Understanding customer complaints
a data-driven approach

Vagionitis, I.

Award date:
2019

Department of Mathematics and Computer Science
Architecture of Information Systems Research Group

Understanding customer complaints – a data-driven approach

Master Thesis

Ioannis Vagionitis

Supervisors:
dr. N. (Natalia) Sidorova (TU/e)
ir. A. (Angelique) Brosens-Kessels, PDEng (Philips)
dr. L. (Laura) Genga (TU/e)

Final version

Eindhoven, August 2019

Abstract

Handling customer complaints is important for every business that provides products. Focusing on the complaints and using them as data to find how customers think about a product is one of the reasons that a company spends a lot of time and resources to understand them. Their content is used as an input in the development process of the product and is needed to know what its impact is. How it can be found even and noted in case there are many complaints for a product from a specific domain is a challenge. The traditional way of handling complaints one-by-one requires resources and it cannot identify the impact of a specific topic. This thesis provides a data-driven approach to import the customer complaint handling process and make a step closer to create a classification that can constantly involve and assist the development process of a product. The approach was applied to customer complaints for the product of Philips Azurion, to create an analysis of them and achieve getting collective insights.

The main research goals (1) identify a clustering technique for the complaints and (2) identify insights for the clusters complaints that could integrate new content, have been reached. A transformation of the customers complaints was applied with the use of the technique Bag-of-Words from text analytics that uses a predefined set of words dedicated for the investigated product. Clustering was achieved with the use of a hierarchical clustering technique and the use of a metric that used as value the outcome of the text analytics. A combination of data graphs and text analytics are providing insights from the customer complaints. Every step was achieved with the use of Python scripts that use as input raw data and provide clustering and insights.

To conclude, the results for the approach have shown that complaints can be clustered and get insights. Using the clusters in the handling process is going to increase efficiency. At the same time creates a reference for the future cases from the suggested method for insights. These can provide a description of the topic with the use of the data mined from the complaints as it is presented in an example which is presented. Complaint handlers can get use the approach to semi-automate parts of their work and assist more efficiently the development process of the product.

Preface

I would like to express gratitude to my supervisor in the university Natalia, my supervisor in the company Angelique, third committee member Laura and my colleague Raisa who assisted me during this project. Their assistance was very important and help me a lot to overcome every obstacle. I would like to thank Philips for the opportunity to do the project in a real working environment.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

The current document is going to present the process and results of my thesis for the masters Business Information Systems at Eindhoven University of Technology (TU/e) with the guidance of the group Analytics for Information Systems (AIS) from the Mathematics and Computer Science department. The project was supported by Koninklijke Philips N.V. and conducted in their offices. The research was focused over complaint handling process and identifying possibilities for its improvement. Through the research we aim to provide insights that can create benefits for the business.

This chapter is going to introduce the context of the thesis in its first section. The following section is going to introduce the research goals that are going to fulfill the needed task . The next is going to explain the methodology that have been followed in order to conduct the research and finally the outline description of the document.

1.1 Thesis context

Innovative industries need always to provide their customers with products that are being constantly developed to cover their needs. By getting feedback from customers they also get new ideas in enchant product's functionality, performance and usability. Handling customer complaints can contribute significantly because their content describes customers' needs. From the complaints' content, it is possible to mine the most important needs and create or upgrade them. Product's users are those who are going to interact with it and can provide the information needed to know what is important to be added in the product and what is redundant in its current form.

A product's design from scratch is a challenging task for system designers. It is even more difficult when the product is intended to be used to perform activities of a different domain of knowledge and the designers need to learn information for the unknown domain before doing their jobs. Domain experts are being summoned to share their knowledge to overcome this challenge. Often businesses have already produced in the past products that provide similar functionality. If older products have been proven successful its insights are being used as a reference of design for a new innovative product that is going to satisfy their customers.

The research investigates an approach that makes use of available data from complaints that are part of the communication of the business with its customers that could introduce insights for product development. The approach included the use of combined innovative techniques used in a way that can produce insightful results for the business.

The project is going to be based on the newest product of Philips IGT Systems. Their solution is a complex system that is used mainly in a hospital to perform a big variety of surgeries. The Azurion was released in 2017 and until now there are installed more than 400 systems around the world in multiple hospitals and clinics. Its main characteristic is the use of X-ray to guide catheters in inpatients and perform the appropriate medical intervention. The Azurion is a system/product that is being distributed throughout the whole globe. Its development need to take in consideration regulations and culture in every different place and make the same product for all to use.

As in every system that is produced from Philips there are received complaints about various reasons. Some of them could have a great impact on the functionality of the system and may even affect the patients. Complaints handling is an important operations that focuses not only on overcoming possible unwanted behavior of the system but also on gaining insights for future development. Complaints are received by the customers or by close monitoring of the clients via sending employees for maintenance operations.

1.2 Research Target

Philips has defined hazardous scenarios after observations and input from the available complains. They are not in a position to know whether a problem has occurred in a system if it was not included in a complaint. If a reported situation has occurred in multiple places makes it easier to structure the tasks that need to be focused first and resolve them as soon as possible. The research is made over the available data from complains that may assist to reduce possible hazardous situations and provide new features, that would increase customer satisfaction and reduce risks for undesirable accidents. The following statement will be used to define research goals for the current research:

- Approach should make use of available feedback for health care devices in combination with previously recorded data to increase risk analysis understanding.

To achieve getting results from the research is going to be efficient to follow steps that their combination is going to introduce a method for processing similar data in health care. It is needed to process natural language text and provide useful insights from it in the final stage of the project. Reaching that point needs several steps that can be split into more concrete goals.

Research goal 1: Identify a clustering technique for complaints.

There are going to be multiple complaints filed as long as a system is operated from a customer. Clusters are formed from similar ones that are going to show the impact and improve the process of handling them. How it is possible to make it by using the initial content and making a cluster when using it as a reference.

Research goal 2: Identify insights for clustered complaints.

A comparable format of the complaints that leads to clusters creates opportunities for efficient processing of the data. Using the clustering outcome in combination with other techniques to provide possible benefits. How this can be applied and what actions are needed for valuable results is going to be the challenge.

Following the direction that the goals set it is going to be introduced a new complaint handling process. It is going to include multiple parts and iterations that could ensure validity and reusability with any other input with similar challenges.

1.3 Method

From the beginning of this research, the methodology of CRISP-DM [38] was followed. CRISP-DM (Cross Industry Standard Process for Data Mining) defines six stages that may exist in every data mining project. This project followed the guideline of CRISP-DM and in the following section it is going to be described how the methodology was applied in the research.

Figure 1.1: Circle of CRISP-DM [38]

The six stages of CRISP-DM can be represented from a circle that is illustrated in the above Figure 1.1. They are placed in such a way to show the need of multiple iterations in a project that the preparation is made and applied later but depending the results there are needed adjustments to achieve the best possible performance from the outcome of the research. The first stage is taking part in the circle and the last stage introduces the outcome in the appropriate format for each situation.

Business Understanding

In this stage, the research goals are going to be identified and also the field of content that is going to use while performing the research. As a text, it describes the introduction and refers to the current section of the thesis.

Data Understanding

Data understanding includes actions that took place to understand the content of the data. Being familiar with the data which in the current research are complaints from the healthcare industry does not include only the structure and format but even the situations that are referring that is currently a medical system which is used in hospitals and clinics. The information from this stage are presented in Chapter 3.

Data Preparation

Raw data may need additional actions to ensure their integration in a suggested technique. These kinds of activities that have been performed are described at the beginning of Chapter 4. This stage is connected with the next one of Modelling because there are usually performed multiple iterations to increase efficiency and develop the approach of the project.

Modeling

In this stage that data is transformed with the use of the techniques of text analytics and machine learning. This transformation is going to produce the output that provides the answers for the research goals. As already mentioned above, there are going to be multiple iterations of this stage. Information for this stage are going to be presented in Chapter 4 and in Chapter 5.

Evaluation

The fifth stage includes activities to ensure the validity of the outcome. In this approach, validation will use some values from the data to validate if the outcome is accurate. With these actions, it is going to be introduced also the impact of data in the business. This stage is going to be complementary to the clustering results and discussed in Chapter 5.

Deployment

The final part of CRISP-DM is going to present the business value of the data in a way that it will be used from the company. The main outcome of this activity is this document that describes everything that happened during the project.

