BIG DATA ANALYTICS FOR VENTILATOR ADVERSE EVENTS IN THE MAUDE DATABASE

By

Gehad Mohammed Ahmed, Neveen Fathy Salama,
Al Zahraa' Mahmoud Ahmed and Shymaa Ali Okasha

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
BACHELOR OF SCIENCE
in
Biomedical Engineering

Under the Supervision of

Prof. Dr. Bassel Tawfik

……..

Professor
Biomedical Engineering and Systems
Faculty of Engineering, Cairo University

Dr. Muhammed Rushdi

……..

Associate Professor
Biomedical Engineering and Systems
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2020/2021

Table of Contents

List of Figures

List of Tables

Abstract

The ongoing global pandemic of the coronavirus disease 2019 (COVID-19) was the reason of the rising need for ventilators. Ventilators help human patients with breathing when such a disease limits the lung capabilities. However, ventilators from different manufacturers can have different features and potential mechanical and electrical problems. In this project, we help healthcare managers and maintenance engineers in making informed purchase, maintenance, and replacement decisions for ventilators. In particular, we carried out a big data analysis for ventilator adverse events in the Manufacturer and User Facility Device Experience (MAUDE) database. We made several illustrated comparisons between different types of ventilators, patient problems, and brands according to the device problems. We extracted files of adverse events of many types of ventilators from the MAUDE database. From these files, we extracted structured information on the Device Problems, Event Type, Patient Problem, Narrative, Event-Description, Brand Name, and Manufacturer. We analyzed this information to understand the variations and types of device problems. We also used manual analysis to understand the unstructured data of the Event Description and Narrative. More importantly, we created a 5000-sample dataset of adverse event reports, and hence trained machine learning classifiers for classifying such reports into 5 categories of device problems. Our results show that a linear SVM classifier achieves an accuracy of 86%.

Chapter 1 Introduction

1.1. Motivation

In December 2019, an outbreak of pneumonia of unknown cause first emerged in Wuhan, and spread rapidly in many regions of this country. Several laboratories identified the causative agent as a novel coronavirus, The COVID-19 has caused shortages of medical devices in many countries, especially ventilators.by spreading at unprecedented rates and causing tens of thousands of fatalities within a few months. Many companies around the world planned to produce ventilators. According to the World Health Organization (WHO), the world needs more ventilators, but scaling up production is more complex than it seems. Around the world, thousands of volunteers are working hard to fill the gaps. There is a board range of estimates of the number of ventilators we will need to care for U.S. patients with Covid-19, from several hundred thousand to as many as a million.

Our study, is an analysis on reports of adverse events related to ventilators, which are submitted to the Food and Drug Administration (FDA) Manufacturer and User Facility Device Experience (MAUDE) database, as the world needs more ventilators under COVID-19. FDA website provides databases for different adverse events all over the years since 1998.

1.2. Background on the Problem

The previous studies are conducted to determine failure patterns of the portable ventilator related to batteries in order to be able to determine the life time average for each battery type. This allowس them to know the optimum battery type for the ventilator, the adverse events caused several event types such as injury, death, and malfunction. Also, adverse events caused by another several device problem categories, such as circuit failure problems, power problems, mechanical problems, calibration problems, and software problems. Our study is conducted to determine failure patterns of ventilators to be analyzed, to identify the root causes behind ventilator failures.

1.3. Objectives

Data analytics to investigate CBK & BTL adverse events records that will help The Egyptian Ministry of Health and medical devices maintenance centers to take right decisions related to which manufacturers are recommended for buying from them ventilators under Covid-19

1.3.1. Overall Objective

The objectives of our study are to

- Investigate ventilator adverse events records of big data obtained from the MAUDE database
- Analyze and identify the possible causes of ventilator adverse events.
- Estimation for the common device's manufacturers who caused adverse events less than others using the MAUDE database.
- create a Multi-Class text classification system for the ICU ventilator adverse events reports in 2020.

1.3.2. Detailed SMART Objectives

In our study, we are going to

1-Introduce number of adverse events reports of ventilators through last five years, number of reports in each device problem category, from that we managed to know the most frequent device problems occur, the trending of adverse events in increase or in decrease, and that by seeing the number of adverse events reports from 2016 to 2020, this work had been done through two weeks, the objective behind that is helping the manufacturers to improve their device technologies for longer use, and giving recommendations on which device to be bought based on having less number of malfunctions, this will help Egyptian Ministry of Health and hospitals in ventilator sell operation, and we could measure all this from 2016 to 2020 during two weeks .

2- Analyze ICU ventilator adverse events reports in 2020 manually to create a Multiclass text classification to predict the device problem category of the input report, we achieve an accuracy of 86% using linear kernel SVM model.

1.3.3. Implementation of the project object

We downloaded data records of (MAUDE) Database from 2016 to 2020, and extracted the data related to (CBK)&(BTL), which are different types of ventilators with different product code, then we select the attributes that we can gain information from such as Event Type ,Device Problems, Narrative ,Event description and Manufacturer columns from this data, these columns includes structured data such as Event Type, Device Problems and Manufacturer columns, and unstructured data such as Narrative and Event description columns. Then we made manual analysis on unstructured data to reduce number of device problem categories to (circuit failure problems, power problems, mechanical problems, calibration problems, software problems) each main category included number of subcategories, that enabled us to convert our problem into supervised problem which can be trained and tested, we applied natural language processing (NLP) for getting the features needed for our multiclass text classification system, we had 5000 row in our dataset of narrative and event description with 5 classes, and hence trained machine learning model for

classifying unseen report before, Our results show that a linear SVM classifier achieves an accuracy of 86%, it was the best model. At the same time, we made analysis on structured data to get relations between manufacturers and Device Problems categories, our results on CBK showed that Philips_Respironics_Inc is the most manufacturer with display and battery problems. From all that we could get valuable information from our data analytics which can help Egyptian Ministry of Health and medical devices maintenance centers to take right decisions related to ventilators under Covid-19.

1.4. Gantt chart

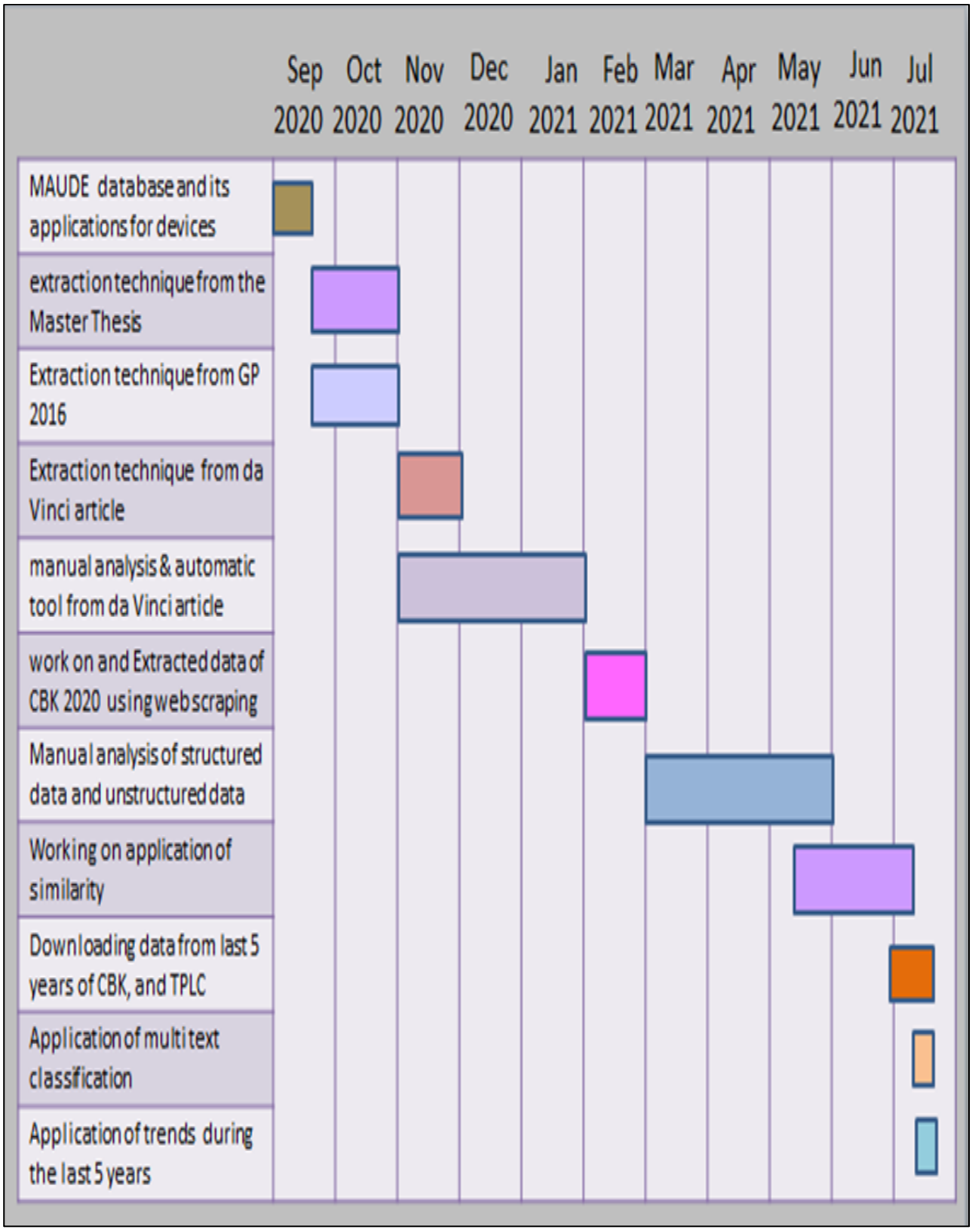


Figure 1.Gantt chart

Chapter 2 Literature Review

By reading many suggested papers that presented data analysis near to our work. Table 2.1 shows a list of the work related to our graduation project [1][2][3]4].

