Lori Stoffers. "The 5 Emotions of PCSing." Military.com, 15 June 2021, https://www.military.com/spousebuzz/blog/2016/09/why-didnt-you-tell-me-the-5-emotions-of-pcsing.html.

30 Asonam 2022 - the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, https://asonam.cpsc.ucalgary.ca/2022/.

31 Liu, Hui, et al. "Prediction of Online Psychological Help-Seeking Behavior during the COVID-19 Pandemic: An Interpretable Machine Learning Method." Frontiers in Public Health, U.S. National Library of Medicine, 3 Mar. 2022, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8929708/.

32 Carron-Arthur B;Reynolds J;Bennett K;Bennett A;Griffiths KM; “What's All the Talk about? Topic Modelling in a Mental Health Internet Support Group.” BMC Psychiatry, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/27793131/.

33 O'Neil, Emma. “Mental Health Support on Reddit.” Medium, Medium, 5 May 2021, https://oneilemma.medium.com/mental-health-support-on-reddit-64e5f3d38923.

34 Dao, Bo, et al. “Latent Sentiment Topic Modelling and Nonparametric Discovery of Online Mental Health-Related Communities - International Journal of Data Science and Analytics.” SpringerLink, Springer International Publishing, 30 Sept. 2017, https://link.springer.com/article/10.1007/s41060-017-0073-y.

35 Weng, Jiahao. “Topic Modeling in One Line with top2vec.” Medium, Towards Data Science, 21 Dec. 2020, https://towardsdatascience.com/topic-modeling-in-one-line-with-top2vec-a413991aa0ef.

36 Cohen, Idit. “Explainable AI (XAI) with Shap - Regression Problem.” Medium, Towards Data Science, 23 Oct. 2021, https://towardsdatascience.com/explainable-ai-xai-with-shap-regression-problem-b2d63fdca670.

37 Cole, Arthur. “The Quest for Explainable AI.” VentureBeat, VentureBeat, 6 May 2022, https://venturebeat.com/ai/the-quest-for-explainable-ai/.

38 Robnett, Anthony. “Cosine Similarity, TOPIC CLUSTER QUALITY VALIDATION, Google Colaboratory.” Google Colab, Google, https://colab.research.google.com/drive/1nnBE6kbCA8VJry4nEq4z0V-seIhd5iSA#scrollTo=TC5BOgHg3fXs&line=2&uniqifier=1.

39 Robnett, Anthony. “Euclidian, TOPIC CLUSTER QUALITY VALIDATION, Google Colaboratory.” Google Colab, Google, https://colab.research.google.com/drive/1rpAlBUFle67apgbmOgg8OniCsPC_ZNi4#scrollTo=TC5BOgHg3fXs&line=1&uniqifier=1.

40 Robnett, Anthony. “1_Final_Large_Reddit_Scrape.Ipynb.” Google Drive, Google, https://drive.google.com/file/d/1qFSnJ-9Hxcjw8_W-EtYsKow2TjNzO4Bh/view?usp=share_link.

41 Robnett, Anthony. “2_Data_Preprocessing_and_Cleaning_Air_Force_Subreddit_Comments.Ipynb.” Google Drive, Google, https://drive.google.com/file/d/19ECZk2OVF2yiXQ9FkcF2zicm1TzLuY3j/view?usp=share_link.

42 Robnett, Anthony. “3_Top2Vec_Model_Subreddit_Comments.Ipynb.” Google Drive, Google, https://drive.google.com/file/d/1wHeeAA9vBKthtz1O4WSjsEWNNxnn57aD/view?usp=share_link.

43 Robnett, Anthony. “3_5_Modified_Top2Vec_Model_Subreddit_Discussions_Export.Ipynb.” Google Drive, Google, https://drive.google.com/file/d/1gx54T2w7v1Elx1IFYWNQFC1jTDvKkuMT/view?usp=share_link.

44 Robnett, Anthony. “4_Time_Series,_correlation,_and_regression_analyses.Ipynb.” Google Drive, Google, https://drive.google.com/file/d/1HfcqRaXGP4N8QXV1qqD6V0jIyoPnpSiL/view?usp=share_link.

45 Robnett, Anthony. “5_Regression_and_SHAP.Ipynb.” Google Drive, Google, https://drive.google.com/file/d/1sq46To0TsD1jVURJ6AAz5ED4Zyv3XSC0/view?usp=share_link.

46 Robnett, Anthony. “DAF_comments.CSV.” Google Drive, Google, https://drive.google.com/file/d/1RxqDgZJxVbgSLmT0fW_d2kDK_wVukdKg/view?usp=share_link.

47 Robnett, Anthony. “AF_COMMENTS_100K.CSV.” Google Drive, Google, https://drive.google.com/file/d/11ROOAKTTEj9djDy1c-ct7I3W2-Jmmby5/view?usp=share_link.

48 Robnett, Anthony. “PCS_DEPRESSION_COMBINED_MONTHLY_TEXTS.CSV.” Google Drive, Google, https://drive.google.com/file/d/1wk9qdCzA0gB8wfB8lGxZc7kV1AK4rNNb/view?usp=share_link.

49 Robnett, Anthony. “Top2vecmodel.h5.” Google Drive, Google, https://drive.google.com/file/d/168C-ZmubipgU9FgL_4hs81PfCdVfk51A/view?usp=share_link.

50 Polanitzer, Roi. “Data Science One on One - Part 19: Interpreting Regression Output in Excel.” Medium, Medium, 10 Dec. 2021, https://medium.com/@polanitzer/data-science-one-on-one-part-19-interpreting-regression-output-in-excel-fc7d49be63f.

51 “Sklearn.feature_extraction.Text.CountVectorizer.” Scikit, https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

52 “6.2. Feature Extraction.” Scikit, https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction.

53 “Homoscedasticity and Heteroscedasticity.” Wikipedia, Wikimedia Foundation, 17 Dec. 2022, https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity.

54 “Coefficient of Multiple Correlation.” Wikipedia, Wikimedia Foundation, 3 Oct. 2022, https://en.wikipedia.org/wiki/Coefficient_of_multiple_correlation#:~:text=In%20statistics%2C%20the%20coefficient%20of,linearly%20from%20the%20predictive%20variables.

55 “Autocorrelation.” Wikipedia, Wikimedia Foundation, 10 Nov. 2022, https://en.wikipedia.org/wiki/Autocorrelation#:~:text=Autocorrelation%2C%20sometimes%20known%20as%20serial,the%20time%20lag%20between%20them.

56 “Simple Linear Regression.” Wikipedia, Wikimedia Foundation, 7 Dec. 2022, https://en.wikipedia.org/wiki/Simple_linear_regression.

57 “Sklearn.linear_model.Linearregression.” Scikit, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression.score.