1.4 Outline

This section will briefly present the structure of the thesis. The second chapter will provide information for all preliminary concepts related to the project. For each, there is a section that presents their aspects focusing on topics that make a significant influence in decisions taken while working on it. The third chapter describes the case related to the project. The description is wide and covers of aspects in the business that provide the data of the project. The fourth chapter introduces the step that should be taken to follow the approach and achieve getting the format of the data that are required to get further. Data preparation and text analytics are described and discussed. The following chapter is going to show steps of the approach that can provide answers to the research goals of the project. Hierarchical clustering and data graphs are presented. In chapter 6 there are included discussions over other existing approaches and limitations that may exist. Finally, there are written conclusions related to the project. Moreover, in the same chapter is introducing possibilities for future development and limitations for the approach.

Chapter 2

Preliminaries

This chapter is going to describe briefly some concepts that already exist and used to assist this project. Firstly, there is some information about complaints handling in the current state. Next, it will follow up some information for Text Analytics that was used during the research. Finally, there are going to be described concepts of Machine Learning.

2.1 Customer Complaints Handling

The following section, it discusses aspects related to complaints handling. Firstly, it is presenting the importance of effective handling for a business. Effective complaints handling has a serious impact on the relationship with the clients and their satisfaction. Secondly, there are going to be discussed issues and concerns of current complaint handling systems. There are already multiple approaches [31][24][21] which are currently used in different businesses.

2.1.1 Impact of complaints handling

Handling complaints can provide a serious impact on the relationship of the customer with the provider [41]. A customer that creates a complaint can refer to a variety of topics. There is the possibility of a malfunction of the product which may or may not have an impact on product usage. Also, the complaints may be made to suggest a feature that the client would like to see in the future; and believes it can increase the value of the product's usage. The way that a business decides to act in such a situation creates different impact to the relationship it is going to have with the clients [16].

Product buyers feel more comfortable getting a product that comes what good support than getting a similar product which includes more features but no support in case something unexpected happens [41]. Their loyalty goes to the business that can provide solutions. The occurrence of a product's malfunction is accepted but if the solution is not provided the negative impact is high. The impact can be described simply by considering the rule of "Word of Mouth"[5]. Domain experts communicate to exchange their opinions on various topics related to their work. For example in the medical sector, there are conferences that medical experts gather and discuss various medical topics they are also discussing the products they are using to make their work easier and efficient. A doctor that is dissatisfied with a product is going to discourage others to use it, which is a negative impact on the product. On the other hand, if the customer is satisfied with the usage, functionality and support it is going to create a positive impact and most probably the user is going to promote it[5].

Other aspects of complaints handling impact are the opportunities that can be mined from their content. According to Frazer [46] there are multiple fields of the business that can gain benefits such as research & development (R&D) and quality assurance (QA). R&D can have great input in their production pipeline with ideas that come straight from the product users. QA could identify and include scenarios in risk analysis that have not been introduced from the internal working team. Such actions are going to create an image of a trustworthy business that respects its clients and continuously take steps to improve their services and products [47]. On the opposite businesses that are not in a position to grasp and take advantage of user complaints, tend to have a lack of a systematic approach in their process that projects over their clients as lack of respect to them [47].

2.1.2 Other approaches for Complaints

A different approach for handling complaints can be seen in most of the organizations. Some of them follow published guidelines and techniques that have been adjusted in their needs. This research introduces an approach base on the case that is described in Chapter 3. The following paragraphs will discuss aspects of different approaches that can be related and assisted to define the steps followed in the research.

The approach of the computerized complaints handling which is introduced from Mitchell [31] suggests an automated system that can have adjusted content with the same functionality. It introduces fixes data values that can be filled and input a complaint in the system. The system contains an archive of reference options that assist with the difficulty of handling the manual records. Taking advantage of this functionality introduces multiple tools that can output report surveys and other statistical results for the complaints made. Although it can be great tool aspects are missing from this approach such as unique situations that have not to be identified and classified as a reference.

The content can provide advantages from its information according to the methodology that is introduced [35]. Making a brief analysis but selecting to handle only those that are expected to provide opportunities for the business. This selective methodology creates challenges discussed in the previous subsection that could affect the reputation of the business to its clients. Similarly, in the case of e-CCH systems [24] that introduces a case-based system with already identified situations that may exist and make a classification of the outcome that happened. Following this technique, are introduced difficulties over the understanding of the cause that the customers were provoked to submit a complaint.

2.2 Text Analytics

The techniques which can be placed under the title text analytics provide the functionality to process text data and return information. It is also a topic related to natural language processing that transforms the text into data [29]. Thought this subsection it is going to be discussed a technique process a text in a format that can be uses for data analysis. Finally, it is going to be discussed the domain knowledge impact.

2.2.1 Natural language processing

Content's text understanding is the initial meaning of Natural Language Processing (NLP) [37]. It refers to all techniques that can assist in making use of text information from computerized systems. One of its concepts that is involved in the project is the natural language understanding.

Data that contain natural language is not possible for a system to understand and easily use them. The difficulty comes from the complexity of the language. There are many cases that the same words are used to provide different meanings and it is also influenced usually from the content before and/or after [37]. One of the existing techniques to handle natural language is the Bag-of-Words that is going to be discussed in the following paragraphs.

2.2.2 Bag of Words

A text has the goal to describe a situation or provide information on how to do something. One of the most commonly used applications of text is searching for information on the internet by using a relevant keyword [40]. This is achieved by applying text analytics over multiple data in order to identify if they are related to the input. On the other hand, some applications follow this process with different goals and sequence of actions. For example, if it is needed to predict the behavior of individuals, text data from social media are analyzed to achieve it [19].

Taking into consideration the previous paragraph it is possible to state that text analytics is the intermediate step between a written text and its data-like format. The concept of bag-of-words is a representation of a text with only some selected words that are targeted. A text is scanned and if there is a match of a specific word the vector position of the work-related with the complaint increases [3]. An example is in the following Figure 2.1.

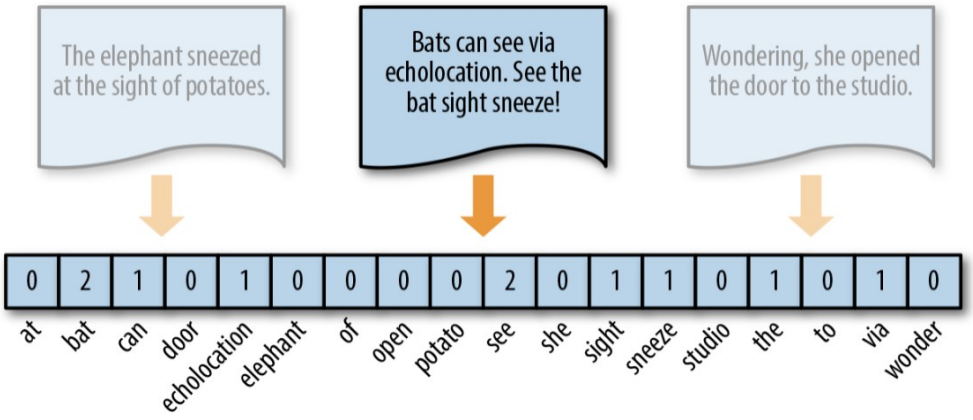

Figure 2.1: Example of Bag-of-Words representation [14]

Through this technique, it is possible to see the content of a text in values but also identify the frequency of words in a text. Taking into consideration that a free text is unstructured there are also frequent words used that may create noise in the data. The 'Term Frequency - Inverse Document Frequency' (TF-IDF) [34] concept provided a statistical approach of 'removing' words that are too frequent in the text since they are not providing value. As a concept works with the percentages of word appearances in a document and its rare presence across multiple documents. These rates end up providing a ranking of words that could potentially provide noise [28] in case there are used.

2.2.3 Domain knowledge impact

To process text data it is important to understand its content. In the case of some fields of expertise, it is needed to know what is going to be described from the text. When a text is analyzed and extraction of information is done but in cases that the goal is a specific type of information, for example, content the refers to components of a product text analytics are limited. An attempt to introduce the concept of dedicated libraries was in 1997 [10] with an approach that performed a discovery based on relevant domain knowledge. Text data have much information to provide but it is a challenge to find only the information that is needed. Making that choice requires understanding in the domain that the text is originated.