Table 1. work related to our graduation project

Authors	Year	Data source	Methods	Result
Alemzadeh et al.	2016	MAUDE DATABASE	By developing an automated natural language processing tool, they performed analysis of the adverse events reported to MAUDE database from 2000 to 2013	they determined the number of events reported per procedure and per surgical specialty, the most common types of device malfunctions and their impact
Abdullah	2018	MAUDE and TPLC DATABASES	Text classifiers using regularized logistic regression, in Data Extraction developed a framework for automated analysis of failure data	he determined the number and rate of (AED) device incidents, identify the causes (the number of battery-related), and patient impact of the AED device malfunctions
Sagy and Magdi	2015-2016	MAUDE Database	they analyzed adverse events with text mining For classification	They extracted the device data related with the ventilator adverse events (AEs). The MAUDE returned 2,785,054 reports in this

			of failures, then Applied Hierarchical Agglomerative clustering, and applied analysis (PCA) to reduce number of features	interval. Among these reports, there are 25,637 (0.9%) AEs reports associated with the ventilators. They cleaned, analyzed, and classified these AEs textual reports using text mining. Moreover, they fit the failure data to several probability distributions to find the best-fit distribution
Alemzadeh et al.	2013	public FDA databases	they investigate causes of failures in computer and their impact on patients by analyzing descriptions of recalls and adverse event reports from public FDA databases	The authors characterize computer-related failures by deriving fault classes, failure modes, recovery actions, and number of devices affected by the recalls. This analysis is used as a basis for identifying safety issues in life-critical medical devices and providing insights on the future challenges in the design of safety-critical medical devices
Kohani and Pecht	2017	MAUDE Database	They used data mining analysis of an electrostatic discharge (ESD) failures in medical devices, over the last ten years, using the U.S. FDA's MAUDE database	Analysis shows the relationship between low RH and occurrence of ESD malfunctions the number of ESD malfunctions during cold (dry) months were 6 times higher than that of summer months. As a result, hospitals must maintain at least 30% RH level, and the minimum RH limit (20%) provided by Addendum D to ASHRAE 170-2010 must be changed back to the 30% in its previous version

Chapter 3 Materials and Methods

3.1. MAUDE database

Manufacturer and User Device Experience (MAUDE), is a Food and Drug Administration (FDA) database that stores Medical Device Reports (MDRs) submitted to the FDA, these reports, submitted by mandatory reporters (manufacturers, importers, and device user facilities) or voluntary (care professionals, patients, and consumers). Each report contains information such as

- Device name
- Event type (malfunction, injury, death, other)
- Manufacture name such as Philips and ResMed
- Brand name which refers to model name such as Trilogy100, AstralTM10
- Device problem
- Patient problem
- Event Description and Manufacturer Narrative fields, written by the reporter without any rules, Therefore, manual analysis of accident reports requires a great human effort. Figure 2 show an example of MAUDE report. Certain types of reporting information are protected from public disclosure under the Freedom of Information Act (FOIA). If the report contains trade secrets or confidential business information, this text shall be replaced by "(b)(4)". If the report contains information on personnel or medical files, this text shall be replaced by “(b)(6) [6].

MAUDE can help customers find out about device problems that other customers have reported. We encourage customers to review the database before making a purchase.

3.2. Data Extraction

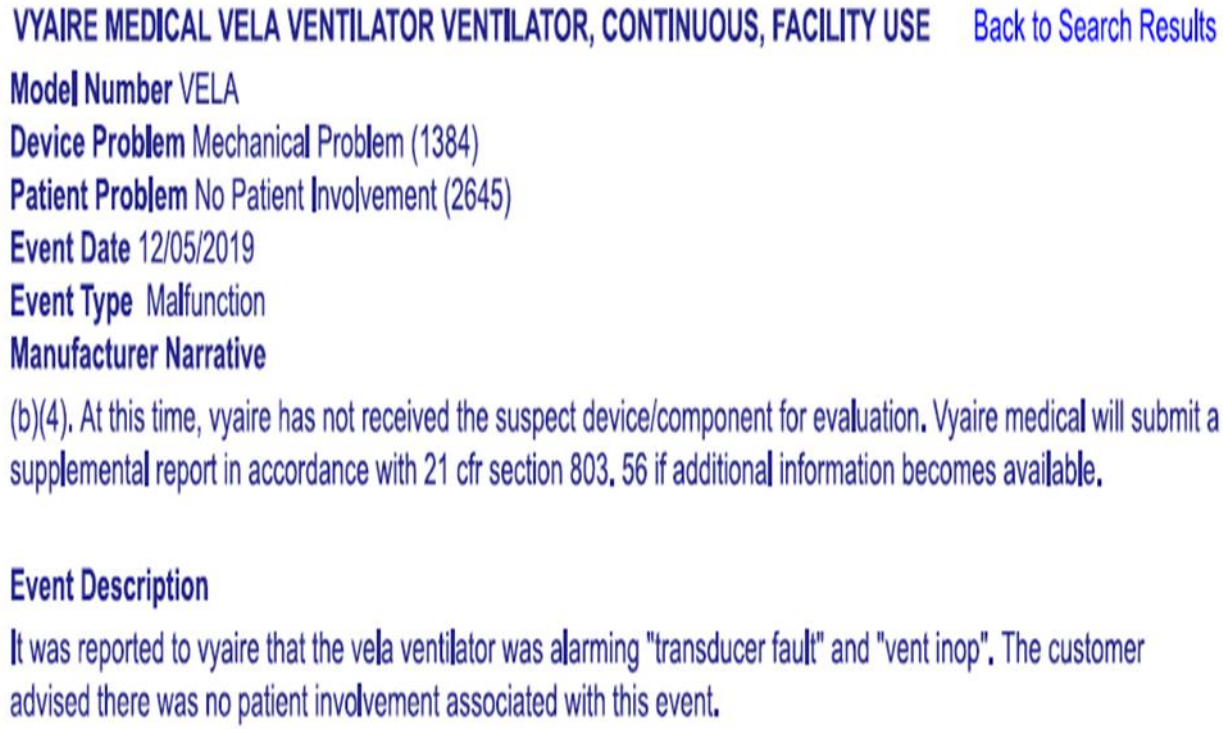
We developed web scraping tool that automatically retrieves all the reports related to two types of ventilators CBK (continuous, facility use ventilator) and BTL (emergency, powered (resuscitator) ventilator) over 5.5 million MAUDE records posted between January 2016 and December 2020 and get 60,000 records. Figure 3 Show data extraction.we extracted the following information from the structured data in the reports:

- the relation between device problems and the name of manufacturer and brand.
- no of each event type each year.

See data extraction steps in Appendix A .

3.3. structured data analysis

using python code to extract each device problem and find the relation between main manufacturers and each problem. Extract each event type each year for CBK and BTL. Using excel to plot the results.



VYAIRE MEDICAL VELA VENTILATOR VENTILATOR, CONTINUOUS, FACILITY USE Back to Search Results

Model Number VELA
Device Problem Mechanical Problem (1384)
Patient Problem No Patient Involvement (2645)
Event Date 12/05/2019
Event Type Malfunction
Manufacturer Narrative

(b)(4). At this time, vyaire has not received the suspect device/component for evaluation. Vyaire medical will submit a supplemental report in accordance with 21 cfr section 803. 56 if additional information becomes available.

Event Description

It was reported to vyaire that the vela ventilator was alarming "transducer fault" and "vent inop". The customer advised there was no patient involvement associated with this event.

Figure 2.an example of MAUDE report5]

