Clustering and Predicting Themes in Reddit Health Posts

Caroline Adams

Executive Summary

Natural language processing (NLP) is being increasingly used for public health surveillance purposes with a focus on online forums and social media sites, as they are rich with information about how individuals experience disease and treatment, as well as information about how many people have or are concerned about specific conditions. To effectively use NLP to glean meaningful insights for public health surveillance purposes, it is critical for researchers to be able to differentiate between posts that literally or figuratively discuss health. To start on this effort, Naseem et al. (2022) used the Reddit API to collect Reddit posts that discussed health and manually classified them into posts that discussed health figuratively and non-figuratively.

This analysis utilized the dataset compiled by Naseem et al. (2022) to see if text clustering and prediction efforts could easily identify between these categories for future efforts trying to distinguish between these types of posts on other platforms or with additional Reddit content. Text clustering was performed using the K-means algorithm, resulting in four clusters focused on general chronic health conditions and symptoms, Alzheimer's disease, allergic reactions, and heart attacks. Each cluster contained a mixture of personal, non-personal, and figurative health mentions, indicating that the clustering algorithm did not match the categorizations completed by Naseem et al. (2022). Prediction efforts utilized the K Nearest Neighbors and Naïve Bayes algorithms to predict whether a post used health terms literally or not. Both algorithms achieved approximately 80 percent accuracy, with Naive Bayes performing slightly better than KNN across all evaluation metrics.

Although the K-means clustering algorithm did not identify clusters that matched the labels added by Naseem et al. (2022), the results did provide useful insight into top health themes discussed on the platform. Furthermore, the high accuracy of the prediction efforts indicates that it may be a useful approach for future efforts aiming to distinguish between literal and non-literal uses of health terms. However, given that a substantial proportion of posts were incorrectly classified, future researchers should involve supplemental qualitative assessment procedures to ensure validity of findings.

Introduction

Social media and online forums have been increasingly used for the discussion of and creation of personal connections around health conditions (Patel et al., 2021). Online platforms have facilitated discussion of diagnoses, symptoms, and treatment among individuals with similar health conditions or with individuals generally in their online networks (Sinha et al., 2018). Researchers have identified these conversations as untapped resources for public health and health research with use cases such as aiding public health surveillance, understanding disease progression and symptoms, and tracking the spread of health misinformation (Patel et al., 2021). To parse through these rich resources, natural language processing has been used as a tool with social media platforms and online health forums to glean additional insight into population health experiences.

Previous studies have used natural language processing (NLP) for various health topics, including COVID-19, mental health, and general public health surveillance (Naseem et al., 2022). For example, Reece et al. (2017) applied NLP to forecast the onset of depression and post-traumatic stress disorder using Twitter data, De Choudhury et al. (2014) leveraged Facebook data to identify mothers at risk of postpartum depression, and De Choudhury et al. (2016) explored indicators of suicidal ideation using Reddit content. Patel et al. (2021) analyzed post content from three online health forums using NLP to detect health discussion trends related to the mental and physical health symptoms of COVID-19.

Limitations of previous research methods in this space mainly center around the use of Twitter data and identifying posts with the appropriate contextual usage of certain words (Naseem et al., 2022). Most studies that have been conducted use tweets as the primary data source, which do not always stay online as users change their privacy settings or decide to delete tweets later. In addition, tweets are short, and users may not be able to fully communicate the meaning of their sickness or symptoms. Long posts from other forums or social media platforms allow users to include context in their health-related posts, which provides more useful data for NLP methods.

In this analysis, I applied NLP to a collection of online posts specific to health conditions to infer what clusters and themes appear across different categorizations of health comments, such as the groupings determined by Naseem et al. (2022) (discussed below), and to see if text prediction methods could accurately distinguish between posts that talk about health metaphorically and those that discuss the health of actual people. These questions are important to investigate as online forums, especially those used less than Twitter for this form of research, are sources of health data and experiences that have been mostly untapped and could yield new insights into population health.

Data

The dataset I am using for this analysis was compiled by Naseem et al. (2022) and contains health-related content from 15 subreddits on Reddit focused on health, daily activities, and fun. Using an API, the authors collected 10,015 unique posts from January 1, 2015, through March 19, 2021. The authors labeled each post as one of three categories: figurative health mention (FHM; discussing health terms and topics figuratively or hyperbolically, not literally), nonpersonal health mention (NPHM; discussion of health condition/symptoms generally), and personal health mention (PHM; discussion of health condition/symptoms in relation to a person). The authors have made this dataset publicly available on GitHub, however, the subreddit and time origination of each post was not included in this public version.

Table 1 displays how many posts in the dataset fell into each of the three categories. The classes were roughly balanced in the dataset, with each including slightly over 3000 posts. However, the average length varied substantially across the three categories, with PHM posts having the longest average length, followed by NPHM posts (Figure 1). FHM posts were on average approximately one third of the length of PHM posts. Term frequency, a component that is calculated as part of standard Term Frequency Inverse Document Frequency (TF-IDF) weighting (used to calculate the relevance of words in a corpus), can often be a function of document length (geeksforgeeks.com, 2022). Therefore, in this case, PHM posts could influence the results of analysis approaches such as clustering more than the other categories, indicating that it may be appropriate to use TF-IDF with normalization.

Table 1. Number of Posts Included in Each Health Content Category

Health Content Category	**Number of Posts**
Figurative Health Mentions	3430
Non-Personal Health Mentions	3360
Personal Health Mentions	3225

To pull out the most frequently discussed health terms across all posts in the dataset, word frequency counts were tabulated, and a word cloud was generated (included below; Figure 2) displaying a distribution of the most frequently included terms. After reviewing the top words, those that were directly health-related were categorized into 3 groups: health conditions, symptoms, and other health-tangential terms.

Of the top occurring terms that identified specific health conditions, “cancer,” “heart attack,” and “stroke” appeared in the most posts with frequencies of 943, 941, and 923 respectively (Table 2). “Alzheimer[’s]” disease, “OCD,” and “PTSD” all occurred in less than 20 posts. The most frequently discussed symptoms were “fever,” “headache,” and “cough,” which appeared in 850, 832, and 760 posts respectively (Table 3).

