Statistical Data Analysis of the Dataset from df_final.csv

Two unsupervised learning algorithms were used for text classification on the clean_reason column in the dataset of df_final.csv. These algorithms were K-Means clustering and Singular Value Decomposition (SVD). They were used since K-Means clustering can handle big data well due to its linear time complexity, and SVD is popular in the field of natural language processing (NLP) to create a representation of the large yet sparse word frequency matrices. Due to limitations of the RAM memory space of the computing computer, the df_final dataset was reduced from 9 columns with 61,799 rows to 9 columns with 14,512 rows by randomly dropping 83% of the rows that contain the Recall value under the type column as there are more Recall values under the type column than the other three categorical values of Safety alert, Field Safety Notice, and Recall / Safety Alert combined. Below documents how these inferential statistics techniques were used to analyze the data of df_final.csv and the inferences made based on the results.

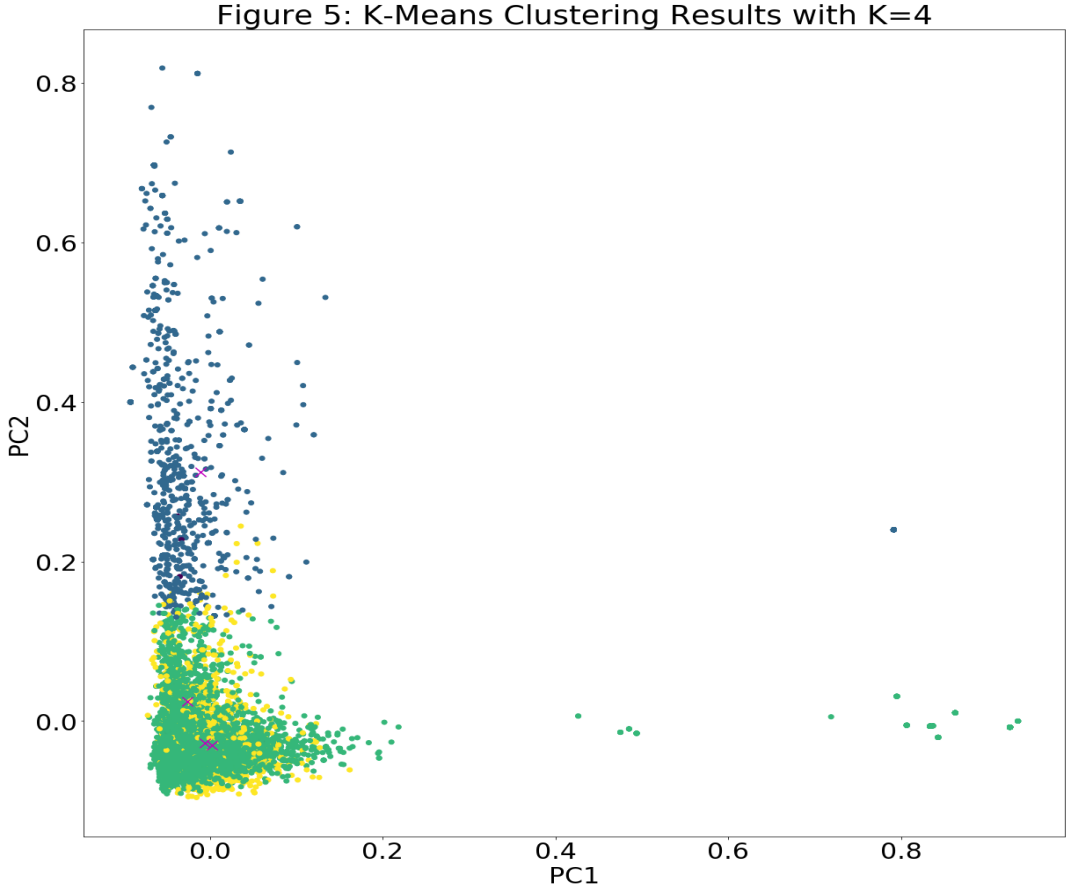
K-Means clustering was first done on the cleaned text under the clean_reason column of the df_final dataframe. In order to do the analysis, TF-IDF (term frequency–inverse document frequency) values were computed to vectorize the text of the clean_reason column in order to create the feature and X matrices. Then, the K-Means clustering algorithm was applied to the vectorized text. The number of clusters in the algorithm was set to 4 since there are 4 event types in the type column in the df_final dataset. The centroids and features of the K-Means clustering analysis were also calculated. The table below shows which clusters the centroids, based on the words of the text in the clean_reason column, belong to.