On the other hand, text analytics can provide multiple benefits for product development and decision making [30]. Text that targets a specific domain can have characteristics that define needs and wants. Being in a position to take advantage of this information provides the opportunity to make targeted business development decisions. An example from a digital content provider which analyzes in customers feedback data to enrich available services according to their wants [19].

2.3 Clustering analysis

This is the task that creates groups of data in a way that all data of one group are mostly similar between them and not with data from other groups [22]. Defining what is similar and what is not it is done in multiple and different ways. Many algorithms [8] use statistical distributions which provide cluster based on different characteristics. The main focus is going to be on the technique of hierarchical clustering and its functionality that is based on distance connectivity.

2.3.1 Hierarchical clustering

As its name implies its operations are performed in a way that the data are arranged in an arrangement. This means that all of the data are somehow connected but their connection is not always direct. There are two ways that the hierarchy is achieved.

- Agglomerative: bottom-up
- Divisive: top-down

Both methodologies are achieved with the use of a greedy algorithm [4] which means that they are always trying to achieve the global optimum for their results. The Agglomerative considers all data as single clusters and tries to make pairs of these until it reaches a point that they are all connected. On the other hand, Divisive considers data part of one cluster in the beginning and the split is achieved recursively to achieve the hierarchy concept. The following Figure 2.2 is going to show the difference between the two methodologies.

Also, the Figure 2.2 is a typical representation used in hierarchical clustering and is called dendrogram. The graph shows the connection between data and how they are connected altogether. This kind of graph can also include additional information in x-axis and y-axis. Usually, x-axis provides the id of the element and y-axis provides the distance value between the connection points of the data.

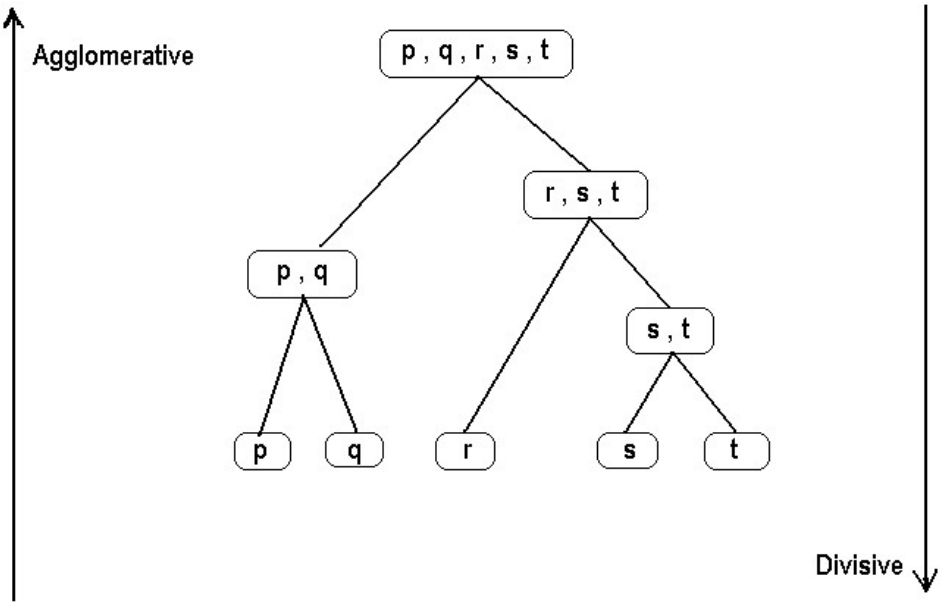

Figure 2.2: Example of Hierarchical representation [2]

The most commonly used is the Agglomerative method. To perform the clustering operations two types of equations are used. Firstly, the metric that calculates the distance between data. Secondly, there is the linkage that is responsible to handle the connectivity between cluster with the use of the distance values from the metric [42].

The following is the euclidean distance equation that is commonly used for the calculation of the metric [6]:

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \tag{2.1}$$

To calculate the linkage there are available multiple criteria. Ward criterion is the most commonly used and its functionality minimizes the variance between clusters [44].

2.3.2 Other Clustering techniques

The following Figure 2.3 is going to provide a graphical representation of the differences between some clustering techniques. Each one has different characteristics and creates clusters by following different guidelines that influence the outcome of the process. A short description for each one is going to be provided in the following paragraphs.

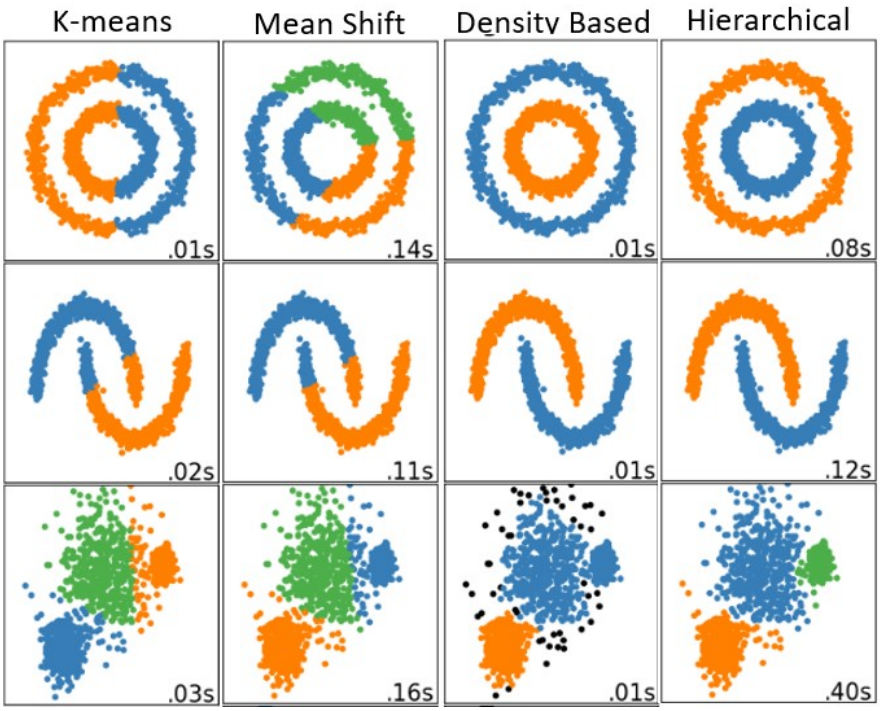

Figure 2.3: Comparison for clustering techniques [1]

K-mean

This clustering approach is trying to create clusters with the use of vectors that separate the data. It calculates the best average values of distances between the data point in an iterative way until the best possible clusters are achieved. It does not consider data as unique and tries to make use only of the cluster centers from the groups of data that vector(s) separate [27].

Mean-Shift

Mean-shift tries to locate high-density data centers for each candidate cluster. In every iteration, it tries to focus the candidate centers until there are representing the focus of all surrounding data points. Duplicate centers are removed and the remaining one represents the center of the new clusters [13].

Density-Based

This algorithm does its calculation by visiting the closest neighboring data point. Every time there is no close enough neighboring point it considers that a new cluster is formed. In case there are not enough neighboring point to form a cluster all the leftovers from the data are being considered as a cluster of outliers that do not fit in any of the previously formed clusters because there are not close enough [25].

Chapter 3

Case: Azurion

The approach was based on complaints related to the Azurion systems which is an innovative product of Philips for medical usage. The following section will present related information that will assist in the understanding of complaint content and the current complaints handling operations. The complaints are part of a complex procedure the company follows and integrates and uses them to increase the value of their product.

- What is Azurion?

Azurion is an iXR system that is used in hospitals and clinics. iXR stands for interventional X-Ray which is a technique for diagnosis and treatment of a disease with the use of images generated with radiation. Radiation is emitted to the patient from the X-ray tube of the system and with use of monitors and other medical equipment doctors can treat patients with diseases such as aortic valve stenosis which is located in the heart and with the use of Azurion the treatment can be applied without an open heart operation.

The patient table is one of the main components of the system. The patient can lay down and the operator can adjust its location mechanically with the use of the Table Side Operator (TSO). The TSO is a collection of buttons and joysticks that can perform various tasks related to the patient and the stand. On the other hand, the Touch Screen Module (TSM) is a table-like module and can perform all X-ray related tasks, like beam collimation and image adjustments. The stand is a C-shaped moving component. The stand can be moved and angulated in order to capture clinically relevant images from the patient. It contains the X-ray tube and the detector. Flexvision is called the main monitor that displays captures and other information even from different medical systems that are being projected. The last main component is the pedals' that is a separate moving device that activates the X-ray emissions. Multiple secondary components are intended to assist the operation such as lead shields that reduce X-ray spread over doctor and technicians [43]. In Figure 3.1 is a representation of the system.