Figure 1. Average Length of Posts for Each Health Content Category

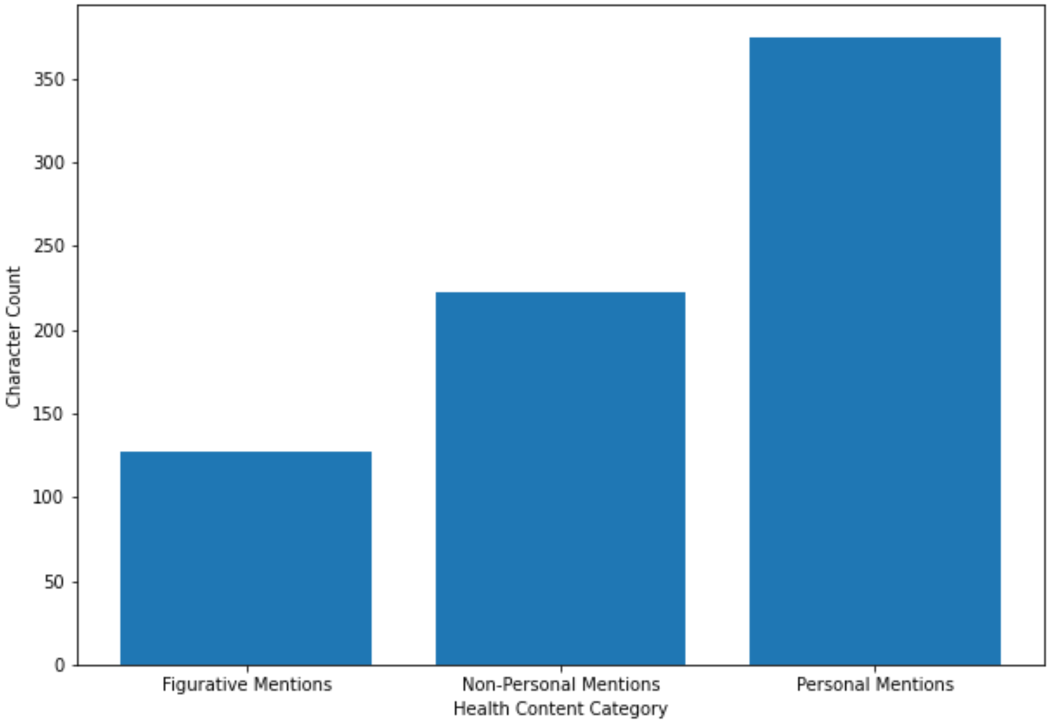


Figure 2. Word Cloud of Most Frequently Occurring Words Across All Posts in Dataset

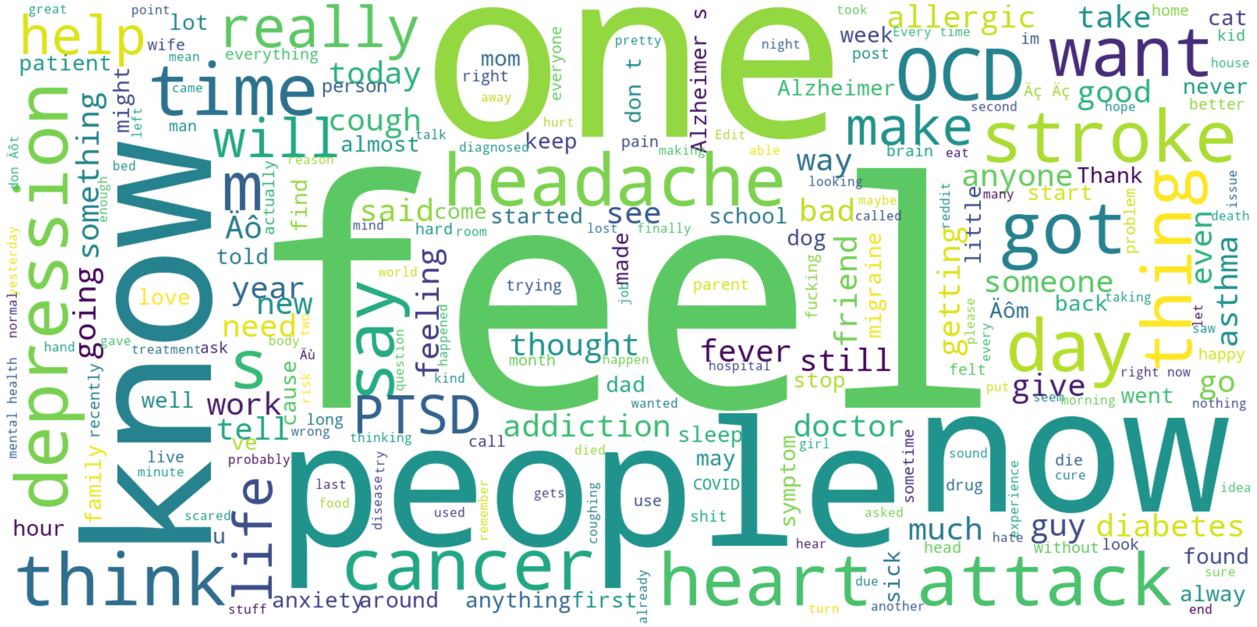


Table 2. Number of Posts That Included Top Health Condition-Related Terms

Health Condition Term	Number of Posts
Cancer	943

Heart attack	941
Stroke	923
Allergic	859
Depression	826
Addiction	690
Diabetes	555
Asthma	428
Anxiety	367
Mental health	128
COVID	85
OCD	19
PTSD	12
Alzheimer	8

Table 3. Number of Posts That Included Top Symptom-Related Terms

Symptom Term	Number of Posts
Feel	1136
Fever	850
Headache	832
Cough	769
Pain	312
Symptom	270
Sick	253
Migraine	252
Coughing	203

Table 4 displays terms that were directly health-related but did not specifically call out a symptom or health condition. Of this other category, “help,” “die,” and “cat” (used in the context of allergies) were the most frequently occurring terms. Examining the most frequently occurring words across the posts was helpful as it 1) validated that the text had been appropriately pre-processed and 2) provided early insight into potential themes that could appear through text analysis approaches.

Table 4. Number of Posts That Included Other Top Health-Tangential Terms

Other Health-Tangential Term	Number of Posts
Help	709
Die	587
Cat (in relation to allergy)	503
Sleep	377
Live	331
Body	330
Risk	293
Brain	268

Diagnosed	240
Drug	223
Died	171
Treatment	159

Lastly, the distribution of the top health-related terms across the three content categories was examined, demonstrating that each category varied significantly in terms of word inclusion (Figures 3-5). As shown in Figures 3-5 below, PHM posts had a more even distribution of the top terms than the other two categories, which were much more concentrated on a narrower set of terms. For example, approximately 600 FHM posts discussed heart attacks compared to around 200 NPHM and PHM posts. In addition, almost 1,000 NPHM posts included the word "feel," whereas this word appeared in less than 300 PHM and FHM posts. This provided insight into which terms may be more useful for clustering and prediction efforts; the dissimilarity is useful for data science techniques that work to distinguish between classes accurately.

