Milestone Report

Problem Statement:

This capstone project aims to identify and study factors via unsupervised learning that influence certain types of negative events that occur with medical devices. These negative event types are recall, safety alert, field safety notice, and a combination of two or more of them. The primary clients this project targets are those that work in or deal with the medical field, particularly those that manufacture or utilize medical devices as well as the patients that receive them. Medical devices are known to save, extend, or made better millions of lives. However, according to the International Consortium of Investigative Journalists (ICIJ), more than 1.7 million injuries and nearly 83,000 deaths are suspected of being linked to medical devices over 10 years and reported to the U.S. alone. Additionally, the ICIJ noted that the U.S. had more than 26,700 device recalls while India — with more than a billion people — had just 14 from 2013 to 2017. It is important to consider though that this could be due to the U.S. having more rigorous testing and mandatory regulations and the possibility that medical devices were utilized more in the U.S. than in India. Patients often are the last to be informed about malfunctioning devices. Thus, this project can help medical practitioners make better decisions on which medical devices to use with the most minimal of risks as well as patients to be better informed on which medical devices they want to receive. Also, the results of this project can give indications to manufacturers on how to better design medical devices with reduction to injury and mortality.

The International Medical Devices Database will be used for analysis in this project. It is licensed under the Open Database License and its contents under Creative Commons Attribution-ShareAlike license. It is also available for download by the link: https://medicaldevices.icij.org/p/download.

Dataset Description:

It was decided that an unsupervised prediction model using NLP would be created to predict whether a medical device will issue a recall, a safety alert, a field safety notice, or a combination of two or more of them. For the purpose of data exploration and of the main analysis of this project, the df_final.csv file was merged and cleaned from three CSV files (devices-1562662526.csv, events-1562662544.csv, and manufacturers-1562662522.csv). Below describes the data wrangling and cleaning steps and methods used to create the final df_final.csv file.

Using IPython Notebook, devices-1562662526.csv, events-1562662544.csv, and manufacturers-1562662522.csv were read into dataframes called devices, events, and manufacturers respectfully. The devices dataframe had 15 columns and 104,066 rows, the manufacturers dataframe had 10 columns and 26,013 rows, and the events dataframe had 30 columns and 109,574 rows. The devices and manufacturers dataframes were merged (in an outer join fashion on manufacturer_id) to create the df dataframe which had 24 columns and 104,137 rows. Then, the df dataframe was merged (in an outer join fashion on device_id) with the events dataframe to create the df_final dataframe which had 53 columns and 109,645 rows.

Columns in the df_final dataframe were removed based on 60% missing percentage criteria as well as whether or not the columns would contribute to the data exploration and the main purpose of this project. By a boolean array, columns with a missing (null) percentage of

60% or more were removed. Then, it was decided that device_id (ID of the device), device_name (name of device), device_country (country where device was created), event_id (ID of event), action_classification (event risk class), event_country (country where the event took place), reason (textual reasons device is under investigation or reported), and type (event type) would be kept in the df_final dataframe while the others were dropped. This left the df_final dataframe with 8 columns and 109,645 rows.

The columns were then checked and reorganized based on inconsistent and missing values. The values of the action_classification (event risk class) were reorganized to make its string values more consistent. For example, "I," "Class I," and "Class 1" all denoted the category, Class 1, under the action_classification column, and thus "I" and "Class I" were changed into the string value of "Class 1" to indicate they were all of the same categorical value. Additionally, there were outlier values of "Unclassified Correction" and "Voluntary recall." As there were only four of them, the rows containing "Unclassified Correction" and "Voluntary recall" under the action_classification column were removed from the df_final dataframe since their removal won't affect the analysis results of this project. Finally, the rows containing the null values of the type and reason columns were removed from the df_final dataframe. This leaves the df_final dataframe with 8 columns and 61,799 rows.