Figure 3.1: Azurion System overview

Complaints that are being received for Azurion systems are part of the systems development which will be described in the following subsection.

3.1 Development Process

As in every big organization in Philips, it has defined already a sequence of actions to perform for the development of the system. There are defined some states that are followed in order to identify the focus of the next project and pick the content that is needed to be changed or developed from the beginning. Giving a brief description there is a main pool that features descriptions are stacked that is fed from three different sources of information. Another state is also connected with all the mentioned components that evaluate the severity of each item and defined the priority in the pool.

- Database of complains

The database contains all available records of the complaints about the systems that have been placed in the field. Complains can be separated in the complaints into two different groups. The first group is all of those that ask for an enhancement of a system feature which is usually contains an idea of expanding the functionality. The second group is related to events that occurred in the systems already placed in the field. Whether it happened in the middle of a procedure or without a patient the operator of the system can contact customers' support and create a complaint. The main difference between the two groups is the outcome of each situation. The first group does not have any outcome to report because there are no concrete data that may represent it. On the other hand, the second group has to include the information that is going to describe the severity of the problem that has happened. Complaints can be filed either from the customers or from people related to the development of the system.

- Usability studies

There are performed studies in order to identify the next steps in the development of the system. The main contribution of those is to research new functionalities and features that can be developed and included in future releases of the system. This is done from a group of people that perform research over the systems with the goal to make it efficient and user-friendly.

- Product monitoring

Many visits are performed in the location of the clients. whether if there are made for maintenance or for a different reason the field service communicates with the users and monitors the interaction with the systems. Every possible information is recorded and analyzed to achieve getting input in the development from the field and understand better the need of the customer.

- Evaluation

It is an intermediate step before all the suggested or defined from scratch features are going to be prepared for development. Depending on the type of the feature there are placed in a different position in the stack. The ones that are very important and may have as an outcome the harm of a stakeholder of the systems there are prioritized and marked for immediate development. Furthermore, all new suggested content is evaluated and placed in a position depending on its functionally and business plan for the system.

- Content for Development

This part takes the role of a pool that contains stacked many features that are going to be chosen for development. It has input from the Database of complaints, usability studies and Product monitoring. There is an intermediate step of evaluation that decides the position in the stack each input feature is going to be placed. Every time a new project is started from this pool thee are chosen contents that are going to be included.

3.2 Complaints

Multiple aspects are taken into consideration for the content of the structure of the complaints. The following part is going to present all the information that can be related to the complaints that are being filed for Azurion systems. Firstly, the stakeholders will be presented and afterward their structure and handling methodology which is currently used.

Three main stakeholders can file complaints. Firstly, are the doctors who perform medical procedures with the use of the system. The content of their complaints is describing things related to medical topics because of their expertise. Another type of stakeholder is the technician from hospitals and clinics that stand by the doctor and operate the system. Their focus is more related to the functionality and ease-of-use aspects which is their job. Latter ones are Philips employees, they are making visits in locations systems are installed and use to monitor the situations and identify future needs and opportunities for ideas to include in future development. at the same time, they are checking if the system operates smoothly. Similarly, field services engineers when there are going for planned maintenance. Their view of the system is significantly different from others. The majority of them have an engineering background and focus on the perspectives that can enhance the system in functionality and usability. Having three major stakeholders that can file a complaint it is going to have an impact on the content focus.

PR ID	Short Description	Customer's Problem Description	Customer Communication	Initiated Date	Date Opened	Initiator Name	Source System Institution	Source
6873756	Distance meas	FeedbackDescripti	Dear	23-06-17	27-06-17		Br	In Perso
6899290	Issue regardin	FeedbackDescripti	Dear	27-06-17	04-07-17		Ge	In Perso
6899292	Automatic rec	FeedbackDescripti	Dear	27-06-17	04-07-17		Ge	In Perso
6899293	Issue with dow	FeedbackDescripti	Dear	27-06-17	04-07-17		Ge	In Perso
6906121	Export node te	FeedbackDescripti	Dear	05-07-17	06-07-17		M.	In Perso

Figure 3.2: Complaint information example (name values are hidden)

The Figure 3.2 provide an example of customer complaints data. There are only a few values visible like the id, some text field, location information and dates.

The first step after receiving a complaint is its documentation. Complaint information includes serial and catalog numbers of the system, the location and the institution that operates it and a description of the problem from the stakeholder that initiated. In order to make sure the initiator has provided all relevant information there is communication in the form of a dialog. After the communication is finished a short description is written. Afterward, a complaint investigator reviews a single complaint and with the assistance of field services engineers (if needed), they identify the cause of the problem. during the investigation, is assigned severity code that categorize complaints by the impact of the outcome of the described situation.

At this point the importance of each complaint influence if it is going to need immediate corrective action or it is going to be placed in the development pipeline or if there is no need for corrective action. This process is done for each complaint separately and there are multiple individuals. Through the process which is currently done it is not possible to identify similar entries and get a clear overview of the scale of impact to the system.

After a complaint is handled there are added more data values in the systems in order to close the investigation. Depending on the chosen action that is needed there are assigned code related to the development and sometimes if it is severe an additional code that relates it with a systems update that has been planned. Additional fields with text are also added that it is described in details the evaluation outcome and the action's details.

Chapter 4

Data Preparation of Customer Complaints

In this chapter, it is going to be explained the method which was used to process the available complaints. Multiple techniques were used in order to achieve results and discover groups of complaints that provide useful insights. The applied techniques helped to identify any barriers that are placed because of the data format and content.

4.1 Guideline

The following section is going to describe briefly the initial steps that are needed to perform in order to follow the approach. This chapter is focusing on the early processing of data and their transformation.

4.1.1 Data Preparation

In every field and business, customer complaints have different structure and data records that can be used. In the process of making a data-driven approach, it is required to make preparation actions to identify which complaint data are going to be used and which not. The goal is get value from complaints data and it is important to select and use information that is initially available before a complaint is handled. All additional information that may exist and record for a complaint should not be taken into consideration but their existence is useful for the later stages of validation if such information is available.

- Identifying and selecting customers' complaints data that are suitable.

After the complaint information is classified, the data should be checked for completeness. Selected values may not be available for every customer complaint. The following step should determine the impact [9] that may be caused if values are not available and complaints that miss value with high impact should be excluded.

- Completeness check over the available data.

The situations that made a customer to file a complaint it is described with a free text value. These data should be included in the selected data and also checked for their completeness. It is

the most important information that is going to be processed further in the approach steps and it is the main focus of the project.

4.1.2 Text Analytics

After selecting the data that are going to be used it is needed to transform them in a format that would make it possible to process them further with the use of intelligent algorithms. The approach is focusing customer complaints that include the text field which describes the topic of the complaint. Usually, text value are processed with the concept of Bag-of-Words [20][23][36]. As explained in the preliminaries chapter following the concept of the Bag-of-Words a vector of values is generated that represents all available words in a text. Applying in a similar way this transformation without any other steps would just provide a different format of the complaints. The approach uses the idea of Bag-of-Words and creates a unique set of words that are going to be used for a specific case would also provide a transformation dedicated to a case. This would eliminate the noise that free text includes because of the complexity of the written language.

- Generate of dedicated Bag-of-Words for the investigated case.

Finding an optimal set of words to use it required an understanding of the domain of the case. Investigating a case of complaints dedicated to a unique product the data for the Bag-of-Words can be found usually in the index of the product's manual. A manual would certainly include terminology that is going to be related to the product and the vocabulary terms are describing its parts, operations and functionality. A further step would include transforming customers complaints in a series of vectors with dedicated vocabulary terms.

- Transform complaints into vectors that represent vocabulary terms.

The following step is going to be the inspection of the transformation results. While transforming would be also efficient to generate statistics for the vocabulary terms. If there are cases of terms that are not being used or recorded only once in the data would be proper to exclude them from the results to reduce the time needs to make further calculations [18].

4.2 Data Preparation

Complains need to be processed in order to use them as input in an algorithm that applies various techniques and output insights for them. Making the first it is needed to extract all relevant complaints of the systems form the database they are stored. It is crucial to make a pre-processing that handles missing information that may affect the outcome of the results [33].