Figure 3. Top Health-Related Terms in Personal Health Mention Posts

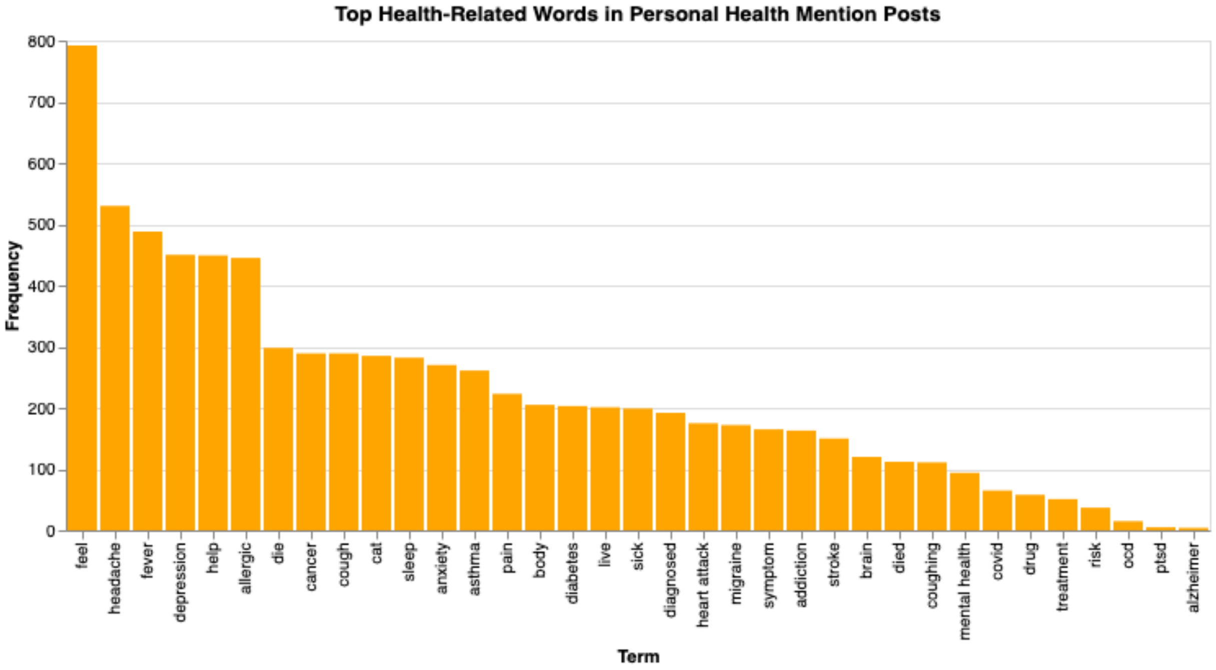


Figure 4. Top Health-Related Terms in Non-Personal Health Mention Posts

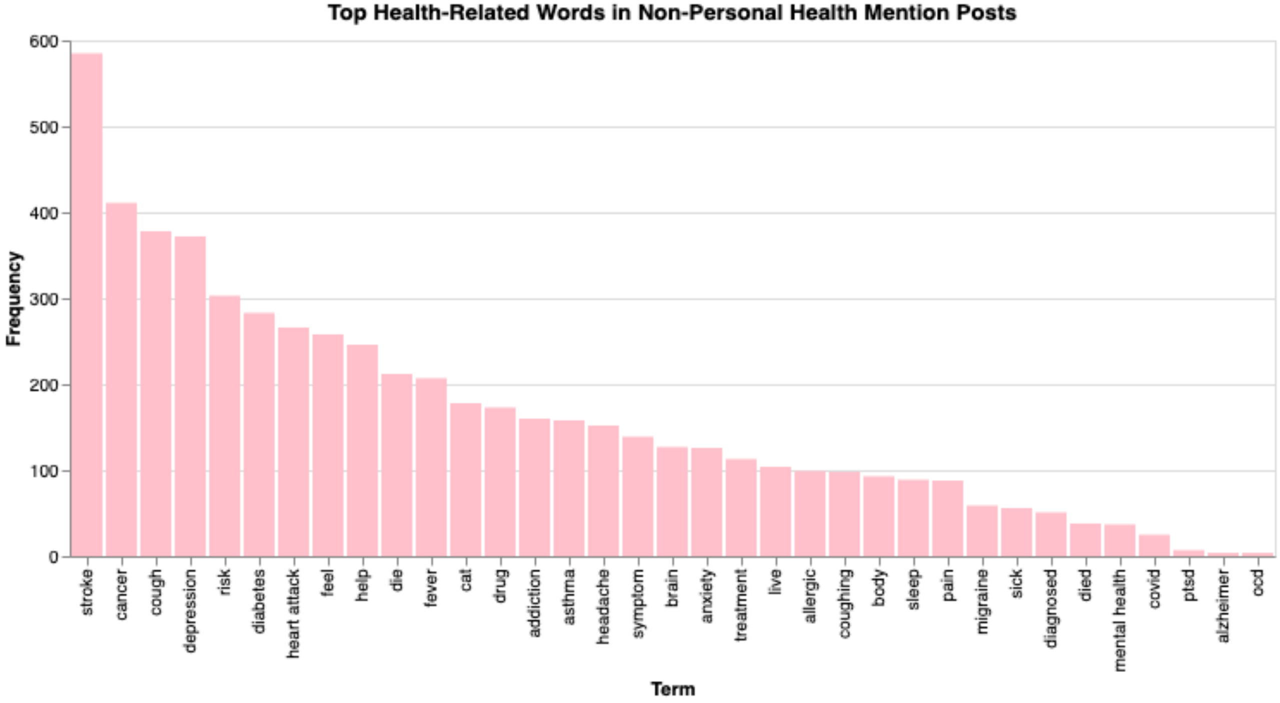
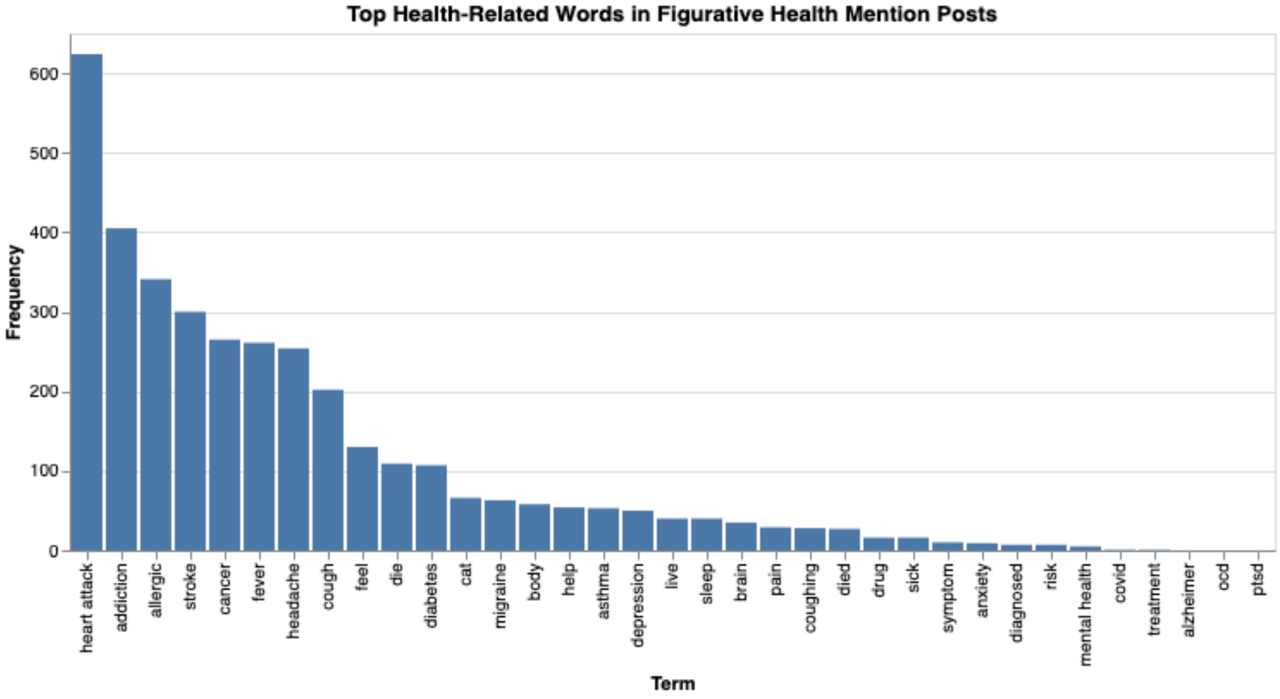