Text cleaning was done on the reason column to prepare the df_final dataframe for the NLP analysis of the project. First, a function called, clean_text(), was created to keep only alphabetical words, remove whitespaces, and convert the text to lowercase before it was applied on the reason column with the new column called, clean_reason, displaying the cleaned text. Next, a function called, stem_text(), was written to help to stem the words of the text under the clean_reason column. Then, the remove_stopwords() function was created to remove the stop words of the text under the clean_reason column. Finally, the rows containing the null values of the clean_reason column were removed from the df_final dataframe. The head() method was used on the df_final dataframe to check how the merged and cleaned dataset came out by seeing the top five rows of the dataframe.

The final df_final dataframe was exported as a CSV file named, df_final.csv.

Initial Findings:

Table 1: Value Count of Event Type (type) Column in df_final

Event Type	Count
Recall	56955
Safety alert	3821
Field Safety Notice	992
Recall / Safety Alert	14

Table 2: Value Count of Event Risk Class (action_classification) Column in df_final

Event Risk Class	Count
Class 2	41373
Class 3	6896
Class 1	3996

Figure 3: Event Types Recorded by Country Medical Device Originated From

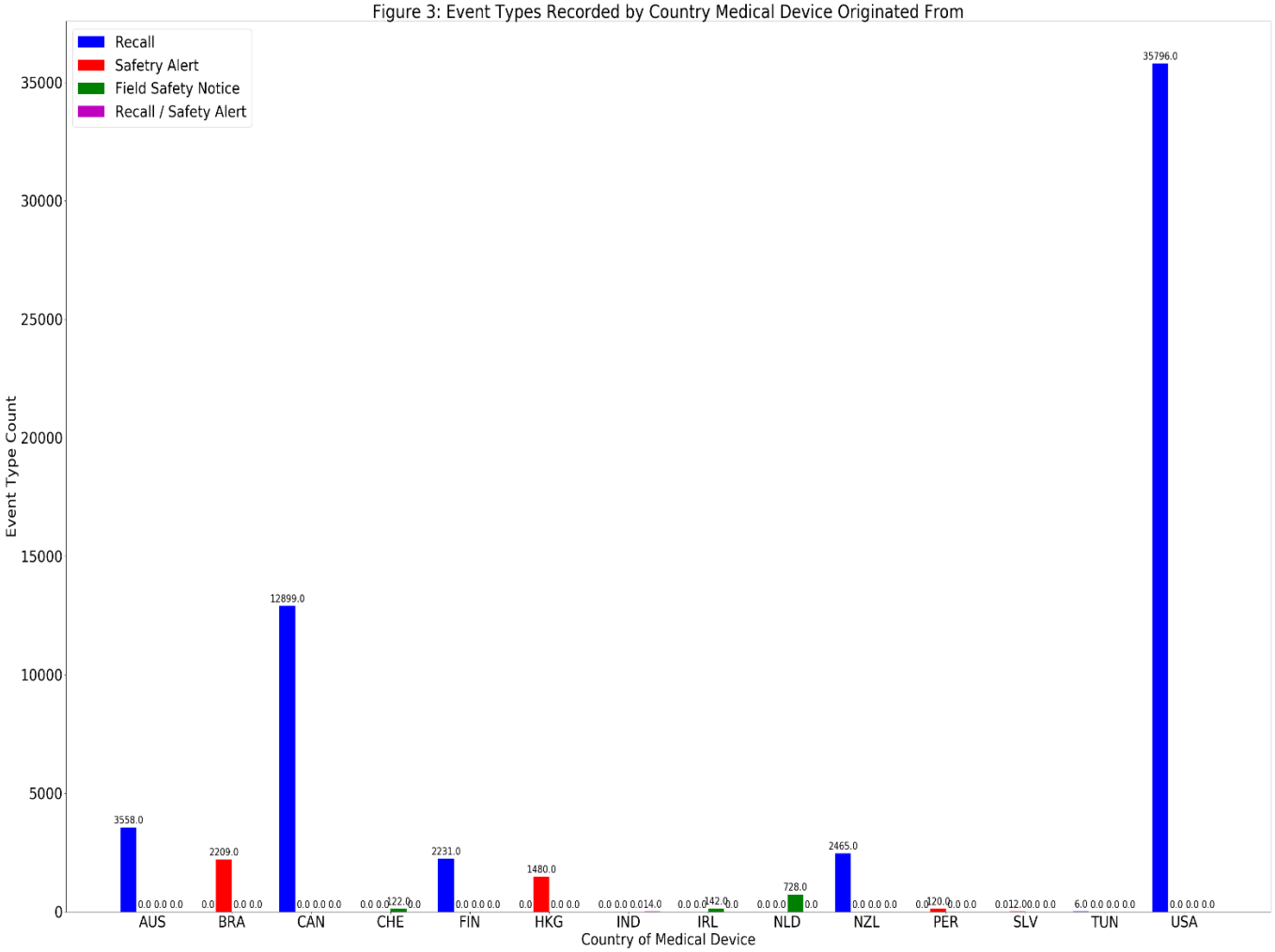
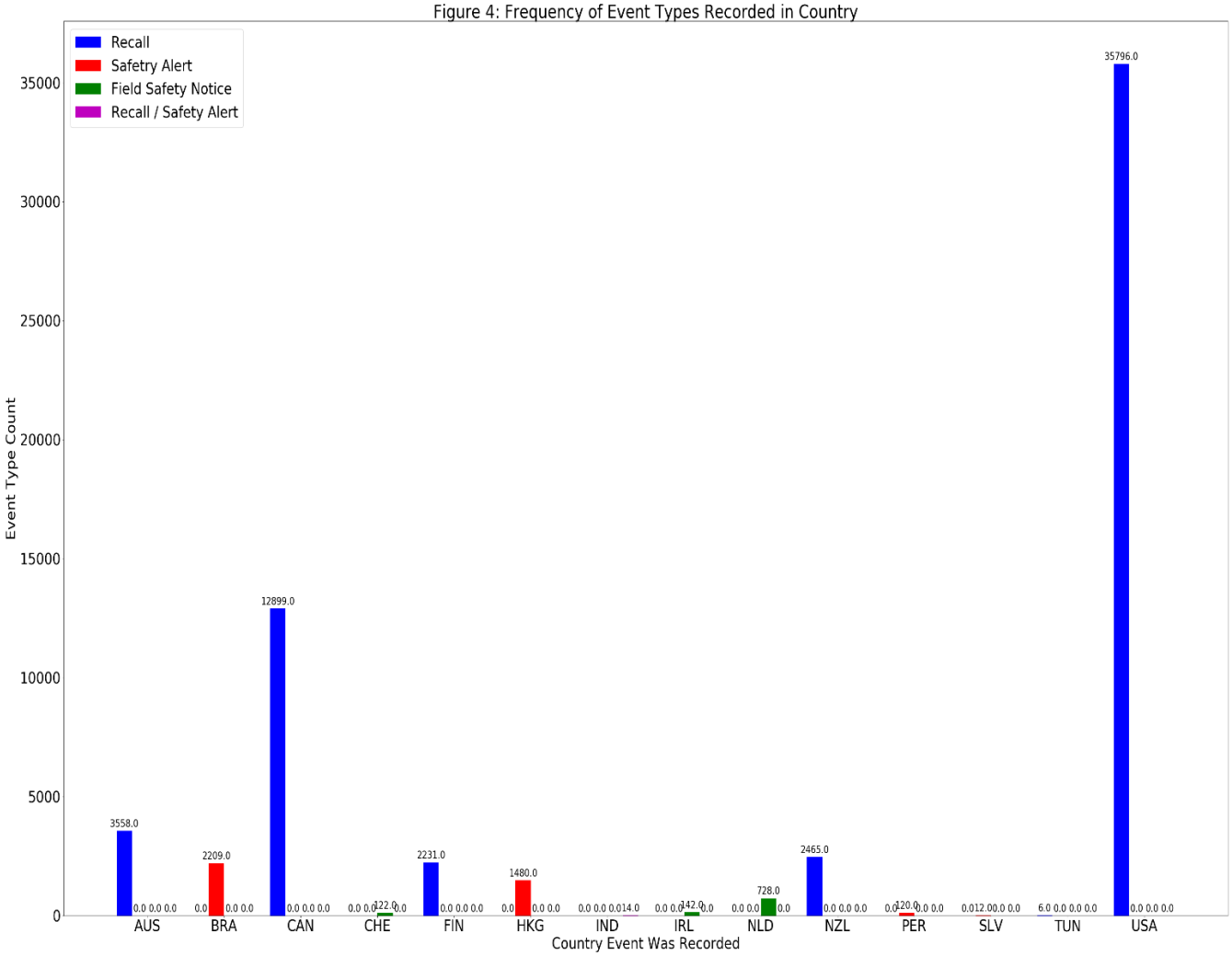