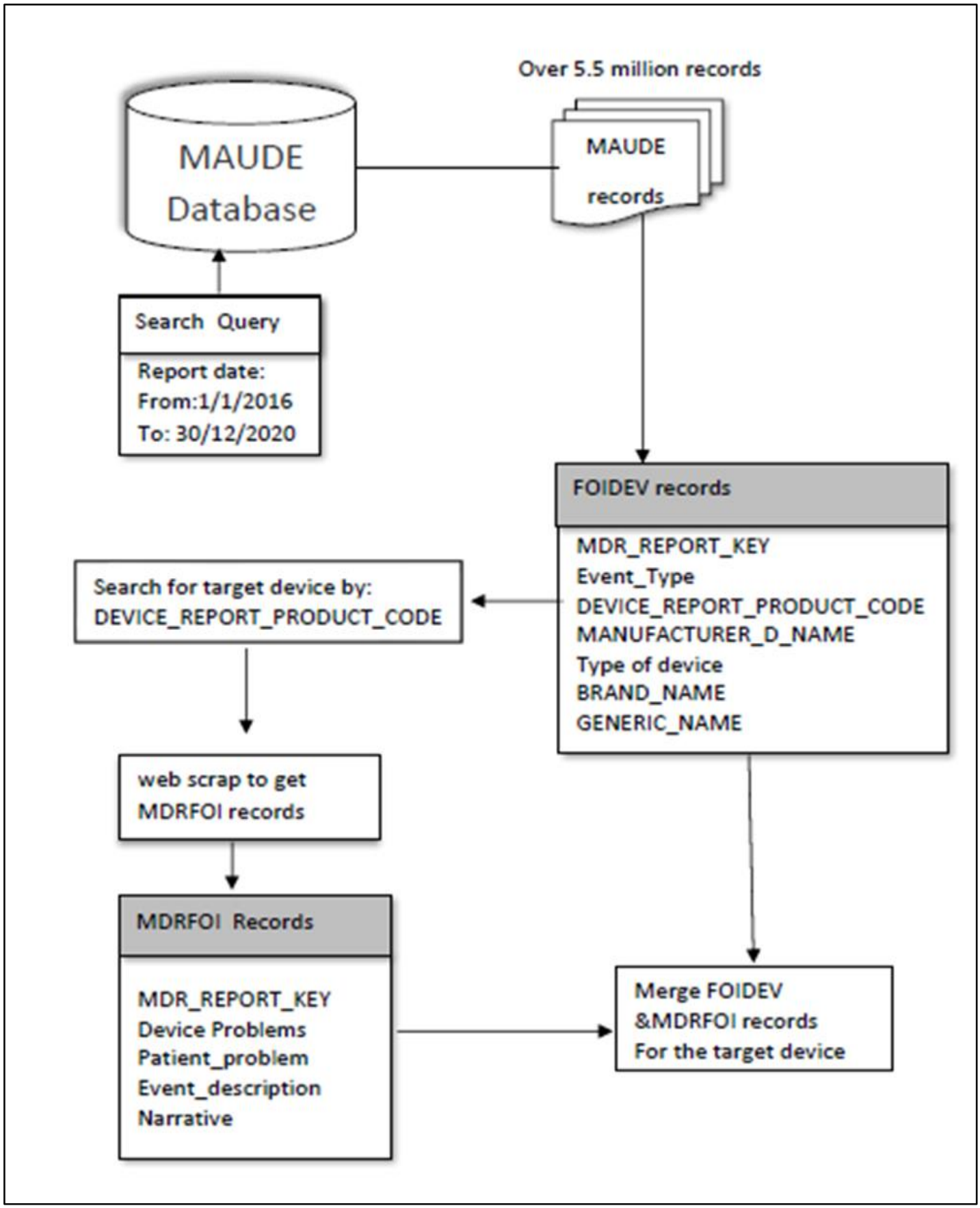


Figure 3.flow of data extraction

Chapter 4 : Results

4.1. For CBK

We extract 60,000 records, about 33% with display problems, 8% with mechanical problems, 6% with battery problems, power, calibration, charge, problems each was 5%, circuit failure and output problems each was 4%, 30% other problems.

The relation between device problems and the name of manufacturer

main global manufacturers of ventilators and their products are:

- ResMed (AstralTM100, AstralTM150)
- Philips Healthcare (Trilogy100, Trilogy200)
- Vyaire Medical (Fabian™ HFO BELLAVISTA™ 1000 VENTILATOR)
- Medtronic (Puritan Bennett™ 980 Ventilator)
- Draeger (Evita® V800)
- Fisher&Paykel Healthcare (F&P 850)

4.1.1. The relation between device problems and the manufacturers

Main manufacturers with display problems Figure 4, with mechanical problems Figure 5, with battery problems Figure 6, ResMed was the most manufacturer with Calibration problems, Philips Respironics was the most manufacturer with charge and Circuit failure problems.