Figure 5. Top Health-Related Terms in Figurative Health Mention Posts



A strength of this dataset was that the authors followed a clear protocol for cleaning and classifying the Reddit text. All posts were evaluated by multiple team members to ensure they were health-related and to ensure the validity of the classification. A limitation of this dataset was that the labeling exercise was subjective. The researchers interpreted the posts to determine which of the three types of information they contained, which could have introduced a level of bias to the classification of the text content, potentially influencing what types of content ultimately fell into the three buckets.

Methodology

This analysis was organized into 3 main phases: data acquisition and pre-processing, application of data science techniques, and validation and interpretation. For this analysis, the two variables included in the dataset generated by Naseem et al. (2022) were used: “text” and “label.” This included the set of Reddit posts and their classification labels. The first phase of the analysis involved loading and pre-processing the data set, performing exploratory analyses (see above) and preparing the documents for text analysis (i.e., constructing the document-term matrix and applying TF-IDF weighting). Two techniques were used for the analysis phase: text clustering and prediction. To complete the analysis, the model inputs and hyperparameters were validated, the accuracy and precision of results were assessed, and the findings were interpreted.

To prepare documents for text analysis, TF-IDF weighting was applied, and a document-term matrix was constructed. Stop words and tokens that involved numbers only were removed from the feature set. See the implementation appendix for more information about how a value for minimum document frequency was set.

For the analysis phase, first, a clustering approach was applied to see if an algorithm was able to generate clusters that were similar to the pre-labeled categories created by the dataset authors or if other clusters may be present across the posts. To do this, the K-means algorithm, an unsupervised learning technique, was used. After the user manually sets a value for K, the algorithm identifies K clusters through a process of identifying center points which minimize the distance between points within a cluster and maximize the distance between points in different clusters (Yildirim, 2020a). Navlani (2022) implemented K-means clustering on a set of news articles, demonstrating the value of this algorithm for instances where users have domain knowledge about how many naturally occurring clusters may be in a dataset. Given that this dataset was pre-labeled with three categories, K-means was selected since a value of K could be selected to match the prior knowledge of potential clusters in the documents. In addition, the algorithm is easy to implement, and validation techniques are well established (Yildirim, 2020a). A primary limitation of this algorithm is that users do have to select K upfront, which can be difficult when users do not have prior domain knowledge to indicate or a rationale for what K should be.

Second, for public health surveillance purposes, it is important to distinguish between posts that figuratively use health terms (e.g., “my apple has cancer”) and posts that discuss the incidence and experience of health conditions for real people. Therefore, for this analysis, the non-personal and personal health mention posts were combined into one category to see if prediction methods could accurately be used to differentiate between posts that discussed human health and figurative uses of health terms.

The prediction algorithms that were used were K Nearest Neighbors (KNN) and Naïve Bayes (NB), which are both unsupervised learning techniques. For text analysis, KNN determines the similarity between documents and uses that information to classify documents into categories. For each document, the algorithm identifies a certain number, K (a value set by users), of the most similar documents. The most frequent classification label of the K similar documents is selected as the document's class (Brodnax, 2022; Yildirim, 2020b). KNN was chosen for this analysis because it is easy to implement and interpret, it works well with large datasets, and it has been used successfully many times with text prediction efforts (Machine Learning Interviews, 2019; Yildirim, 2020b). Potential limitations include that KNN can be sensitive to outliers and that users have to select K upfront. Trstenjak et al. (2013) successfully used KNN to categorize documents into the following topic areas: sports, politics, finance, and daily news. The algorithm was better at predicting some categories than others but had a high accuracy rate overall. The study provided evidence that KNN and TF-IDF together can achieve high accuracy for the classification of text (Trstenjak et al., 2013).