Figure 4: Frequency of Event Types Recorded in Country



Both Figure 3 and Figure 4 show the exact same numbers of each event type being recorded for each country. Perhaps, this is indicative that the devices were used in the same country that they were made from. In other words, they were likely never exported to other countries to use on their patients. Another interesting fact to take note of from Figure 3 and Figure 4 is that the USA has the largest number of recalls for its medical devices than any other country. This might be considered unusual for a developed country. However, it could be for this very reason that more recalls on medical devices were recorded as the USA could have more strict medical policies concerning health hazard and safety. Also supported by Table 1 and Table 2, "Recall" is the event type that is most recorded in the df_final dataset, yet Class 1 which is considered the event risk class with the highest severity in health risk is the least recorded. Thus, there are likely other factors to take into consideration for a medical device to be recalled other than health safety and health hazard. Perhaps the reason column in the df_final dataset could have those other factors to accurately predict what event type would likely occur for a medical device via NLP methods.

Figure 1: 100 Most Frequent Words of "reason" in df_final

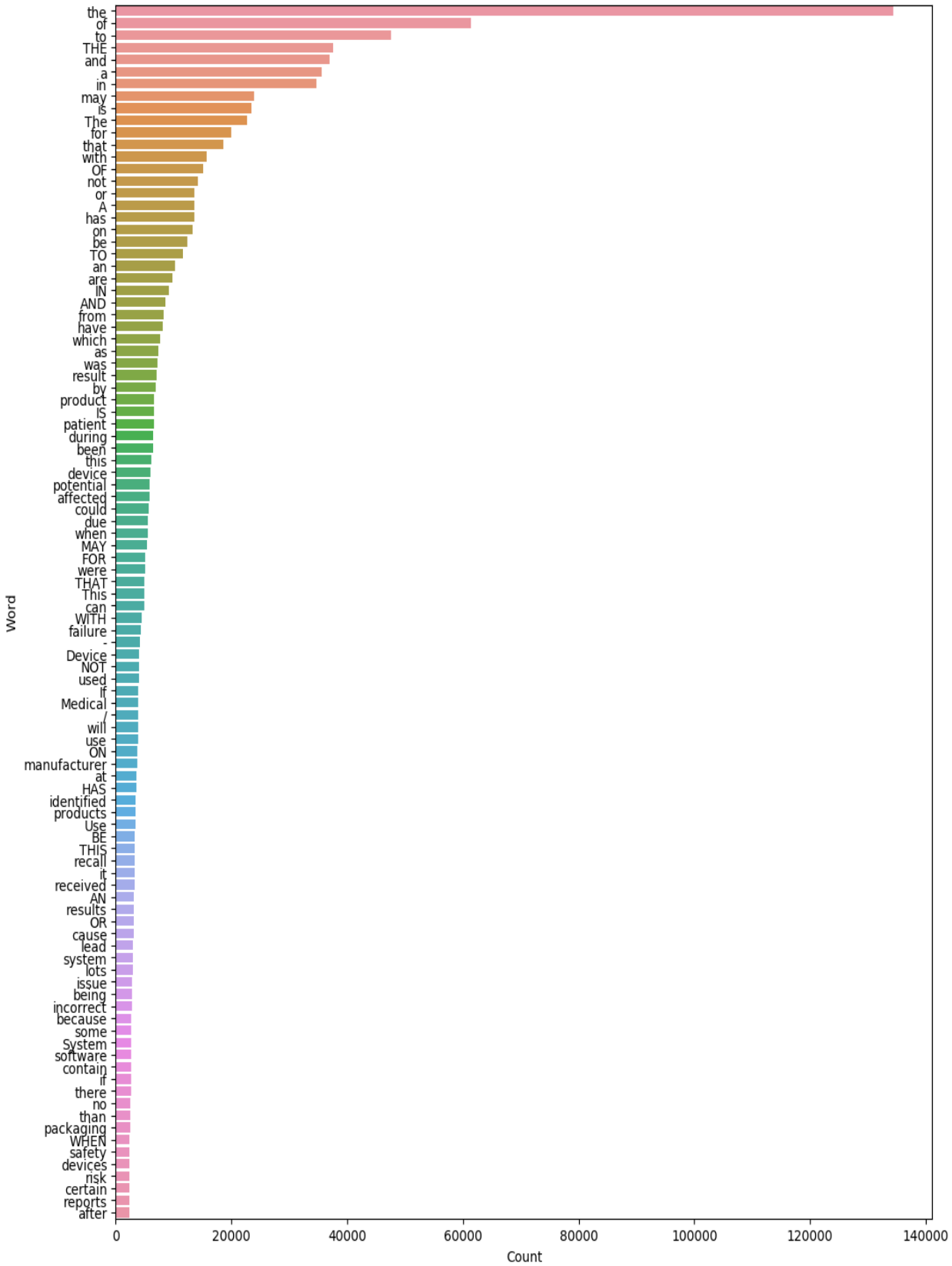
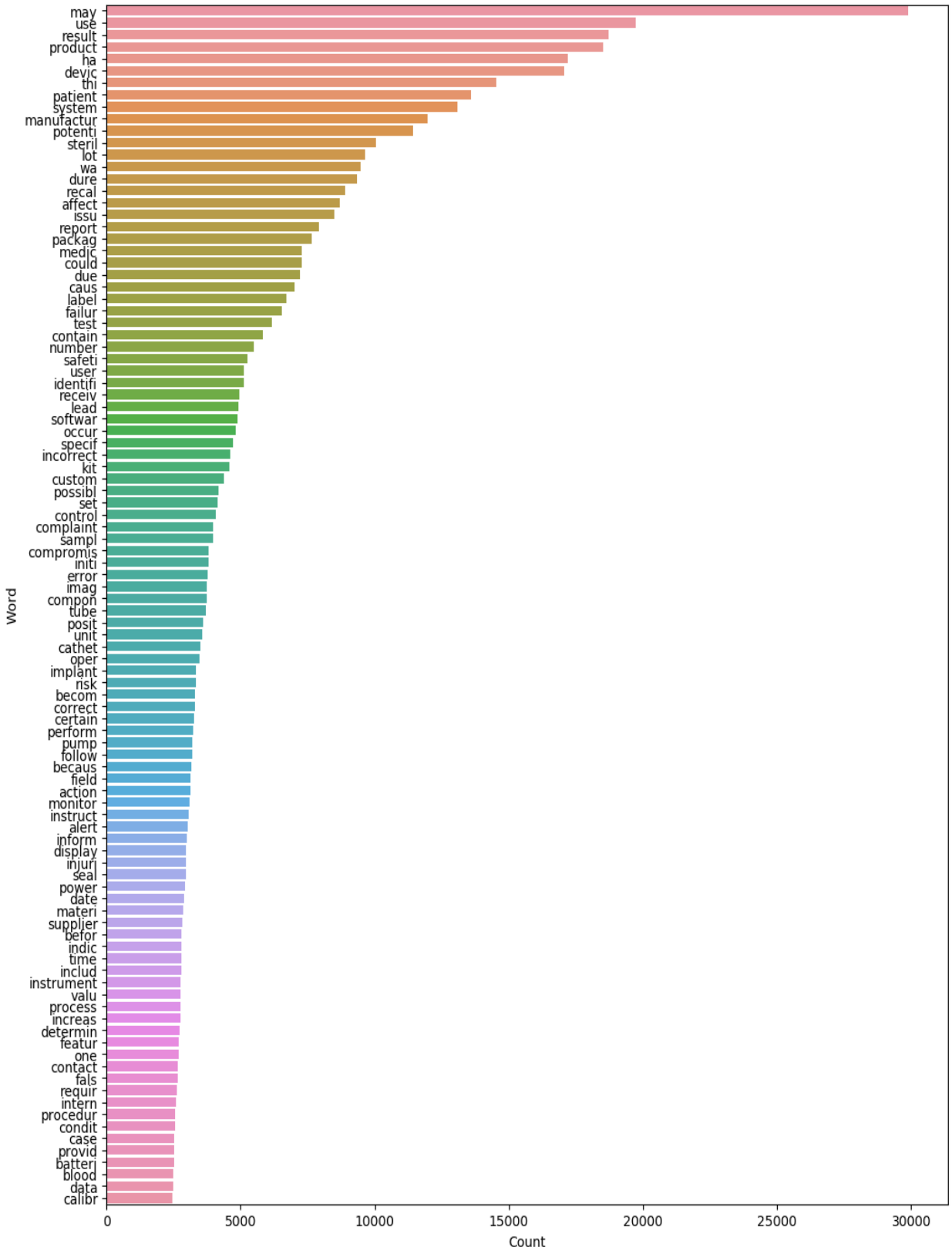