4.1.2. The rate of adverse events last five years

Figure 7 show that

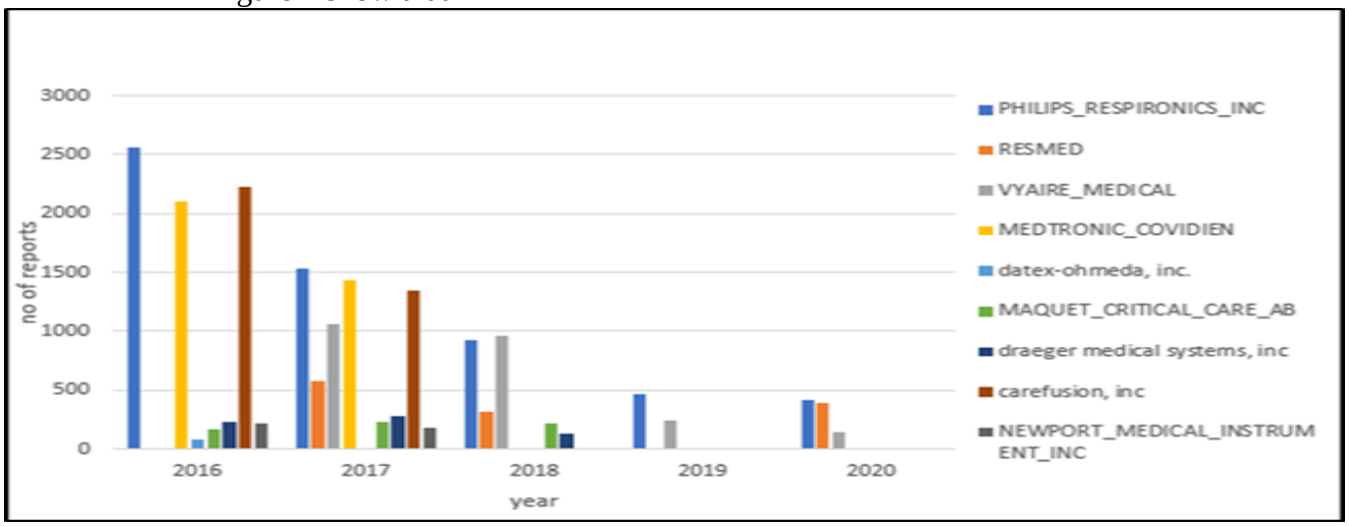


Figure 4. Manufacturers with display problems

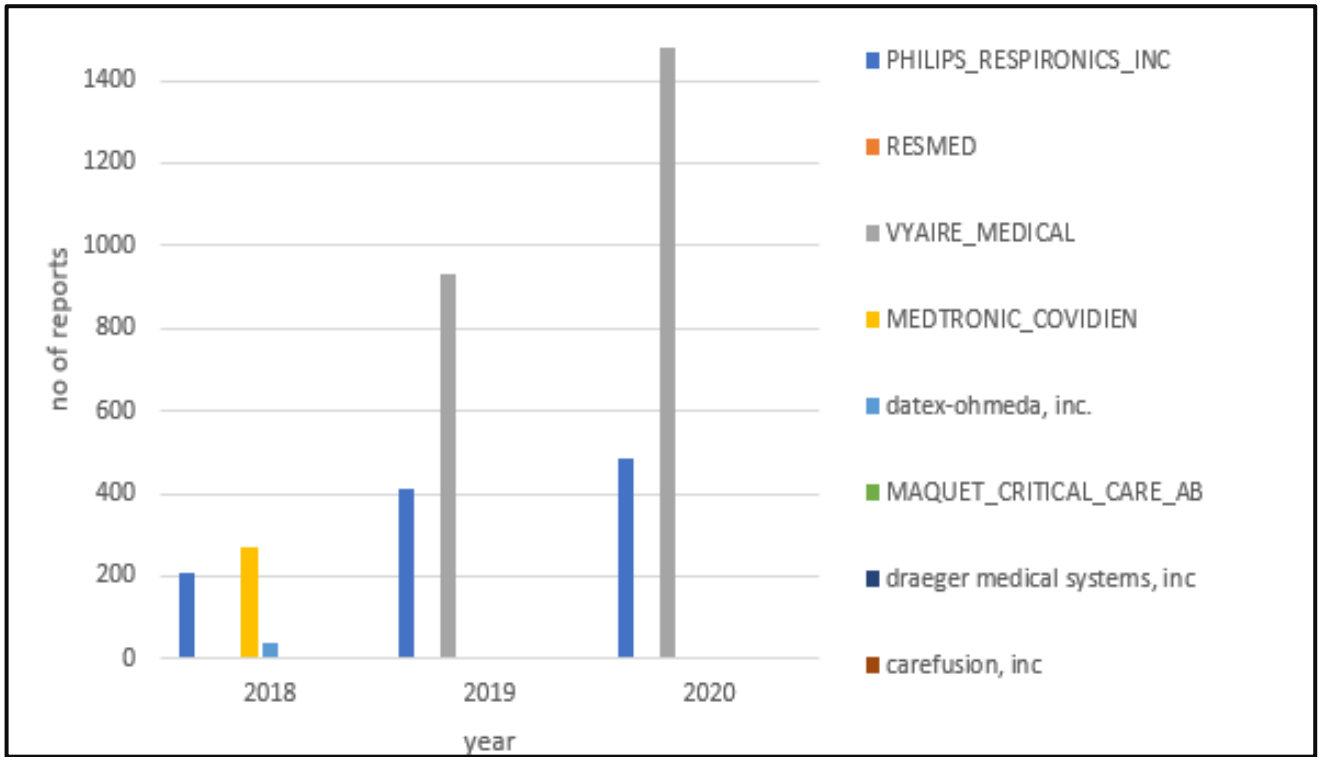


Figure 5. Manufacturers with mechanical problems

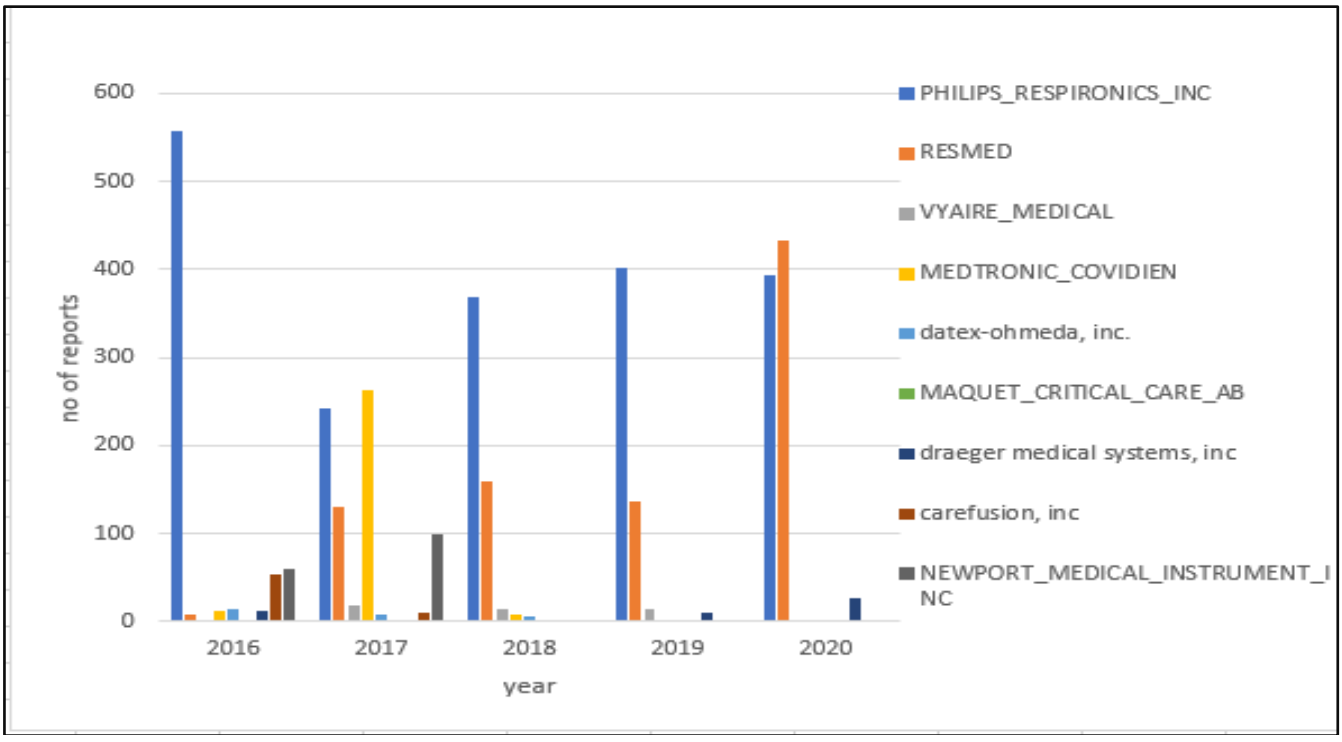


Figure 6. Manufacturers with battery problems

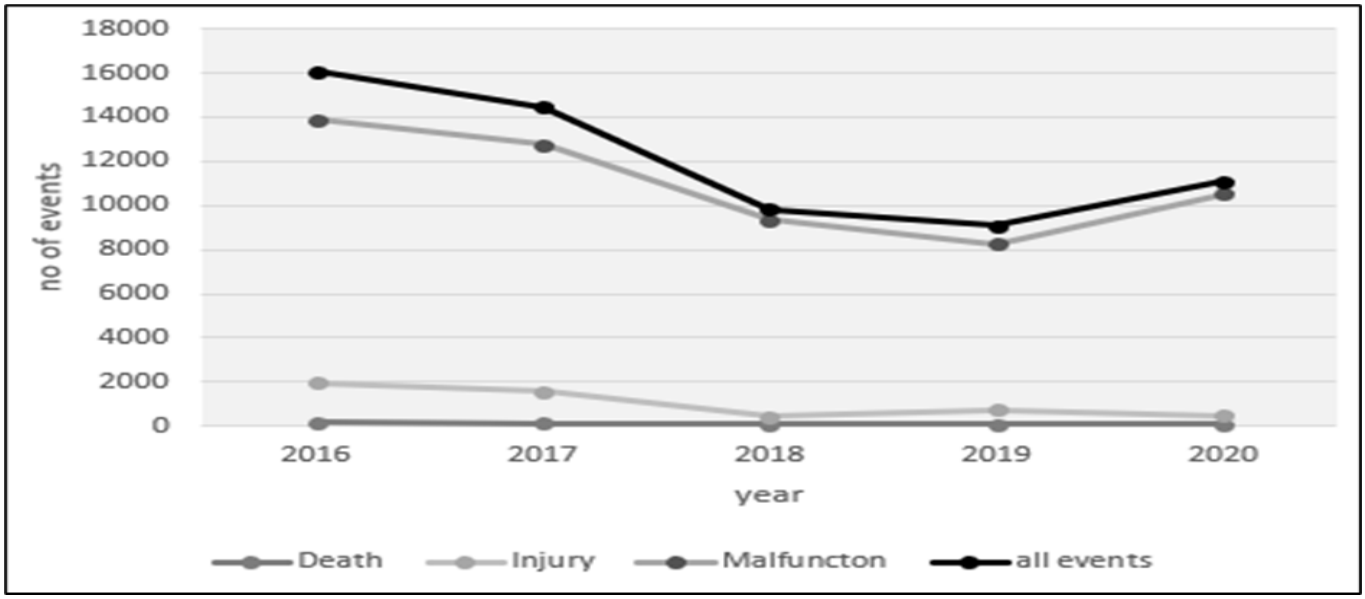


Figure 7. adverse events rate

4.2. For BTL

We found 8 cases of deaths and five of them their manufacturer was smith medical group, 60 cases injury and 54 cases of them their manufacture was smith medical group. We found over 1,174 records that 26% of device problems was break problems, 25% was alarm problems, 16% was pressure problems, 7% was volume problems (Volume Accuracy, Tidal Volume Fluctuations) and 26% other problems.

4.2.1. The relation between device problems and the manufacturers

Figure 8 , Show that.

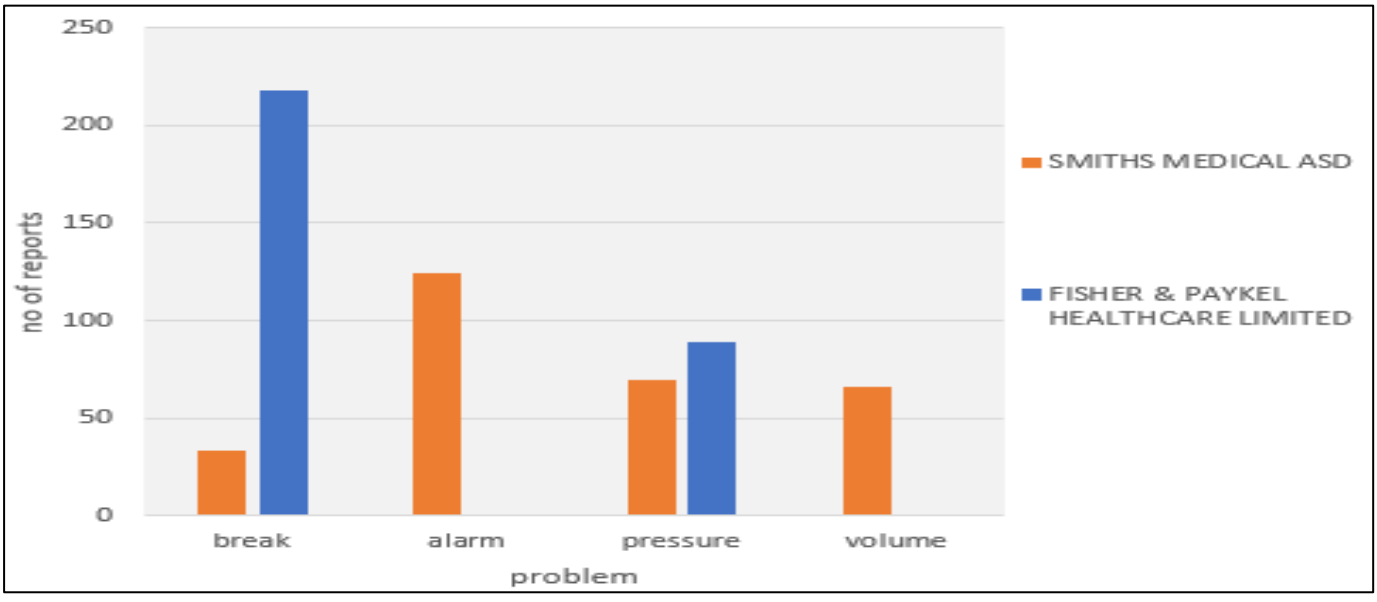


Figure 8. BTL problems & manufacturers

4.3. Manual Analysis

According to the extraction results of the device problems from the entire extracted adverse events, we got 745 different categories of device problems.
First, we selected the first 200 categories of the highest frequencies.

Second, we extracted the event description and narrative data of each device problem category

Third, after reading many reports we extracted key words and stop words of each, we found that there were three types of names of the categories

1-categories called by only one name of a device problem and a special code.

(We called them mono categories)

Failure to Charge (1085)

2-categories called by two names of device problems with a special code for each.

(We called them compound categories)

Device Displays Incorrect Message (2591); Unexpected Shutdown (4019)

3-categories of three or more device problems with a special code for each problem.

(We called them compound categories, too)

Thermal Decomposition of Device (1071); Failure to Power Up (1476); Battery Problem (2885)

The Problem we faced to categorize all the subcategories to main categories was that of how to deal with the compound categories. In many cases, the compound categories included two names or more of different device problems. So, we thought a method to reduce these compound categories in the beginning. Our method was that according to the results (keywords and stop words) of reading most of the event description and narrative reports of each compound category, we decided to append each compound category to a specific main category depending on the cause of the problem. Applying the same method on the mono categories, too.