Data preparation is needed for both types of data that are going to be used in the following parts. The first type is the complaints data that are mainly going to be investigated for their completeness according to a description that is going to follow. Secondly, that data will include terms related to the discussed system to be used in the further step of the analytics that is going to be made from the complaints.

4.2.1 Complaints of the System

Taking into consideration the structure of the complaints that contain many different values it is needed to select which are needed and which not. Some of the value is going to be used to perform the research and some others to validate the result. Primary information is going to include text data that have been recorded before the complaints were investigated and secondary information is going to be values that were added after the complaint was handled and its investigation closed. Primary information fields include unique information and are needed to have records for every value. On the other hand, secondary information may not be completed in every field since not all of them are applicable for the solution of a case [17][32].

Primary Information:

- **PR ID:** unique identifier for the complaint.
- **Catalog Number:** version of the system.
- **Serial Number:** increasing number of item version.
- **Short Description:** summary of the content.
- **Customer's Problem Description:** customer initial text.
- **Customer Communication:** discussion with customer for details.

Secondary Information:

- **Date Opened**
- **Location**
- **CR#**
- **Engineering ID Number**
- **FCO#**
- **Severity level**

All values of the primary information need to exist. It is not possible to generate data and complete the missing value with the use of commonly used techniques [32]. The last three are going to include free text values that can not be generated without creating noise in the results. Catalog and Serial numbers combined provide the information for the location and the SystemUID code that is the identifier for the event log data of each system. Taking into consideration these information complaints that lack of these data have to be excluded from the sample and are filtered out.

The complaints refer to systems that are out in the open market. There are many cases of clients who choose only to buy the system without making arrangements for further maintenance. There is also the possibility that a system is sold from a client to a second-hand market. Keeping in mind this information the pre-processing of the data needs to include a step of a validation type that the complaint's referred system is filed and there are available its information. Using the Catalog and Serial numbers to validate their data from the ISDA database of Philips makes possible to overcome these situations. One of the main benefits of this step is going to be the availability of event log and details for the system that is stored in the database which provides plenty of options for analysis.

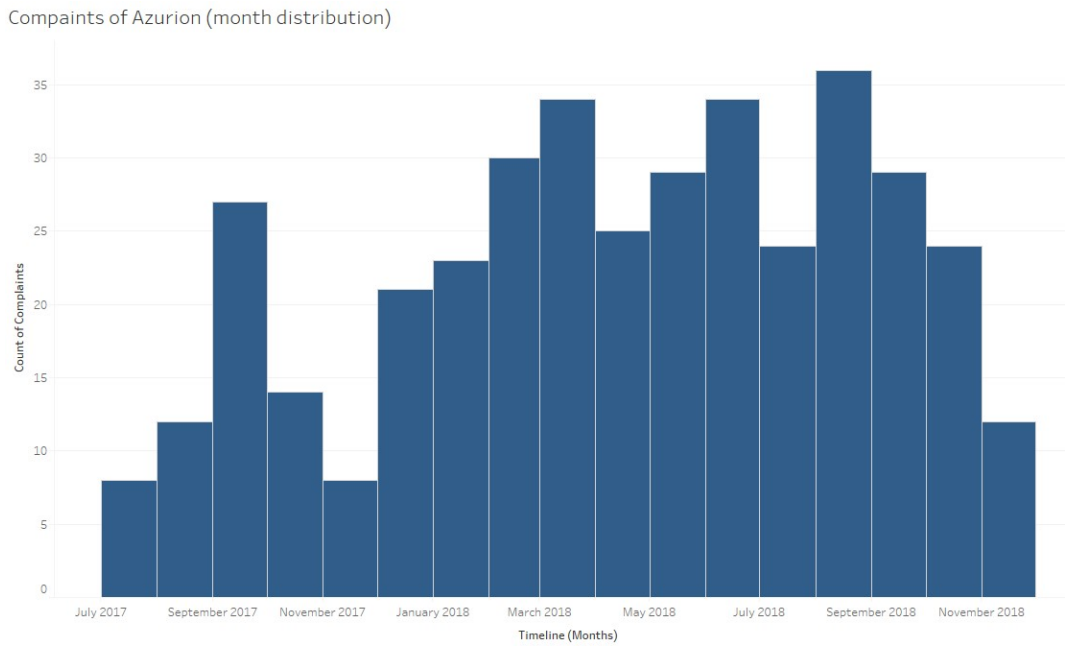

Figure 4.1: Opened complaints each month.

Having a sample for a period that starts 27 June 2017 up to 12 November 2019 that includes 581 complaints the number reached after pre-processing steps is significantly lower. Applying step by step the final sample is going to be 390 complains which is shown in the Figure 4.1 and the following Figure 4.2 shows of complaints per system version.

- Only 67% of the complaints stayed in the sample.

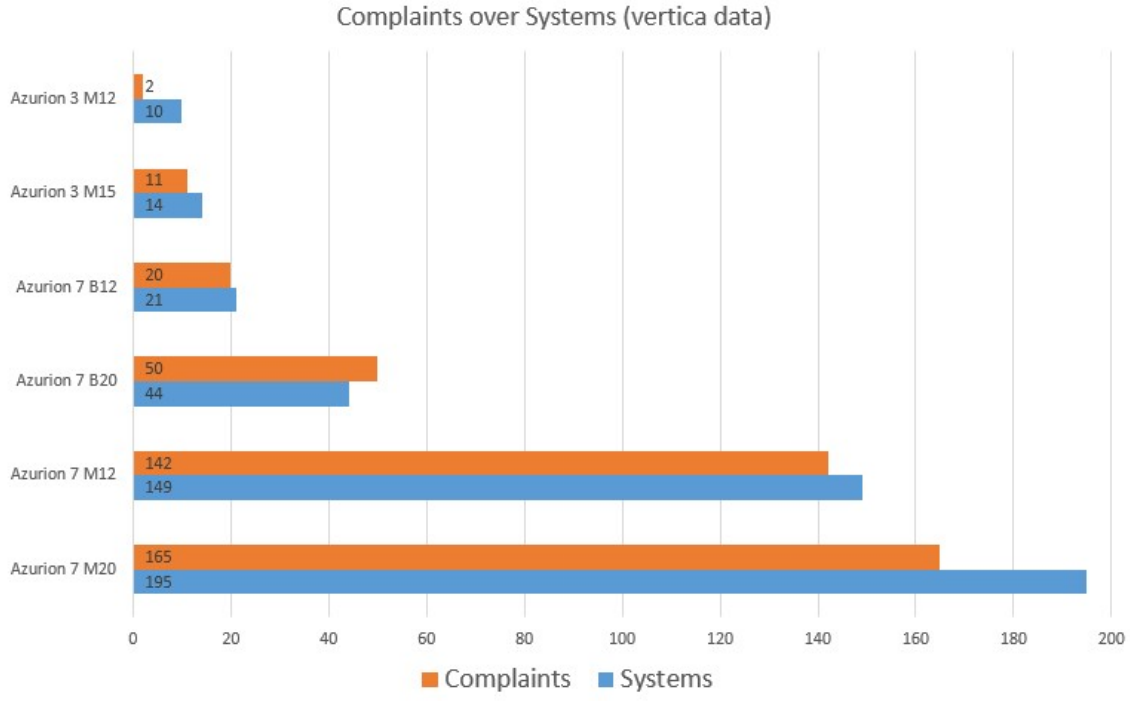

Figure 4.2: Number of complaints for each system type from database.

4.3 Understanding Complaints

Making the first steps to process data from complaints there is a difficulty identified from their content. Initial content contains natural language text that comes with difficulties [7]. To increase the understanding of the data there are three ways.

- Reading its specification.
- Interaction with the product.
- Visiting location its is used.

All three contribute differently and provide concrete information that can help to understand the business and the content of the customer complaints. Reading the specification of the system provides information for its components and its functionality. In that kind of documents, everything is described in a way that even an unfamiliar person with the systems is able to grasp the notions that come with it. Documents like the 'Instruction for Use' [43] describe how to operate the system and how to take advantage of all available features. That information will be beneficial for the interaction with the product.