Figure 2: 100 Most Frequent Words of "clean_reason" in df_final



According to Figure 1 and 2, the word frequency plots of both the reason and clean_reason columns of the df_final dataframe show how well the text cleaning methods worked in removing the punctuations and stop words of the text as well as stemming them. Also, Figure 2 shows how much more meaningful words appear in the word frequency count of the text from the reason column after text cleaning was applied to it.

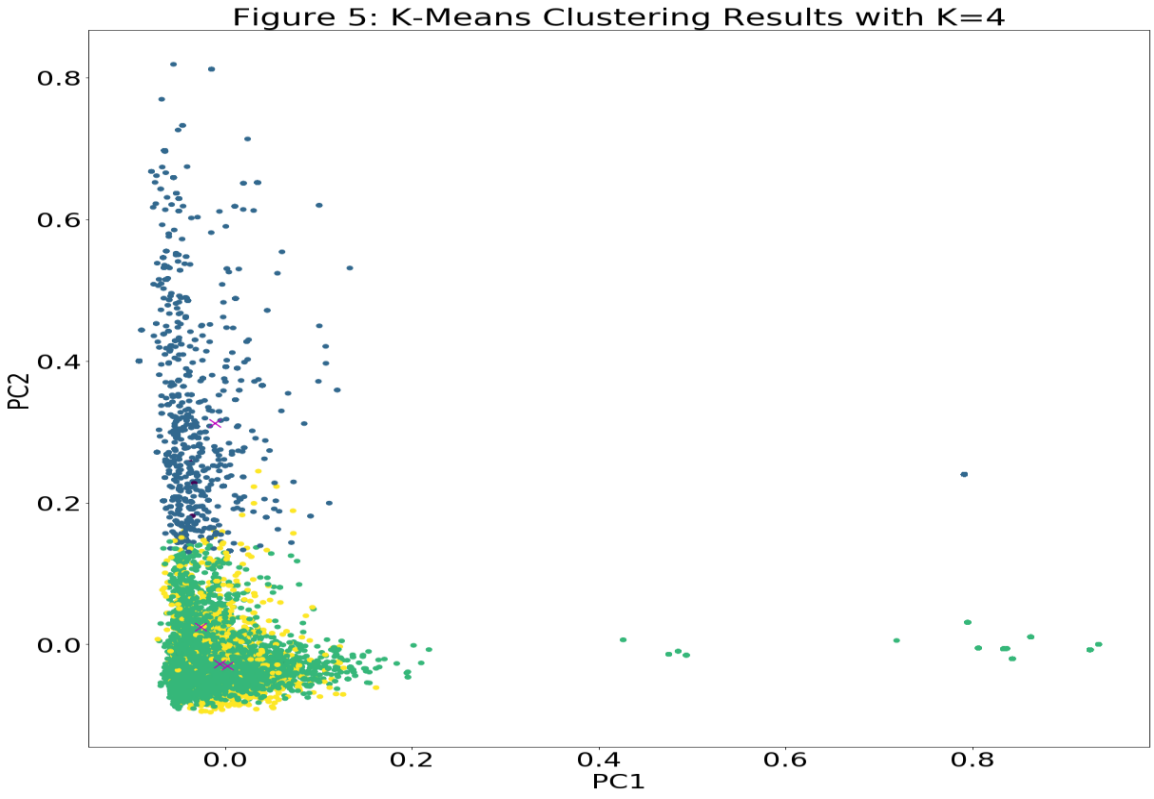
Two unsupervised learning algorithms were used for text classification on the clean_reason column in the dataset of