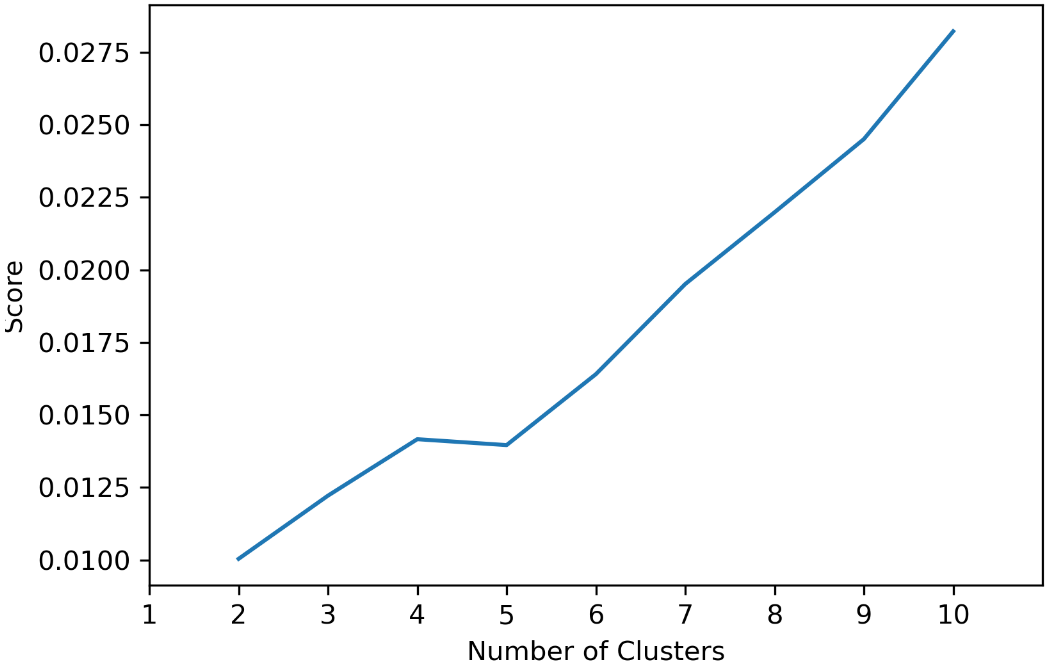
Naïve Bayes is a classifier that predicts the probability that a document is part of a specific class given information about which terms are present. The algorithm uses Bayes theorem to calculate those conditional probabilities (Yildirim, 2020c). Using the information in the feature matrix (i.e., the terms in each document), the algorithm generates probabilities that a document is part of each class and ultimately selects the label with the largest probability. Naïve Bayes was selected for this analysis due to its demonstrated success for text classification efforts. The potential limitations of the algorithm are that it works best on smaller dictionaries and assumes all words are independent. Yeasmin et al. (2022) utilized Naïve Bayes alongside other models to predict user concern about COVID-19 through classification of Twitter posts about COVID-19 by positive, negative, or neutral sentiment. The algorithm was one of the most successful in the study and the authors achieved approximately 90 percent accuracy for the Naïve Bayes models with and without TF-IDF weighting.

Both clustering and prediction efforts were evaluated for quality and accuracy of results. To determine a value of K for the K-means clustering algorithm, domain knowledge and plots of silhouette scores were utilized. The top words occurring in each generated cluster were subjectively reviewed and interpreted; this interpretation was compared to health domain knowledge to validate if the resulting cluster themes and top words made sense within this context. In addition, cluster labels were compared to the labels added to each post by Naseem et al. (2022) to see if the K-means algorithm was able to generate clusters that matched the categorization exercise performed by Naseem et al. (2022). Both prediction algorithms were subjected to five-fold cross validation, and the following metrics were assessed: accuracy, precision, and recall scores. The hyperparameter K for the KNN algorithm was tuned using a validation curve.

Findings

For text clustering, using the K-means algorithm, a value of 3 was initially selected for K to match the pre-labeled categories created by the dataset authors (personal health mentions, non-personal health mentions, and figurative health mentions). Silhouette scores were also plotted against values of K to identify if there were a more ideal value to use for K that would represent a naturally occurring number of clusters in the data set (Figure 6). The plot was used to help determine visually the number of natural clusters in the data by looking for a place in which there is an "knee," which would be an obvious peak or dip in the plotted line. In the below knee plot, we can see a distinct "knee" or peak at 4 clusters. Therefore, the K-means algorithm was used again with K set to 4.

Figure 6. Silhouette Score by Number of Clusters using K-Means Algorithm



To evaluate the quality and content of the clusters, number of posts and top words per cluster were calculated for both values of K. Clusters 1-3 did not change significantly in terms of the thematic content of the top words whether 3 or 4 clusters were generated. Table 5 displays the number of posts in each cluster for both values of K. In both cases, cluster 1 had the greatest number of posts. A review of the top words in this cluster (Table 6) demonstrated that this cluster appeared to be a general grouping of posts discussing symptoms and various chronic health conditions. Cluster 2 focused on discussions of heart attacks and strokes (Table 6). Cluster 3 included terms related to the brain, memory, Alzheimer's disease, and research on cognitive decline and dementia (Table 6). The difference between the two clustering results was that when 4 clusters were generated, cluster 4 seemed to pull posts from the original cluster 1 (when increasing K from 3 to 4, the number of posts in cluster 1 decreased while the number of posts in clusters 2 and 3 remained essentially the same) that appeared to be related to allergic reactions and allergens.

Table 5. Number of Reddit Posts in Each Cluster Using K-Means Algorithm

Cluster	Number of Posts (K=3)	Number of Posts (K=4)

1	8632	8056
2	696	695
3	574	574
4	--	690

Table 6. Top Words Generated for Each Cluster Using K-Means Algorithm with K=4

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
cancer	heart	alzheimer	allergic
just	attack	disease	reaction
like	gave	patients	peanuts
ocd	mini	new	just
fever	nearly	does	make
stroke	just	don	cats
depression	attacks	brain	people
addiction	got	know	like
headache	saw	remember	nuts
cough	like	change	water
ptsd	risk	test	food
diabetes	having	lightbulb	think
don	time	study	cat
time	stroke	just	eat
people	today	having	time
feel	work	drug	deathly
got	gives	people	peanut
know	dad	forget	im
ve	äôs	memory	allergy
asthma	die	research	got

Subjective interpretation of the top words from each cluster indicated that clusters were formed around clear health themes. The themes seemed to be condition specific, except for with the general first cluster, and did not appear to match the categories generated by the dataset authors. While the clustering provides important insight into the content of the Reddit posts, it does not appear as though the K-means algorithm was able to identify natural clusters based on how health-related terms were used (i.e., in a personal, non-personal, or figurative mention). This was confirmed by examining a sample of posts that fell into each cluster; each cluster contains a mix of PHM, NPHM, and FHM posts. Example posts from each cluster are included below in Table 7. Thus, the algorithm did not pick up on the nuances between posts that discussed personal health, someone else's health, or that used health terms in a non-literal sense.