We got 108 sub categories.

Table 2 shows an example of a subcategory and its keywords and stopwords:

Table 2. an exampleof subcatgory and its key words and stop words

Device Problem	**Keywords**	**Stopwords**
Battery problem	internal battery, replaced, external battery , error message, power, battery died, failure, reduced level, capacity, failed, charge, Inoperable, powered down	It, was, reported ,to resmed ,that ,an ,astral, device ,displayed ,error message ,There , no , patient ,harm ,serious injury, a , result ,of ,this ,incident , device , returned ,to , and , evaluation , confirmed, complaint, address , issue, serviced , fully ,tested , before ,it , customer, risk , analysis , failure , mode ,concludes , acceptable ,late ,due , transition , vmsr ,program, Establishment , unaware ,fda ,letter ,dated ,september , cbk ,procode ,status , change , in ,vmsr ,until notified , directly , december , following , submission ,our quarterly , summary , on

Finally, we could categorize the subcategories to 5 main categories:

1-Circuit Problems
2-Power Problems
3-Mechanical Problems
4-Calibration Problems
5-Software Problems

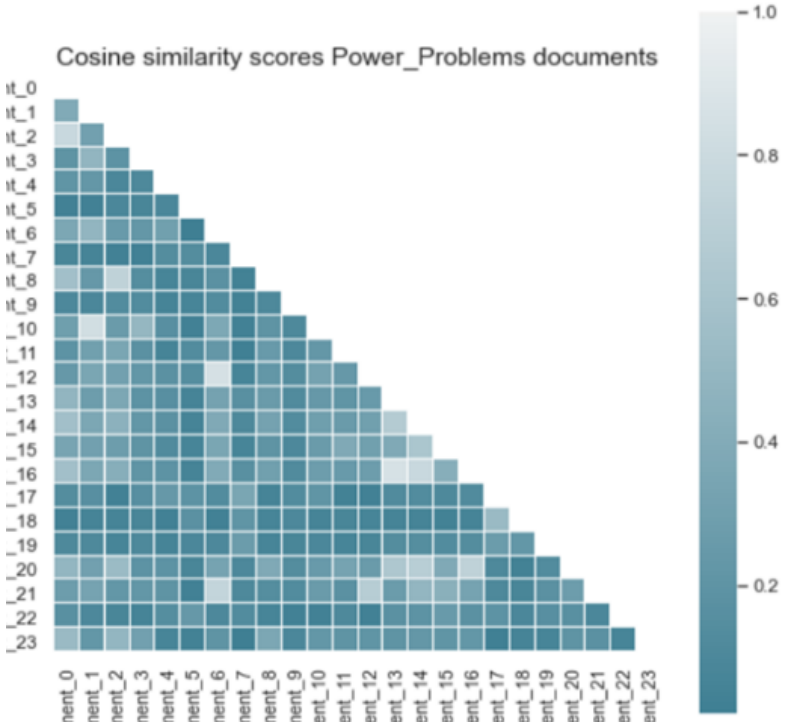
4.4. Document Similarity

"Two documents are similar if their vectors are similar". To illustrate the concept of text/term/document similarity, we will use adverse events reports with same device problem category to construct a corpus of documents. Suppose that we searched for "ventilators with power problem reports" and got back several event_description_narrative documents . We can then collect these documents and store them in a list. This will serve as our corpus.[8]

Goal: The goal here is to show how we can leverage NLP to semantically compare documents.

The process for calculating cosine similarity can be summarized as follows:

- Normalizing a corpus of text.
- Vectorizing a corpus of text using TfidfVectorizer.
- Calculating the cosine similarity between documents/vectors.
- Plotting cosine similarity using a heatmap.



From heatmap on the left side of the page, we can see that the most similar documents are document_13 and document_16. Whereas, the most dissimilar documents are the one's with similarity score of 0.0. One such example of documents that have no similarity is the pair document _7 and document _23. Shown below are the titles of these .documents

Figure 9. cosine similarity score power_problem documents

	pair	similarity
Max value	(Document_13, Document_16)	0.862946
Min value	(Document_7, Document_23)	0.021861