Centroids in the Four K-Means Clusters

Cluster	Centroids
0	may result use product patient system ha thi potenti dure
1	devic affect manufactur medic safeti product alert supplier pleas ha

2	steril packag compromis product seal may integr pouch due barrier
3	featur befor failur devic use manufactur materi pack compon label

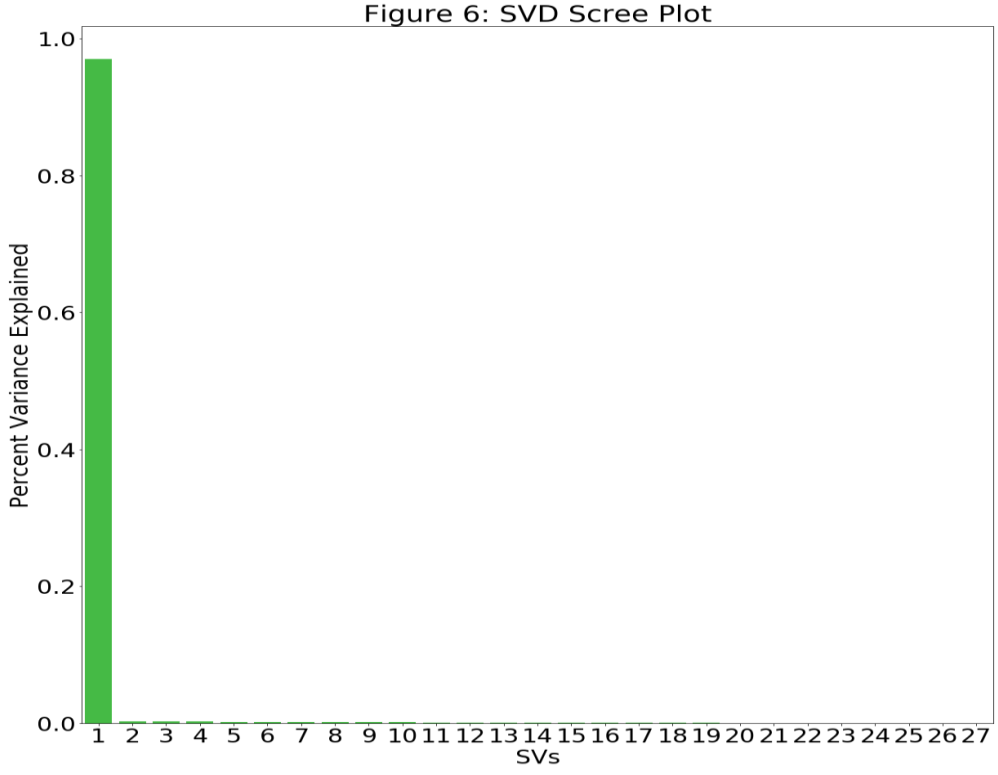
Figure 5: K-Means Clustering Results with K=4



The K-Means clustering plot in Figure 5 was made by creating an instance of K-Means and then using the PCA function in Python 3 to reduce the calculated features and the cluster centers to 2D. According to Figure 5, the scatterplot shows the points overlap instead of forming into four distinct clusters. A homogeneity score was used to check how good the clustering model of the labelled text data of the clean_reason column is in Figure 5. The homogeneity score ranges between 0 and 1 where 1 stands for perfectly homogeneous labeling. The score was calculated to be about 0.129. As it is close to 0, this indicates that the labeling of the text data under the clean_reason column based on the discrete values of the type column is not very homogenous. Even if the Silhouette Coefficient is generally used for unlabeled data, it is also used to check the clustering model in Figure 5. The best value of the Silhouette Coefficient is 1 while the worst value is -1, and values near 0 indicate overlapping clusters. As the Silhouette Coefficient was calculated to be about 0.006, the clusters very much overlap rather than being distinct from each other. The calculations of K-Means so far indicate that the clean_reason column in the df_final dataset does not have the distinct factors to accurately predict what event type would likely occur for a medical device.