Table 7. Example Reddit Posts from Each Cluster Using K-Means Algorithm with K=4

Cluster	**Health Mention Type**	**Example Posts**

1	Personal	After years of eating tons of junk food and sugar, I decided to quit cold turkey a week ago. Since then, I've experienced extreme cravings, headaches, stomach aches and a drop in my mood. I seriously wasn't expecting all this to happen so suddenly. I didn't even realize I had an addiction. I thought I would miss it a little, but I wasn't expecting the whirlwind my body has gone through in the last 7 days. Here's hoping it gets better this week onwards.
	Non-Personal	Corona and mental health. Indiana's 211 hotline went from receiving roughly 1,000 calls a day regarding mental health — including suicidal ideation — to 25,000 calls a day. And calls to Indiana's addiction hotlines went from an average of 20 a week to 20 a day.
	Figurative	My mom may have a nesquik addiction
2	Personal	Last night my heart was hurting to the point where I had to lie on my back and not move. I have asthma but it was worse than an asthma attack. I was afraid to go to sleep cuz I didn’t want to die. I fell asleep around midnight when it started to get better. It’s 10am and my heart still hurts a bit From the aftermath.
	Non-Personal	Adults with severe obesity (BMI >35) and a prior heart attack who undergo weight-reduction surgery may lower their risk of a second heart attack, heart failure and have less than half the risk of death compared to those who did not have surgery
	Figurative	Eating chocolate everyday until I get diabetes and a heart attack Day 1: I got diabetes and a heart attack.
3	Personal	Bill Gates Reveals His Father Suffers From Alzheimer's Disease – and He's Committing $100 Million to Stopping It
	Non-Personal	Experimental antibodies for Parkinson's, Alzheimer's may cause harmful inflammation
	Figurative	I’ve been battling my addiction to the ‘Hokey Cokey’ dance for a number of years now.. It’s been a long and hard challenge, but I’ve turned myself around and that’s what it’s all about.
4	Personal	I don’t like my psychiatrist he barley talks to me and gives me meds. I’m looking for a new one but I’m wondering if anyone would know about this?
	Non-Personal	Cannabis reduces OCD symptoms by half in the short-term: People with obsessive-compulsive disorder, or OCD, report that the severity of their symptoms was reduced by about half within four hours of smoking cannabis, according to a recent study
	Figurative	This is how my McNuggets came. Someone in the kitchen has OCD

As noted above, KNN and Naïve Bayes algorithms were used for text prediction purposes. Eighty percent of the data was using for training and 20 percent was used as test data. As noted earlier, personal and non-personal health mention posts were grouped together as one class, and figurative health mention posts were used as the other class.

For KNN, a validation curve was plotted to determine what value of K to use to minimize overfitting and maximize test performance. Figure 7 displays the validation curve below, which includes a kink at 15 nearest neighbors. For values lower than 15, the model was not very accurate at making predictions on the test data, and for values higher than 15 there was not much gain in test accuracy. Therefore, a value of 15 was used for K in this analysis. An instance of KNN was initialized and five-fold cross validation was applied, yielding an average accuracy score of 79 percent. A confusion matrix was also generated from the predictions that were made using the test set (Table 8). Using the confusion matrix values, precision was calculated to be 83 percent and recall was 84 percent, meaning that the model was slightly better at predicting which posts fell into the figurative health mention class than those that fell into the literal health mention class.

Figure 7. Validation Curve for Values of K in KNN Algorithm for Text Prediction

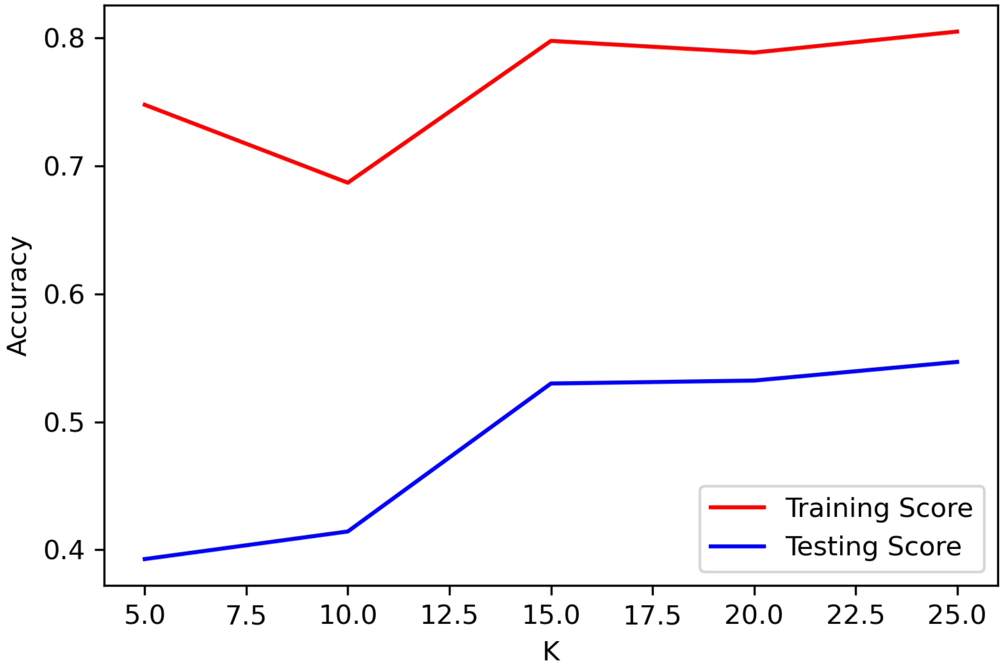


Table 8. Confusion Matrix for Text Prediction Using KNN Algorithm with K=15

	Predicted Figurative Health Mention	**Predicted Literal Health Mention**
Actual Figurative Health Mention	462	225
Actual Literal Health Mention	214	1102

Similarly for Naïve Bayes, an instance of the algorithm was initialized, and five-fold cross validation was once again applied, yielding an average accuracy score of 80 percent. A confusion matrix was also generated from the predictions that were made using the test set (Table 9). Precision for this algorithm was 77 percent and recall was 96 percent, indicating that the model was much better at correctly classifying figurative health mention posts than literal health mentions.

Table 9. Confusion Matrix for Text Prediction Using Naïve Bayes Algorithm

	Predicted Figurative Health Mention	**Predicted Literal Health Mention**
Actual Figurative Health Mention	314	373
Actual Literal Health Mention	53	1263

Relying on the mean accuracy scores from the five-fold cross validation, both models were successful at predicting whether a post was discussing health literally or figuratively most of the time. The Naïve Bayes model was slightly more successful than the KNN model (80 percent accuracy versus 79). Overall, it appears that the use of the KNN and Naïve Bayes algorithms alongside pre-processing steps including TF-IDF weighting, validation of minimum document frequency, and validation of K for the KNN model allowed for meaningful prediction of post content type. However, approximately 20 percent of posts were incorrectly classified, which is not insignificant. Therefore, manual classification of future collections of Reddit posts discussing health topics may still be warranted for researchers or public health practitioners who are interested in pulling insights out of forum content. While it is possible that other algorithms may be more successful than KNN and Naïve Bayes at making this type of prediction, the results of this analysis indicate that human validators are still needed in combination with machine learning techniques to detect nuances in the text that the reviewed algorithms were not able to catch.