The next step is to practice and interact with the system. To interact required to spend some hours on it and reproduce the read instructions with the guidance from experienced employees, it is possible to increase the understanding of the expected content of the data. A brief diary from the use tests is in Appendix A. The latter option included visiting the location that a system is used from clients. Visiting a hospital is a different experience; information that is being gained included the way of the mindset of stakeholders that interact with the system. The content of a complaint text is generated subsidiary of their interaction with Azurion. A brief diary from the visits is in Appendix B. In all three ways provided information from a different perspective but there was one element in common and that was the vocabulary used to describe actions or parts of the system.

Using common vocabulary provides an advantage that can be used to cluster the complaints.

Available customer complaints data provide information that includes the description of the complaint, a discussion of the complaints handler and the complaint creator and a short description. These fields have described the cause and the outcome of the complaint. These are three of the six primary information that was introduced in the previous section. Each one contains text in an unstructured format of natural language [26] and has the goal to describe the content of the complaint focusing on different aspects. The first one includes the description from the stakeholder as he understands it, the next one includes a dialog from employee of complaint handling team with the stakeholder that has the goal to retrieve information that are not included in the first field and the latter one is a summary that has to inform briefly for the topic.

In order to define a complaint it can be stated that:

Free text handling is a challenge since every person may use a different way to discuss a topic but in case of a system that it has already named its components and functionalities, these are going to be included with the same vocabulary. Creating a ranking system that each complaint is scanned if it includes the selected terms would result in a first approach for which is the frequency each term is included in a complaint. Also, it would be possible to create groups of complaints that include the same vocabulary terms.

4.3.1 Terms related with the System and usage

Research should Identify the vocabulary terms that should be included in the Bag-of-Words and it is a challenging activity. These should represent perfectly the discussed system and be very accurate. The first attempt was to create after having interaction with the testing systems and with the help of communicating with Philips employees. This would end up to be very time consuming and can be stated it is outside of the scope of the research. The process would also need mastering the domain of the product which is extremely difficult because of the absence of domain knowledge background from the medical field. Next option was to find an already made one that can describe sufficiently the system and use it.

One of the first and most appropriate places to search is the documents related to those systems. These documents have been written and have as a goal to describe a system in a way that is expected from its developing team. It is expected that the context of the complaints related to a system is going to include the terminology that was introduced in one of the related documents. In the current case, the system is going to be Philips's Azurion the most relevant document is going to be the "Instructions for Use" that are provided for it [43].

The document with instruction for the system includes a section with the Index of words and terms defined of it. Mining these words and terms it is going to provide a list that can be used for the purpose to make analytics from the complaints. The final list included 526 terms that are going to be used to generate the analytics and creating a vector for each complaint with the term included in it like in the Figure 4.3.

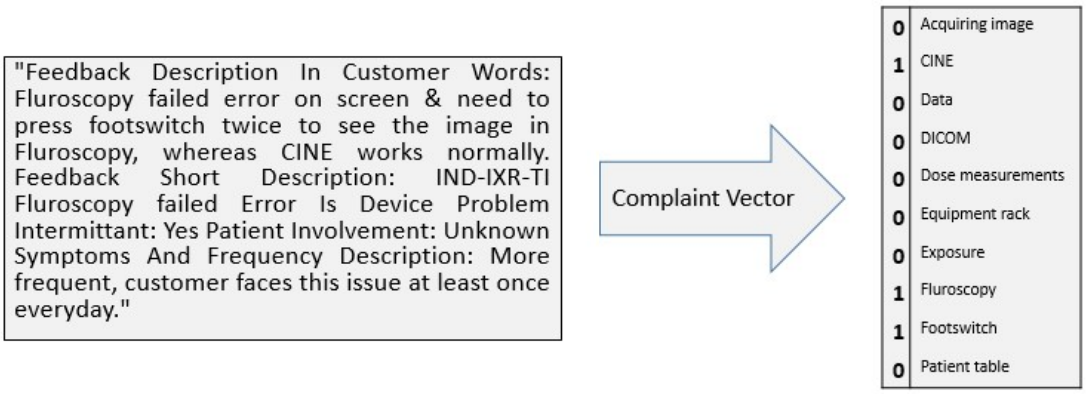

Figure 4.3: Example of complaint vector.

It is achievable by using this code:

```
for i in range(complaints.shape[0]):

    for y in range(terms.shape[0]):

        aField = textCheck(term.loc[y][0],complaints.loc[i][1])
        bField = textCheck(term.loc[y][0],complaints.loc[i][2])
        cField = textCheck(term.loc[y][0],complaints.loc[i][3])

        if (aField or bField or cField):
            results.loc[i][y] = 1
        else:
            results.loc[i][y] = 0
```

Listing 4.1: Code that traverse complaints to find vocabulary terms.

```
def textCheck(a,b):
    """
    a=vocabulary_term b=text_field_value
    """
    if not b is np.nan:
        a = a.lower()
        b = b.lower()
        if a in b:
            return True
        else:
            aSplit = a.split()
            result = True
            for w in aSplit:
                if not(w in b):
                    result = False
                    break
            return result
    else:
        return False
```

Listing 4.2: textCheck().

The Listing 4.1 introduces a double loop structure that searches complaints about terms matches. In the case of Azurion, it checks three different text fields and if in one of them the vocabulary term is present it is marked in the related position in results. The second Listing 4.2 provides the operation of text checking. Since there is a case that one of the text fields may not exist there is an if structure to check if the value is not NaN. Afterward, text and term are converted to lower case and checked of the match. If a vocabulary term is a combination of multiple words it is also split and each word is searched; if all words from the term are included in the text a match is returned from the operation.

4.3.2 Outcome of Text Analytics

The above-suggested notation uses as an input the available 390 complaints and the 526 terms mined from the document [43] there are some interesting results. The following Figure 4.4 is going to present a representation that in horizontal axis are numbers that represent how many complaints had in their text a term. Additionally, the vertical axis is going to have a scale to show the number of terms related to each rate.

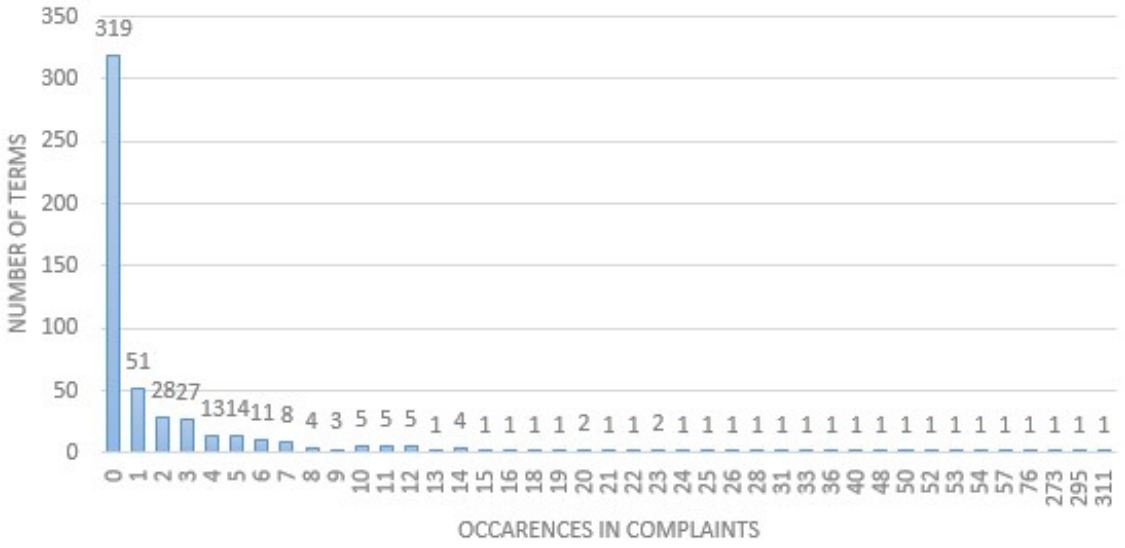

Figure 4.4: Occurrences of vocabulary terms.

The first impression from the results clearly shows that there is a significant amount of vocabulary terms that are not included in any of the complaints. Also, terms that have been included in the text of only one complaint can be considered and added to the amount with zero occurrences since there are going to be as noise in any further action that will result in the creation of complaint clusters. Another notable fact from the graph is the last three number the represent vocabulary terms with a very high frequency of occurrences. These may end up adding noise in the clusters that will be generated after taking into consideration the all three are being present at least in 70% of the complaints.