df_final.csv. These algorithms were K-Means clustering and Singular Value Decomposition (SVD). They were used since K-Means clustering can handle big data well due to its linear time complexity, 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. Table 3 below shows which clusters the centroids, based on the words of the text in the clean_reason column, belong to.

Table 3: 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

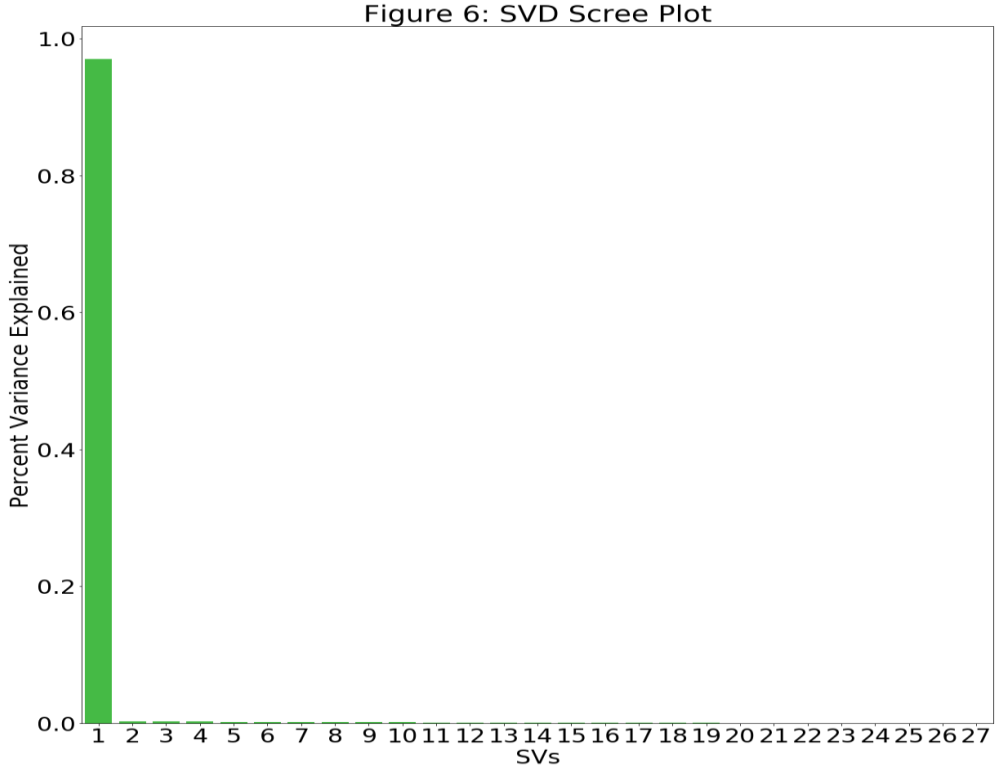
	alert supplier pleas ha
2	steril packag compromis product seal may integr pouch due barrier
3	featur befor failur devic use manufactur materi pack compon label