Conclusion

The findings from this analysis have important implications for the use of natural language processing to draw out information from online social media platforms related to health that may not be available elsewhere. First, the clustering exercise provided meaningful insights into which health topics are discussed most on the Reddit platform, which would likely be of interest to public health practitioners. Examination of the generated clusters demonstrated that many chronic health conditions such as substance use addiction, cancer, Alzheimer's disease, allergies, and mental health conditions are discussed thoroughly on the site, in addition to acute health events such as heart attacks and strokes, in both literal and hyperbolic manners. The thematic findings could be used by public health professionals looking for specific domains on which users are having in-depth discussions about experience of and treatments for conditions. For example, given the cluster analysis findings, an Alzheimer's researcher may want to investigate if discussions about Alzheimer's disease on Reddit include useful qualitative information about disease progression from the perspective of family caregivers.

Second, the K-means clustering algorithm did not identify clusters that matched the labels added by Naseem et al. (2022) which classified posts as having personal, non-personal, or figurative health mentions. Given the high accuracy of the prediction efforts, prediction may be a more useful approach in comparison to clustering for distinguishing between literal and non-literal uses of health terms; however, given that a substantial proportion of posts were incorrectly classified, future researchers should involve supplemental qualitative assessment procedures to ensure validity of findings.

The analysis was constrained by a few limitations. Notably, the dataset authors did not include metadata in the publicly available dataset about when each post was created or in what Reddit or Subreddit it was posted. Without this information, it is impossible to know if posts were pulled from Subreddits with a specific theme that would influence the content of the posts and the resulting clusters that would be identified in the entire set of data. Furthermore, as noted earlier, the average length of post varied substantially across the three categories (Figure 1). It was noted that TF-IDF weighting can be a function of document length and that TF-IDF with normalization should be applied to ensure that longer posts did not carry more weight in the analysis than shorter posts. Unfortunately, due to time constraints, TF-IDF with normalization was not explored. Lastly, results of this analysis are not generalizable to other social media platforms or forums given the uniqueness of the Reddit platform in terms of post length and the types of conversations that occur within some of the niche subreddits.

Some words and topics were discussed figuratively or hyperbolically more often than others across the set of posts, which makes it difficult to assess the validity of clusters for public health surveillance purposes. For example, an entire cluster with a heart attack theme appearing does not necessarily mean that heart attacks are one of the top four most important health topics to Reddit users. Instead, it likely means that heart attacks are more commonly used in a figurative sense on the platform. Future work should focus in on a single category of posts such as personal health mentions and re-conduct the cluster analysis. This approach may provide more meaningful information about what health conditions and topical themes appear in posts that just discussing a person’s health.

Bibliography

Brodnax, N. (2022). Prediction with Text [PowerPoint slides]. McCourt School of Public Policy, Georgetown University. https://georgetown.instructure.com/courses/158040/pages/module-7-learning-materials?module_item_id=2878410

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared Facebook data. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 625–637. https://doi.org/10.1145/2531602.2531675

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI Conference, 2098. https://doi.org/10.1145/2858036.2858207

Geeksforgeeks.com. (2022). Understanding TF-IDF. https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

Machine Learning Interviews. (2019). How does KNN algorithm work? What are the advantages and disadvantages of KNN? https://machinelearninginterview.com/topics/machine-learning/how-does-knn-algorithm-work-what-are-the-advantages-and-disadvantages-of-knn/

Naseem, U., Khushi, M., Kim, J., & Dunn, A. G. (2022). RHMD: A Real-World Dataset for Health Mention Classification on Reddit. IEEE Transactions on Computational Social Systems, 2339-924X. https://doi.org/10.1109/TCSS.2022.3186883

Navlani, A. (2022). Text clustering: Grouping news articles in Python. https://machinelearninggeek.com/text-clustering-clustering-news-articles/

Patel, R., Smeraldi, F., Abdollahyan, M., Irving, J., & Bessant, C. (2021). Analysis of mental and physical disorders associated with COVID-19 in online health forums: A natural language processing study. BMJ Open, 11(11). https://doi.org/10.1136/BMJOPEN-2021-056601

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. Sci Rep, 7(13006). https://doi.org/10.1038/s41598-017-12961-9

Sinha, A., Porter, T., & Wilson, A. (2018). The Use of Online Health Forums by Patients With Chronic Cough: Qualitative Study. J Med Internet Res;20(1),E19. https://doi.org/10.2196/JMIR.7975

Trstenjak, B., Mikac, S., & Donko, D. (2013). KNN with TF-IDF Based Framework for Text Categorization. Procedia Engineering, 69, 1356–1364.

Yeasmin, N., Mahbub, N. I., Baowaly, M. K., Singh, B. C., Alom, Z., Aung, Z., & Azim, M. A. (2022). Analysis and prediction of user sentiment on COVID-19 pandemic using tweets. Big Data and Cognitive Computing, 6(65), 1–15.

Yildirim, S. (2020a). K-Means Clustering — Explained. https://towardsdatascience.com/k-means-clustering-explained-4528df86a120

Yildirim, S. (2020b). K-Nearest Neighbors (kNN) — Explained. https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3