Figure 10. Max and Min value for similarity

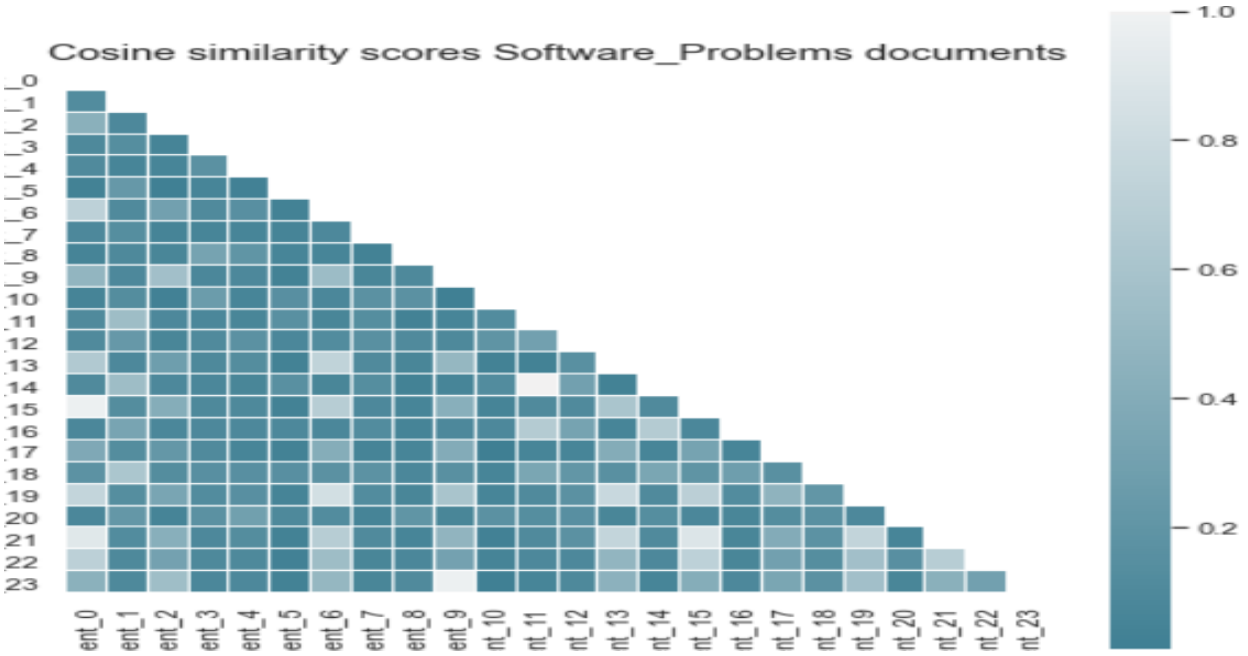


Figure 11. Software_problem documents

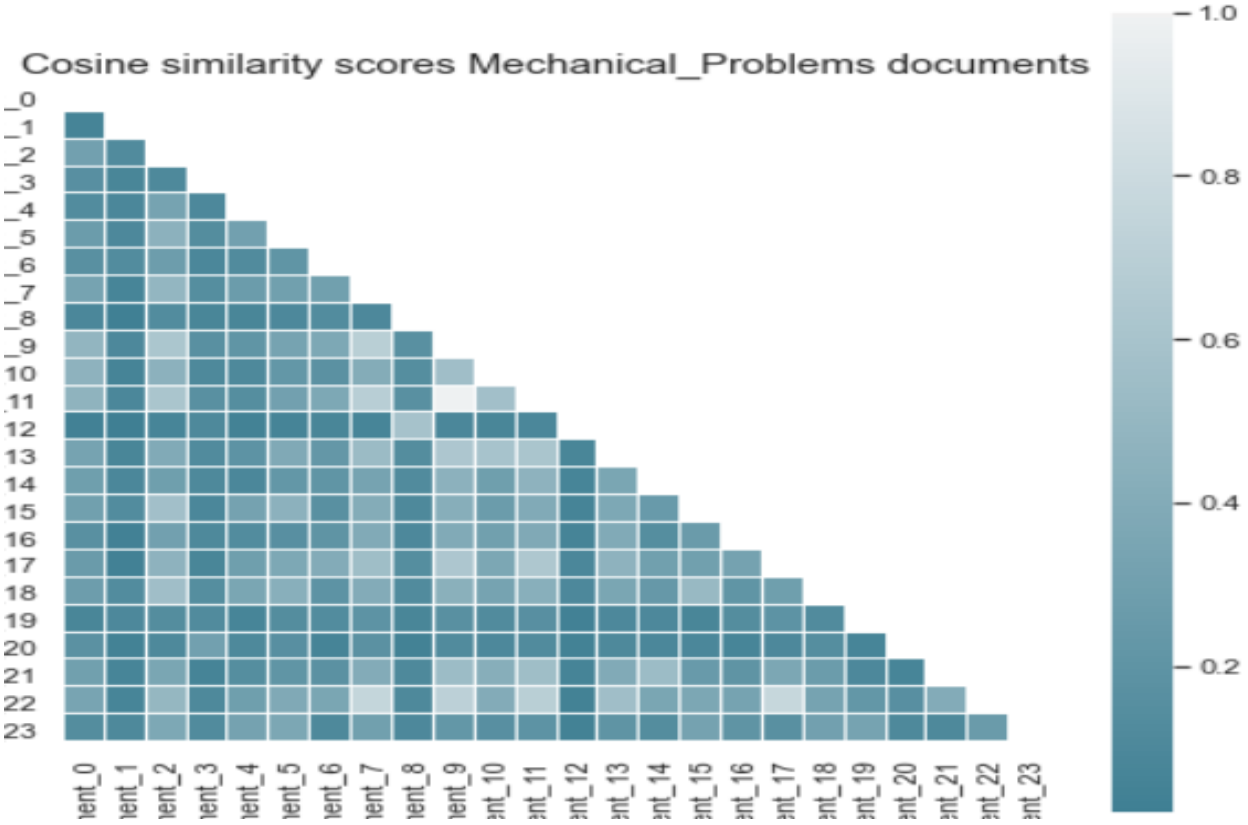


Figure 12. Mechanical_problem documents

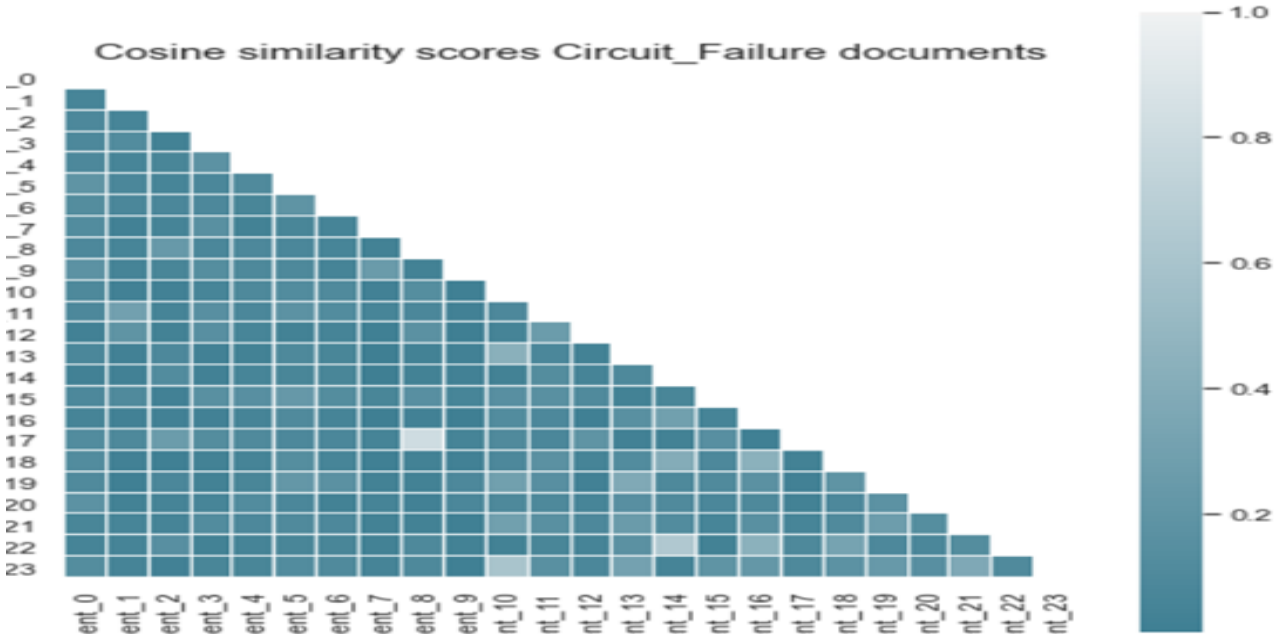


Figure 13. Circut failure_problem documents

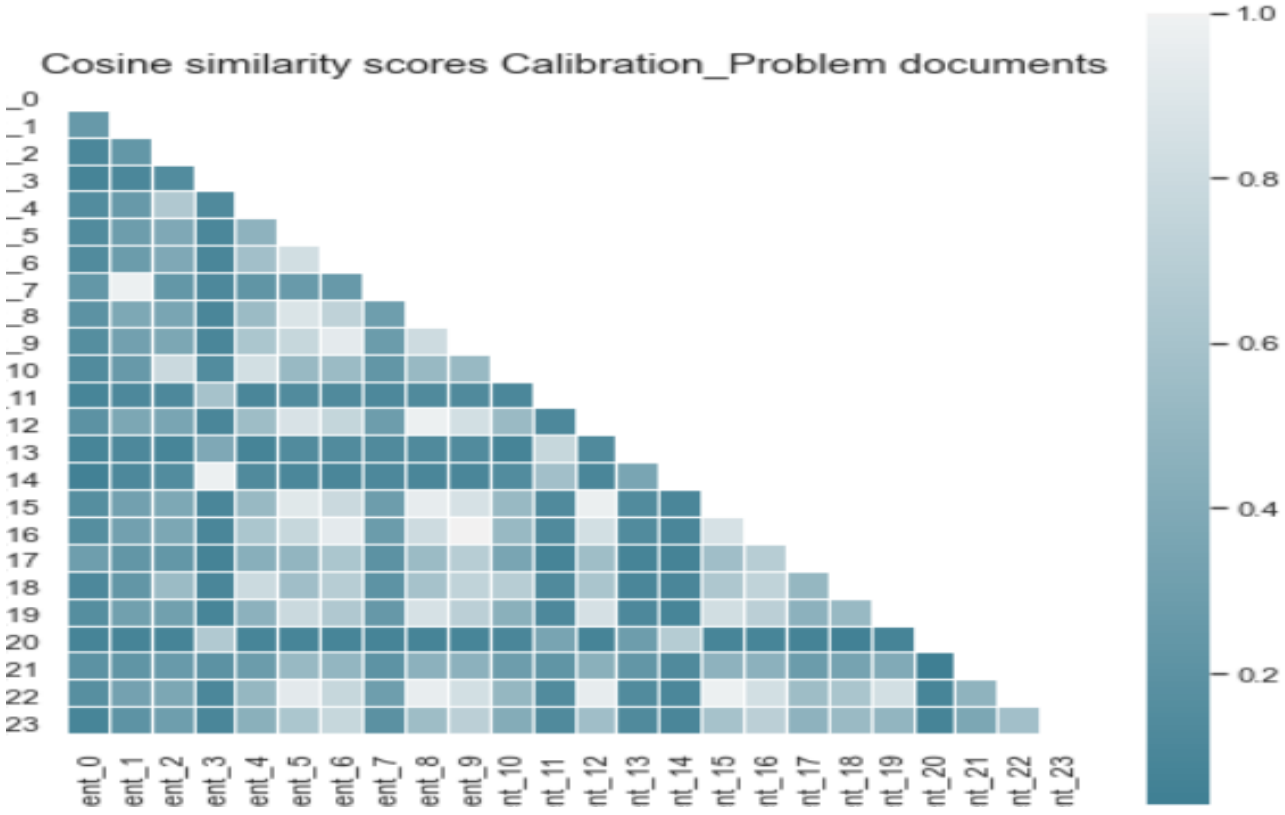


Figure 14. Calibration_problem documents

4.5. Multi-Class Text Classification

Our application is Multi-Class text classification in the medical device adverse events world.7]

4.5.1. Problem Formulation

The problem is supervised text classification problem, and our goal is to investigate which supervised machine learning methods are best suited to solve it.

Given a new event_description_narrative comes in, we want to assign it to one of 5 device problem categories [Mechanical, Circuit failure, Calibration, Power, Software]. The classifier makes the assumption that each new complaint is assigned to one and only one category. This is multi-class text classification problem.

4.5.2. Data Exploration

Before diving into training machine learning models, we should look at some examples first and the number of event_description_narrative in each class.

5383

	Device_Problems	Event_description	Narrative	Event_description_Narrative
0	Calibration_Problems	It was reported to resmed that an astral device failed to complete its internal self-test. There was no patient harm or serious injury reported as a result of this incident.	The device was returned to resmed and an evaluation confirmed the complaint. The non-return valve (nrv) assembly was replaced to address this issue. The device was serviced and fully tested before it was returned to the customer. Resmeds risk analysis for this failure mode concludes that the risk is acceptable. (b)(4).	It was reported to resmed that an astral device failed to complete its internal self-test. There was no patient harm or serious injury reported as a result of this incident.The device was returned to resmed and an evaluation confirmed the complaint. The non-return valve (nrv) assembly was replaced to address this issue. The device was serviced and fully tested before it was returned to the customer. Resmeds risk analysis for this failure mode concludes that the risk is acceptable. (b)(4).
1	Circuit_failure	The customer reported that the volume delivery of the flow valve was high on the ltv (lap top ventilator) 2150. It was set to 500 ml but was delivering 561 ml. At this time, there is no information regarding patient involvement associated with the reported event.	Vyaire medical file identification: (b)(4). At this time, the suspect device has not been returned for evaluation. Therefore, root cause has not been determined yet. Vyaire medical will submit a supplemental report in accordance with 21 cfr section 803. 56 if additional information was received.	The customer reported that the volume delivery of the flow valve was high on the ltv (lap top ventilator) 2150. It was set to 500 ml but was delivering 561 ml. At this time, there is no information regarding patient involvement associated with the reported event.Vyaire medical file identification: (b)(4). At this time, the suspect device has not been returned for evaluation. Therefore, root cause has not been determined yet. Vyaire medical will submit a supplemental report in accordance with...
			Our field service engineer (fse) replaced the air filter with a new one. The ventilator passed pre-use check and was cleared for clinical use. No	It was reported that the air filter that is connected between the ventilator and the air compressor was defective. There was no patient involvement.