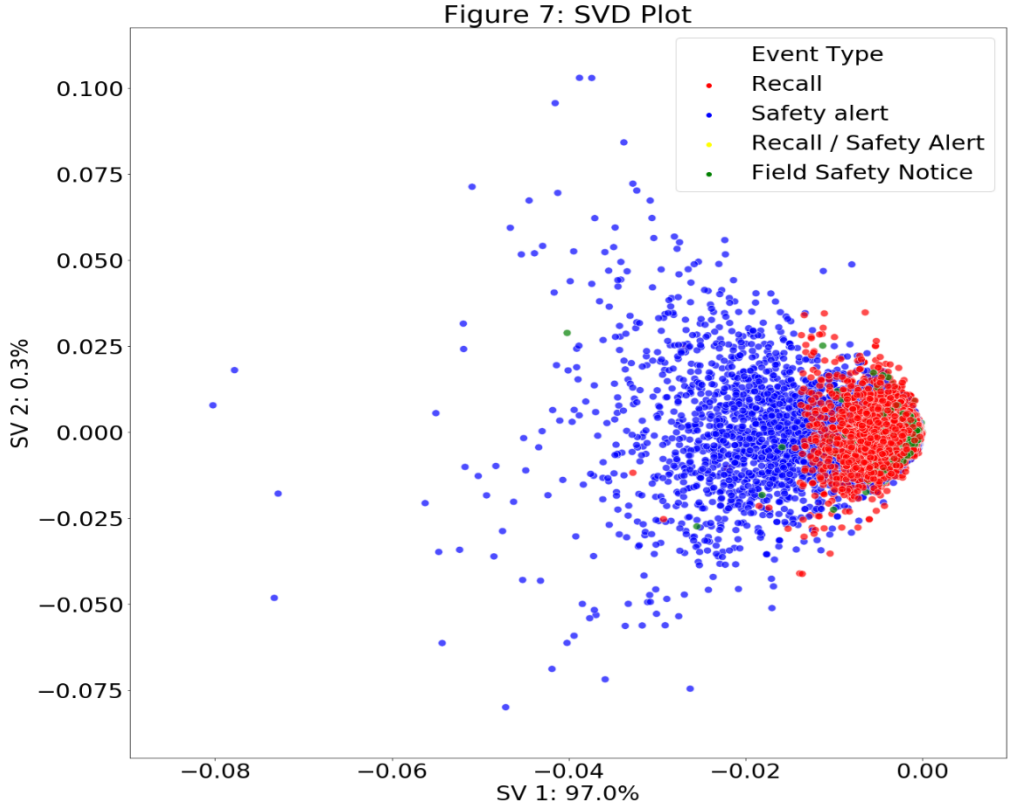
Single Value Decomposition (SVD) was also used to analyze the text under the clean_reason column in the df_final dataset. The bag-of-words approach was used to create a sparse word matrix of the clean_reason column to be analyzed by SVD. NumPy's linalg module's svd function was used to do the SVD with "full_matrices" set to True to get all singular vectors. SVD created three matrices denoted as u, s, and v. The matrices u and v contain singular vectors while s contains singular values. A scree plot was created to visualize the percent variance explained by each singular vector or PC (principal component) and to understand the structure of the text data in the clean_reason column in the df_final dataset.

Figure 6: SVD Scree Plot



The scree plot in Figure 6 shows the percentage of variance explained by each singular vector, and the first vector in Figure 6 explains most of the variation in the text data of the clean_reason column. To better visualize the SVD analysis, a scatter plot of the SVD was created. This was done by creating a data frame containing the first two singular vectors and the meta data from the type column in the df_final dataset before making a scatterplot of this dataframe.

Figure 7: SVD Plot



According to Figure 7, SVD also calculated the data of the text under the clean_reason column of the df_final dataframe to largely overlap and not form any distinctive clusters. So far, K-Means clustering and SVD show that the text under the clean_reason column does not have any distinct classifications to properly predict the event type to likely occur for a medical device. However, there might be a few more statistical methods to properly categorize the text data of the clean_reason column to accurately predict the event type occurring for a device. Perhaps performing k-fold cross validation and a multi-label classification method would improve the classification of the text data in the clean_reason column in the df_final dataset to create a prediction model that accurately predicts the event type that would happen to a medical device.