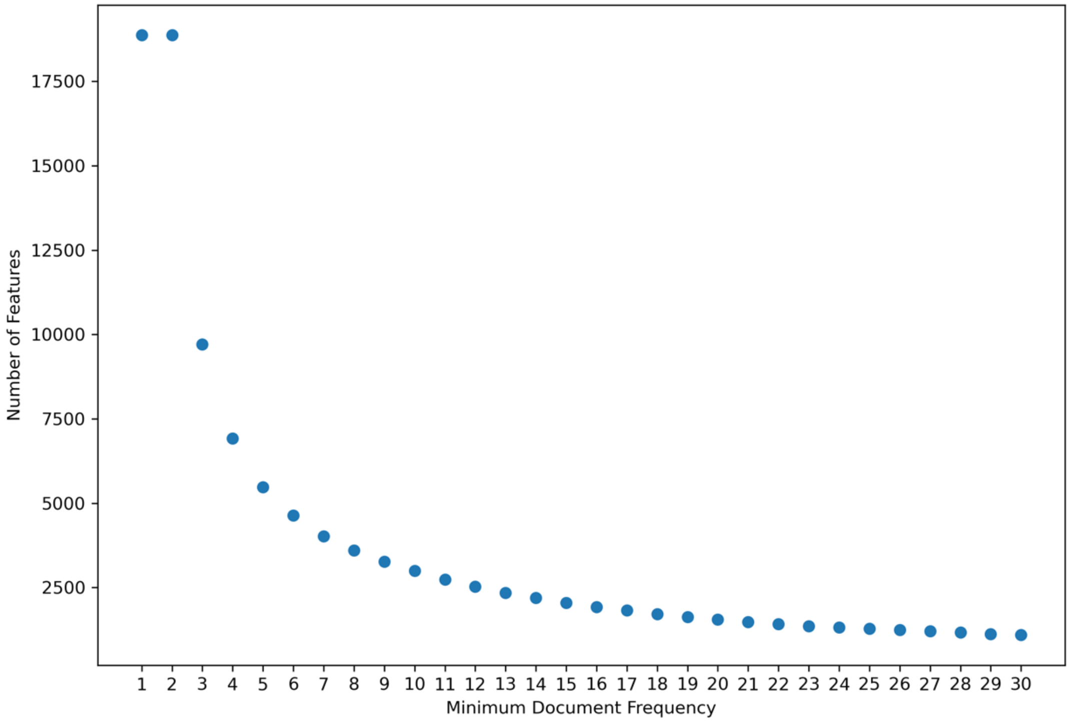
Yildirim, S. (2020c). Naive Bayes Classifier — Explained. https://towardsdatascience.com/naive-

bayes-classifier-explained-50f9723571ed

Implementation Appendix

Minimum document frequency for TF-IDF: For both clustering and text prediction purposes, the minimum document frequency that was set while implementing TF-IDF was validated by plotting the number of features generated for various values of minimum document frequency ranging from 1 to 30, which can be seen in Figure 8, below. The inflection point of the curve on the plot appeared to be around a minimum document frequency of 8. Before a minimum document frequency of 8, the increase in minimum document frequency was associated with significant drops in the number of features. This pattern changed after a minimum document frequency of 10, as further increases of minimum document frequency did not yield as large of decreases in the number of features. Therefore, a minimum document frequency of 10 seemed to be a reasonable value to set for vectorization and was used in this analysis. After removing stop words, tokens with just numbers, and setting a minimum document frequency of 10, the number of features used in the analysis was 2705 (reduced from over 17,000).

Figure 8. Number of Features by Minimum Document Frequency

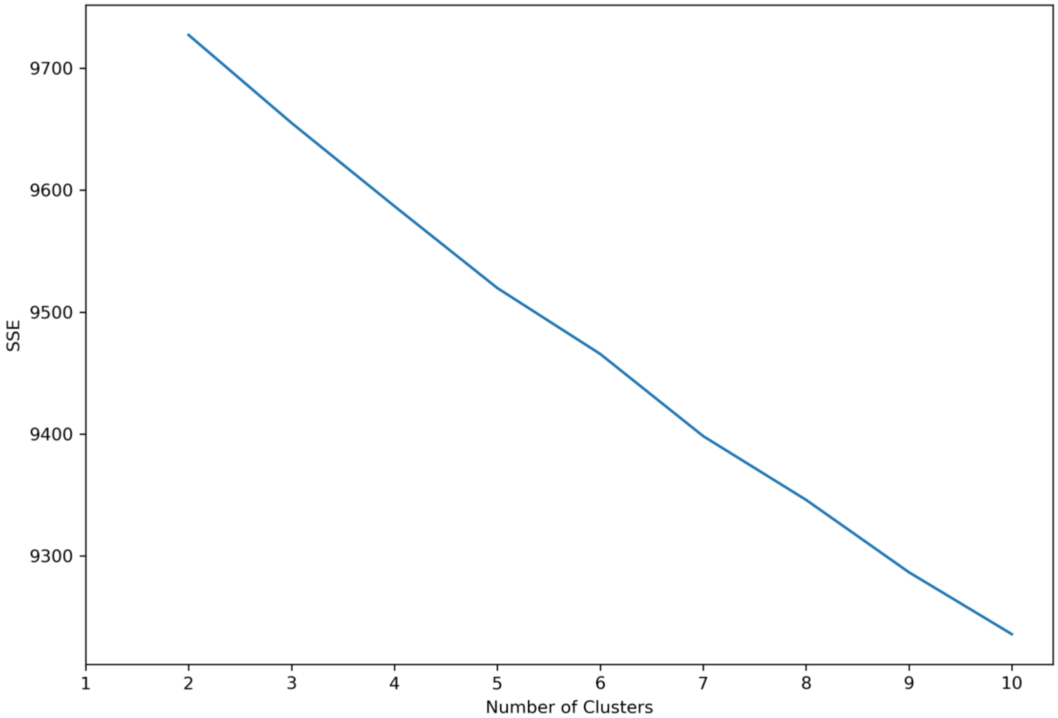


Selection of K for K-Means Clustering Algorithm: Two plots were generated to aid in the selection of K for the K-means clustering algorithm. For each plot, a new algorithm instance was instantiated for K values in the range of 2 through 10. First, a plot of silhouette scores for each instance versus number of clusters was generated (included and described above; Figure 6). Second, a plot of sum of squared errors (SSE) versus number of clusters was also generated

(Figure 9, below). SSE is used to help measure the cohesion in clusters. Minimal cohesion (indicated by minimizing SSE) indicates that our clusters are split into many sub-clusters (higher separation) while higher cohesion would mean there is poor separation, indicating that sub-clusters are merged into one cluster of fewer clusters. The elbow plot below shows SSE versus number of clusters (Figure 9). As expected, SSE is minimized as the number of clusters (K) increases and there is more separation (less cohesion) between the identified clusters. The elbow plot can be interpreted to help determine visually the number of natural clusters in the data by looking for a place in which there is an “elbow,” which would be an obvious bend in the plotted line. In the below elbow plot, there was no obvious elbow. Thus, it was not used to inform which value of K was selected, unlike the plot of silhouette scores (the knee plot).

Silhouette coefficients also represent information about both the cohesion and separation of clusters. Similar to the elbow plot, the knee plot can be interpreted to help determine visually the number of natural clusters in the data by looking for a place in which there is an “knee,” which would be an obvious peak or dip in the plotted line. In the knee plot included earlier (Figure 6), a distinct “knee” or peak was present at 4 clusters.

Figure 9. Sum of Squared Errors (SSE) by Number of Clusters for the K-Means Algorithm


SSE
Number of Clusters