Figure 15. Data Frame Exploration

For this project, we need only three columns:

"Device_problems",event_description and "Narrative".We compined event_description and "Narrative" together in one column.

- **Input:** event_description_narrative

Example: "It was reported that the pressure line pops out easily by itself from the connection port on the y-piece. Ther was no patient harm. Manufacturer ref. #: (b)(4).It was reported that the pressure line pops out easily by itself from the connection port on the y-piece. The root cause could not be determined since no product was returned for investigation. (b)(4)."

- **Output:** Device_problems

Example: Mechanical_Problem

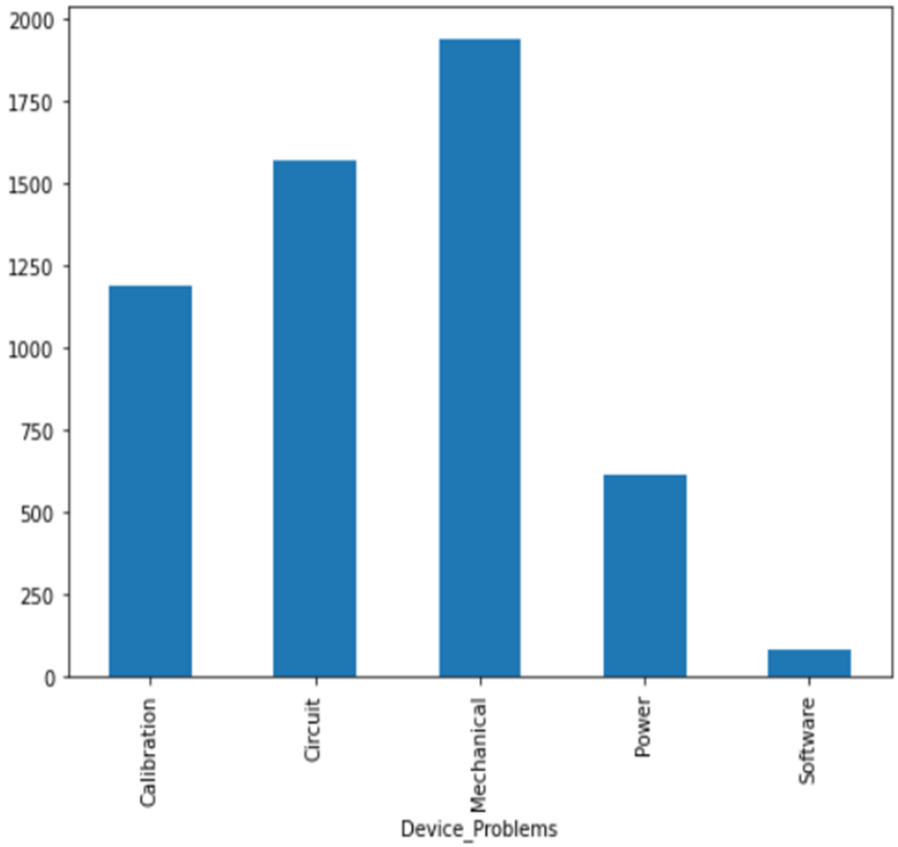
We will remove missing values in "event_description_narrative" column, and add a column encoding the product as an integer because categorical variables are often better represented by integers than strings.

this is samples for 3 rows of the data we will be working on:

Device_Problems	Event_description_Narrative	category_id
Mechanical_Problems	The customer reported blower failure and inspiration valve or device leaky active alarm on a bellavista 1000 unit during start up. A smell of something burned and a bit of smoke appeared. The customer confirmed that there was no patient involvement associated with the reported event.Vyaire file identification: (b)(4). Technical support has received the log files sent by the customer and being reviewed. The log file shows that there are multiple instances of blower failure and inspiration val...	0
Calibration_Problems	It was reported to resmed that an astral device failed to complete its internal self-test. There was no patient harm or serious injury reported as a result of this incident.The device was returned and an evaluation confirmed the complaint. The non-return valve (nrv) assembly and the main circuit board were replaced to address the issue. The device was serviced and fully tested before it was returned to the customer. (b)(4).	1
Mechanical_Problems	The customer reported that the ltv 1150 unit missing pixels. As of this time, there was no information about patient involvement associated with this reported event.Vyaire medical file identification: (b)(4). A third party service technician evaluated the ventilator. Found out that the unit was due for 10k/2yr pm. The unit need a main board replacement. Vyaire medical will submit a supplemental report in accordance with 21 cfr section 803. 56 if additional information becomes available.	0

Figure 16. Event_Description_Narrative column

4.5.3. The number of event_description_narrative per Device_problem



	Device_Problems
Mechanical	1937
Circuit	1566
Calibration	1186
Power	612
Software	82

4.5.4. Text Representation

The classifiers and learning algorithms cannot directly process the text documents in their original form, as most of them expect numerical feature vectors with a fixed size rather than the raw text documents with variable length. Therefore, during the preprocessing step, the texts are converted to a more manageable representation.

One common approach for extracting features from text is to use the bag of words model: a model where for each document, a event_description_narrative in our case, the presence (and often the frequency) of words is taken into consideration, but the order in which they occur is ignored.

Specifically, for each term in our dataset, we will calculate a measure called Term Frequency, Inverse Document Frequency, abbreviated to tf-idf. We will use [sklearn.feature_extraction.text.TfidfVectorizer] to calculate a [tf-idf] vector for each of event_description_narrative complaint narratives:

- sublinear_df is set to True to use a logarithmic form for frequency
- min_df is the minimum numbers of documents a word must be present in to be kept.
- norm is set to l2, to ensure all our feature vectors have a euclidian norm of 1.
- ngram_range is set to (1, 2) to indicate that we want to consider both unigrams and bigrams.
- stop_words is set to "english" to remove all common pronouns ("a", "the", ...) to reduce the number of noisy features.

Fig Code for converting sentences into vectors using TF-IDF

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(sublinear_tf=True, min_df=5, norm='l2',
                        encoding='latin-1', ngram_range=(1, 2),
                        stop_words='english')
features = tfidf.fit_transform(df["Event_description_Narrative"]).toarray()
labels = df.category_id
features.shape
```

```
(5383, 5145)
```

Now, each of 5383 event_description_narrative is represented by 5145 features, representing the tf-idf score for different unigrams and bigrams.We can use sklearn.feature selection.chi2 to find the terms that are the most correlated with each of the device problem:

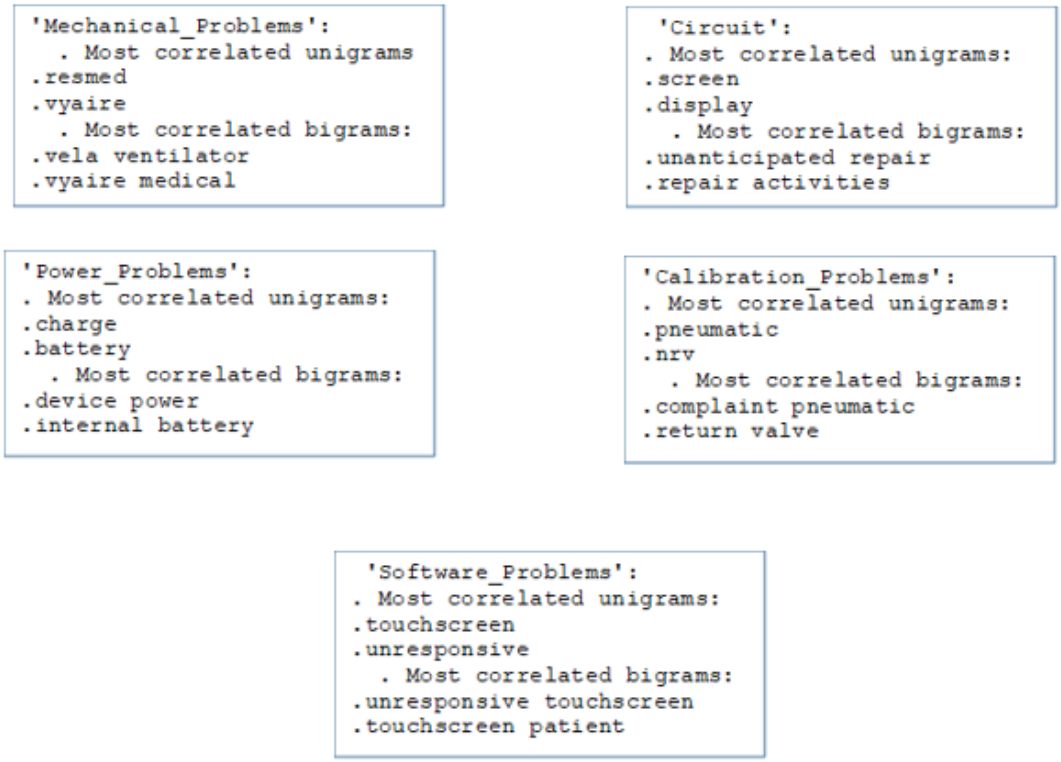


Figure 17. terms that are the most correlated with each of the device problems

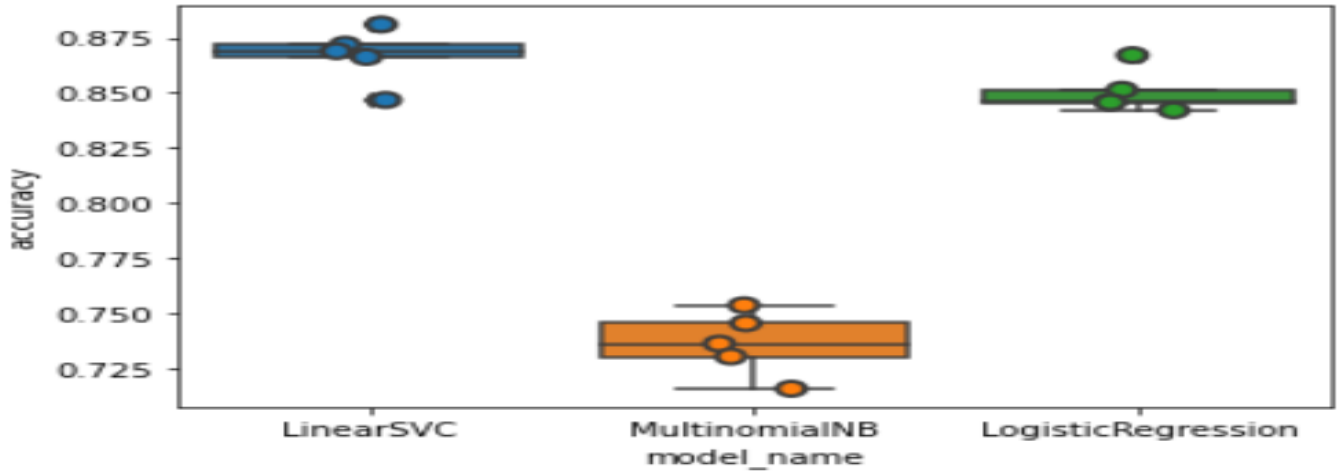
4.5.5. Multi-Class Classifier: Features and Design

To train supervised classifiers, we first transformed the "event_description_narrative" into a vector of numbers. We explored vector representations such as TF-IDF weighted vectors. After having this vector representations of the text, we can train supervised classifiers to train unseen "event_description_narrative" and predict the "product" on which they fall. After all the above data transformation, now that we have all the features and labels, it is time to train the classifiers. There are a number of algorithms we can use for this type of problem.

4.5.6. Model Selection

We are now ready to experiment with different machine learning models, evaluate their accuracy and find the source of any potential issues. We will benchmark the following three models:

- Logistic Regression
- (Multinomial) Naive Bayes
- Linear Support Vector Machine