The goal is to create clusters of complaints vocabulary terms that have been recorded less than two times are either unique and can provide noise to the continuation of the process and it is going to be more efficient to filter them out of the data. On the other hand, the three vocabulary term that has been a record in more than 70% of the customer complaints it can be stated that is being part of the TF-IDF [34] and it will be discussed further in the next section of the document.

Chapter 5

Analysis of Customer Complaints

The following chapter is going to introduce the main part of the approach. The first section is going the describe the suggested clustering step and why some actions were performed in such a way that the results differentiate afterward. The second section is going to describe how the topic of each cluster is going to be introduced with the use of the text analytics and the clustering output.

5.1 Guidelines

The following section is going to show how goals can be reached by using the outcome of the previous steps as an input of various techniques. The focus is going to be to cluster the data and provide additional insights.

5.1.1 Clustering

Clusters of data can be made with many different techniques. The techniques try to create clusters of data by using different approaches. At this point, it is important to remind that the text analytics results represent actual vocabulary terms and not various types of numeric values. Techniques like k-means, mean-shift and density-based clustering provide similar functionality. K-means calculates the shorter distance between data [27], also does step by step calculations with the mean of the data [13] and density-based uses a system that tries to identify the closest data point and not the most similar one [25]. On the other hand, Hierarchical (agglomerative) performs clustering by trying to group those with the most common data values. This is performed until all data points reach to be included in one universal cluster that can be seen in the dendrogram graph. This technique is the one best suited for the operations that are needed to be performed.

- Identify suitable clustering technique.

The next actions after applying hierarchical clustering is generating a suggestion for the number of the clusters in the form of a graph. Is should be investigated if these are going to be meaningful or not. In case there is no meaningful cluster it is going to be needed to make a new iteration of the algorithm until the results are sufficient.

- Apply clustering until there are identified at least some meaningful clusters.

Clusters of complaints can be meaningful and vague at the same time. It is important that these candidate clusters to be validated from the data or from an inspector that is familiar with the domain. Making a validation for the first candidates it is very important since the later steps are going to be based on calculations that may occur because of them. Continuing the steps and by using the following equation it should be identified as a threshold that validated cluster are going to differ significantly from the others.

$$A = \frac{\sum_1^i (t_i)}{n \cdot t} \ , \ i = complaint\ in\ cluster \tag{5.1}$$

The above equation is representing the average of a cluster is going to be the sum of all unique terms in all complaints divided by the multiplication of the complaints in the cluster and the sum of unique terms that exist in the cluster. Having that I represents a complaint in the cluster, n the sum of the complaints in the cluster and t the sum of unique vocabulary terms in the cluster.

- Investigate the threshold values.

Afterward, the threshold and the other restrictions are known it is possible to use a recursive technique and find more candidate clusters. Starting with two clusters and increasing them by in every iteration. Every cluster should be checked and if it is above the threshold should be extracted from the data as a candidate cluster. In the case of an iteration that a candidate is identified the number of clusters should not be increased at that turn. Every time a cluster has only one complaint included this one should be extracted since the algorithm classified as unique since it was placed in a cluster alone.

5.1.2 Graphs

Identified clusters from the steps in the previous subsection and having information known for each of them but not in a format that can show its impact. With the use of data graphs, it is possible to create representation for cluster data. Creating the graph it is going to need some preparation steps. Firstly, It is needed to select the data that are going to be represented in the graph. The focus of the approach is the vocabulary terms and their impact which was used to create a cluster of complaints. Making nodes in the graph that represent all vocabulary terms found in a cluster is going to generate a graph that is going to show insights focus on a cluster. After selecting the nodes it is needed to find how there are connected. Generating a connection matrix from the text analytics results related to the complaint included in the investigated cluster.

- Select data that are going to be represented in the graph.

The previous step provides all data that are going to be included in the graph. Nodes and connection matrix to generate the graph. The generation of the connection matrix was made in order to show the level of connection between each node and not just if there are connected. By creating a weighted data graph [45] it is going to be easier to identify which of the terms are making the major impact on the topic of the cluster and which ones the least. Having available this information it is going to be the choice of the user if he wants to generate an all connected graph or a version that shows only vocabulary terms with high impact.

5.2 Hierarchical Clustering

The following part is going to describe a suggested approach for clustering customer complaints with the use of the machine learning technique of hierarchical clustering and the introduction of a metric that measures the similarity of the complaints in a given cluster. There are introduced more than one iterations of the process that are going to provide different insights that affect the outcome of the process significantly.

5.2.1 Introduction of technique

The procedure of applying the hierarchical clustering implies the need to generate a dendrogram graph to show the possible connections between data. This dendrogram connects data until it has reached a point that they are all connected to a single cluster. Then it colors differently the section that is most appropriate to be as one cluster.

The image in Figure 5.1 provides a dendrogram that at the bottom of it they have represented all data points and the edges that start to connect them represent possible clustering levels. The suggested cut of data from the technique according to the following figure is to create two clusters and separate them into these. Creating only two clusters it is not going to create any value from the data as it is going to be discussed later in this subsection. The data represent customer complaints of a specific system that is considered complex enough to get only two variations of complaints.

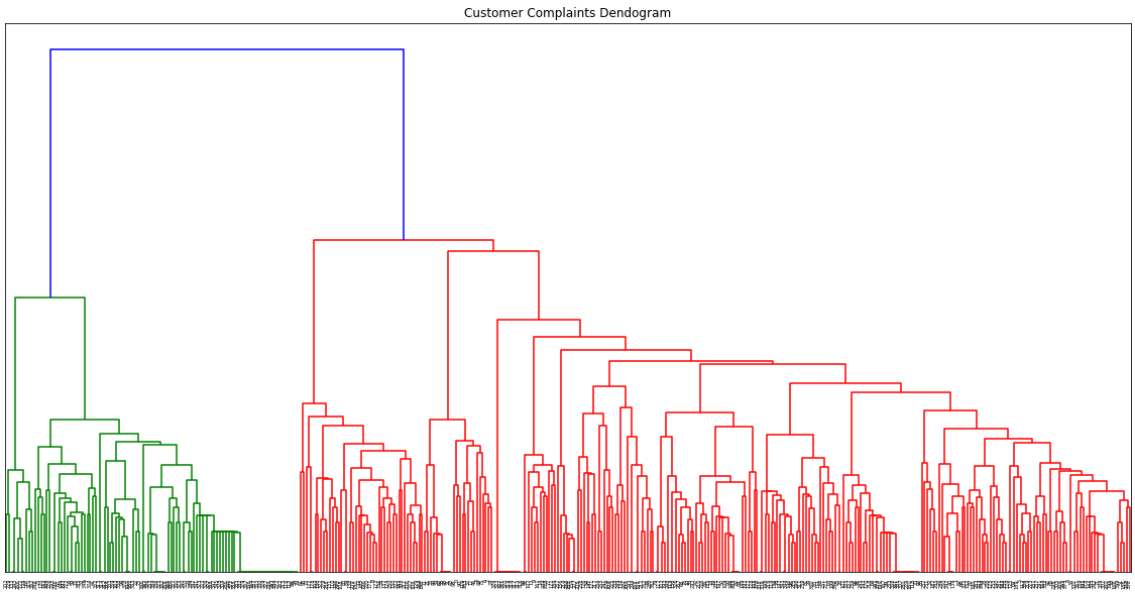

Figure 5.1: Default dendrogram from Hierarchical clustering.

Taking a closer look at the suggestion from the technique it is needed to apply the Agglomerative Clustering operation [39] that gets as an input the data values and the number of the desired clusters and outputs to which cluster each data point is intended to be part of. Following these steps, the next figure is presenting information for the two clusters. The letter **n** provides the number of complaints of the cluster and the **t** the number of vocabulary terms that have been a record at least once in a complaint of the cluster. The chart in the x-axis provides terms that have been records from the complaints and the y-axis is the value of occurrences divided by the

number of complaints in the cluster.