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.

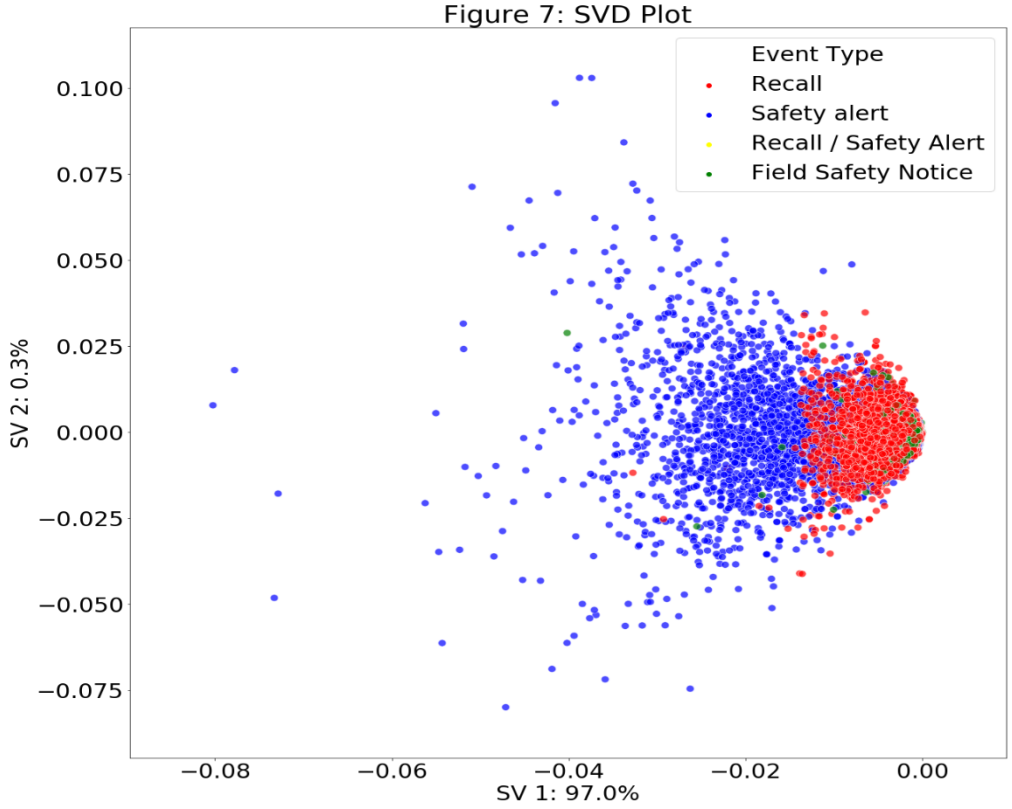
Single Value Decomposition (SVD) was also used to analyze the text under the clean_reason column in the df_final dataset. The bag-of-words approach was used to create a sparse word matrix of the clean_reason column to be analyzed by SVD. NumPy's linalg module's svd function was used to do the SVD with "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.