```
pd.DataFrame(cv_df.groupby('model_name').accuracy.me
```

model_name	accuracy
LinearSVC	0.866992
LogisticRegression	0.850457
MultinomialNB	0.736394

Figure 18. Models accuracy

LinearSVC and Logistic Regression perform better than the third model, with

LinearSVC having a slight advantage with a median accuracy of around 86%.

4.5.7. Model Evaluation

Continue with our best model (LinearSVC), we are going to look at the confusion matrix, and show the discrepancies between predicted and actual labels.

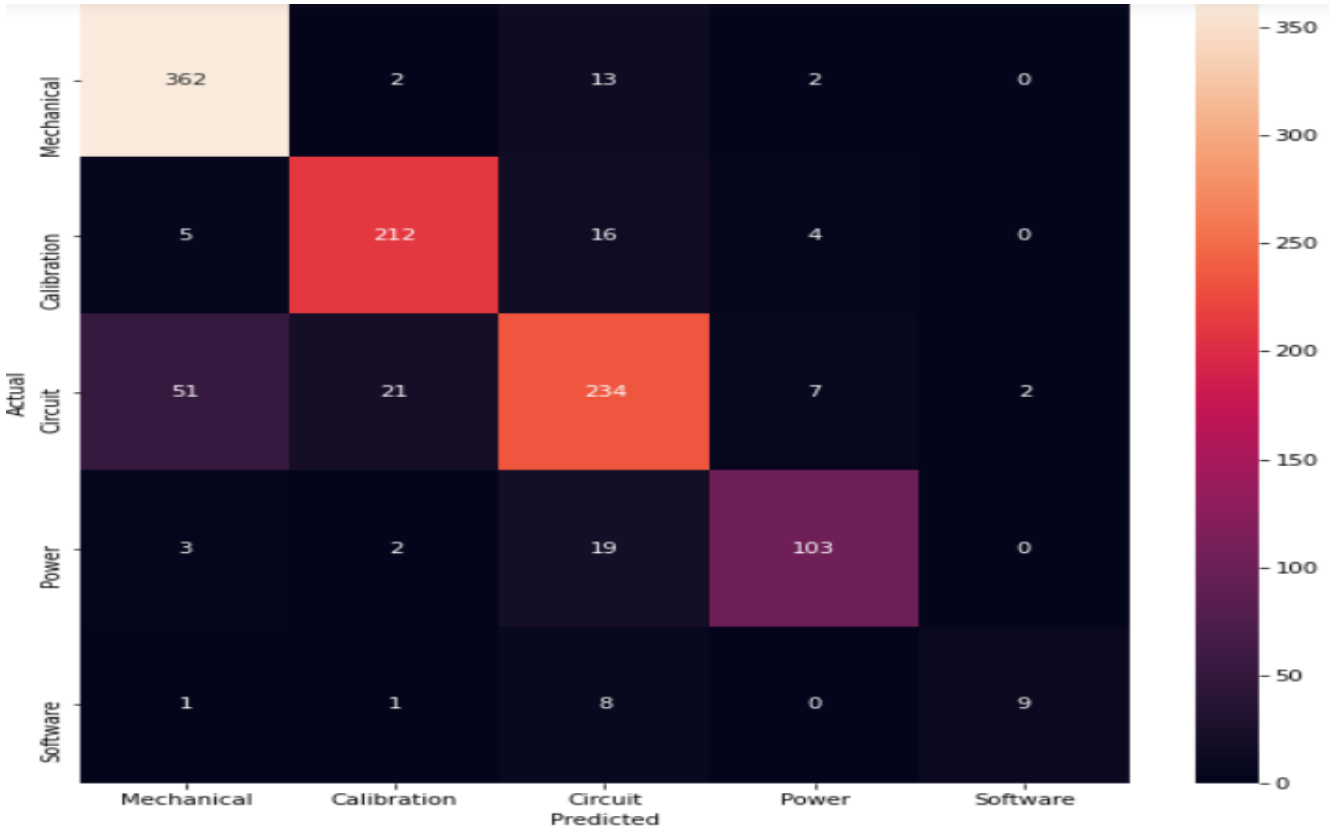


Figure 19. Confusion Matrix using SVC model

The vast majority of the predictions end up on the diagonal (predicted label = actual label), where we want them to be. Finally, we print out the classification report for each class:

accuracy 0.8542246982358404

	precision	recall	f1-score	support
0	0.86	0.96	0.90	379
1	0.89	0.89	0.89	237
2	0.81	0.74	0.77	315
3	0.89	0.81	0.85	127
4	0.82	0.47	0.60	19
accuracy			0.85	1077

Figure 20. classification report for each class

Conclusion and future work

we introduced the FDA MAUDE database which contains information about issues with medical devices that are on the market in the united_states, we introduced how to extract this data, we analyzed its structured data and we got a lot of plots which indicates to which manufacturers have specific device problem category than others, we analyzed its unstructured data manually to make data transformation, this transformation helped us to apply NLP technique for check the similarity between a document with the same class, this transformation which was based on manual technique helped us for applying multiclass text classification to predict for unseen reports to be classified with its device problem category, we end up with a model which gives us an accuracy with 86%, this model was a linear kernel support vector machine, and for future work, we'll need to reduce the loss of data during manual analysis for data transformation to get more precise results, also for future work we can work on the unstructured data analysis for more than one year and that is for gaining more experience.

Appendix A

Data extraction step_by step

Step1: download devices files from maude database

Step 2: Extract records according to product code from device files

Step3: download web addresses file

Step 4: web scraping of web address of each MDR report key to get MDR (medical device report) file.

References

1. M. Kohani and M. Pecht, "Malfunctions of Medical Devices Due to Electrostatic Occurrences Big Data Analysis of 10 Years of the FDA's Reports," in IEEE Access, vol. 6, pp. 5805-5811, 2018

2. Alemzadeh H, Raman J, Leveson N, Kalbarczyk Z, Iyer RK (2016) Adverse Events in Robotic Surgery: A Retrospective Study of 14 Years of FDA Data. PLOS ONE 11(4): e0151470.

3. S. Sagy and Y.Magdy," Analysis of Ventilators Failure Patterns", from the department library.

4. Hashem," ANALYSIS OF AUTOMATED EXTERNAL DEFIBRILLATOR FAILURES IN THR FDA MAUDE DATABASE"

5. MAUDE Adverse Event Report: VYAIRE MEDICAL VELA VENTILATOR VENTILATOR, CONTINUOUS, FACILITY USE (fda.gov)

6. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm

7. https://analyticsindiamag.com/multi-class-text-classification-in-pytorch-using-torchtext/?fbclid=IwAR2AKjqCfJ3rDUS3wSD006fTU7UHizfyMAMd5kM02XkBzh2iiJVFN3rhlKY

8. https://www.nltk.org/_modules/nltk/tokenize/regexp.html?fbclid=IwAR2CTTyxbn4kqkQzsZxdAXFVfxXi6epc_khUklOCKNYq5l_fw3ClTbCFCp4

9. https://searchenterpriseai.techtarget.com/definition/natural-language-processing-NLP?fbclid=IwAR2FJcXlnKOWfrtuv37Zt_Tg-cPcjvbobI3umls6S59gWr89nNNg0bl5Aoc

10. https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/?fbclid=IwAR3r6wYPk5UkDaOPBDiPzyaHqJsecs2JaA5YOq4PDABvrHiqKYvPHmmquSU