Reading the values from the charts in Figure 5.2 three terms are significantly common in the first cluster. On the contrary, there are no terms that have been identified to be common in the second cluster since the highest in value is present in only 29% of the complaints. Since there are three common terms for the first cluster these are:

Vocabulary Terms frequently common in First Cluster:

- Format 97%
- System 96%
- System Information 93%

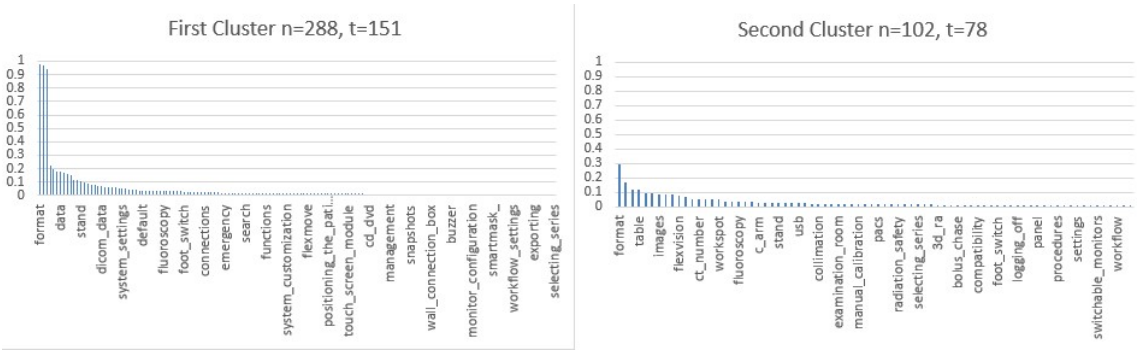

Figure 5.2: Default term distribution in clusters.

Elaborating more over the initial results from the use of the technique's suggestions is that the outcome is not accurate enough for this case. At this point, it is needed to identify how or what is going to be the information in order to create clusters from the use of the technique but by using a different kind of method to identify the possible cluster that may occur from the data. For that reason, it is needed first to find clusters and identify a "common ground" for them.

The following subsection is going to present some an iteration that there were some insightful clusters and provided the solution for the need described above.

5.2.2 First Iteration

In the previous subsection, the need for a way to identify if a cluster ends up to contain common complaints was introduced. Applying the same procedure but choosing to create twelve clusters of complaints the following dendrogram that is introduced can provide it.

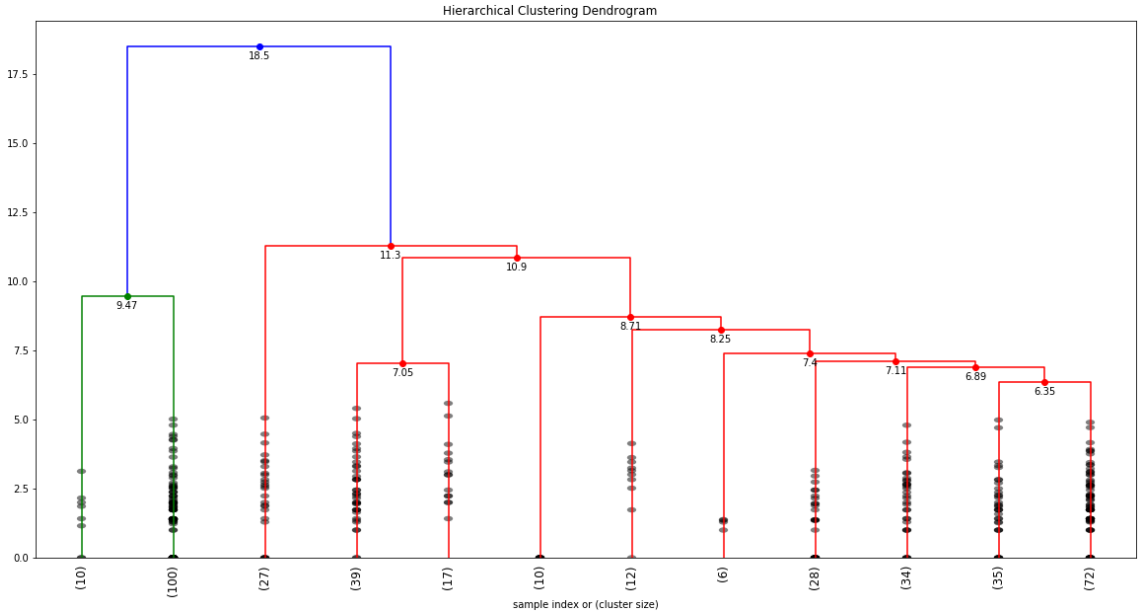

Figure 5.3: Dendrogram for 12 clusters from first iteration.

The Figure 5.3 provides a dendrogram that contains a suggested distribution of complaints about twelve clusters. There are calculated and added in the representation of some additional elements. Firstly, in its connection, it is calculated and note the average range between the combined leaves of it and their scale is on the y-axis. Secondly, in the x-axis, it is noted the number of complaints that are expected to be part of each cluster and the black markings denote possible splits that would occur if the dendrogram was not configured to show only the first twelve separation levels.

The figure below is going to provide four clusters and the rest are going to be included in the Appendix C. Firstly, the 3rd and 9th clusters have the same characteristics as the one that was created when the data was sorted in only two clusters and was discussed earlier. The 8th and 10th clusters are significantly different from others. The number of complaints included is relatively small compare to the others and there are many vocabulary terms in common for all cluster members. This denotes that something is different. At this point it is needed to investigate if the complaints in the groups are related to the same topic and/or their investigation was handled with the same solution. Taking advantage the secondary information (refer Chapter 3 for more information) that are information added in complaints after there are handled it is possible to identify that complaints from each of these cluster share same information and more explicitly in the 8th cluster all complaints are indented to be resolved with the same engineering change (Engineering ID Number is common) and in the 10th cluster all complaints are resolve from the same document number that is included in one of the data values (FCO # common).

Following similar actions the following Figure 5.4 is going to show insights from four clusters after applying the Agglomerative Clustering operation [39]:

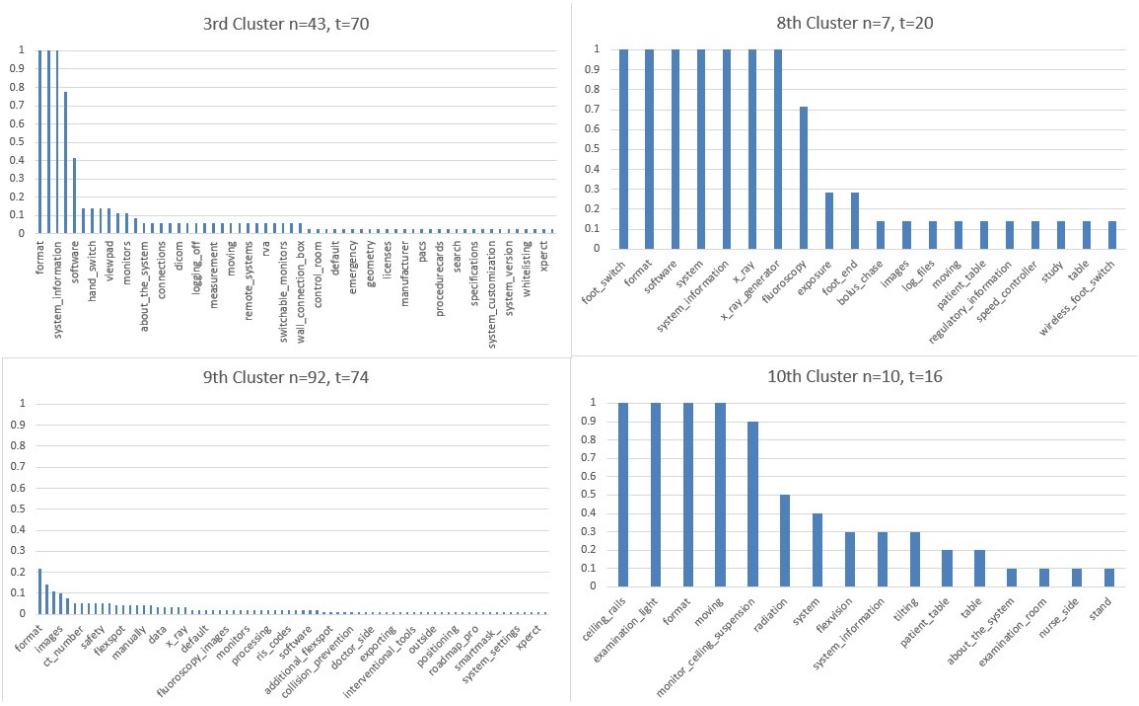

Figure 5.4: Term distribution in 3rd, 8th, 9th and 10th clusters.

At this point, there have been identified clusters that can be also validated from the available data and there are going to be used to identify the reason for being common from the data side of information. Introducing the following definition of a data measure average:

Definition of A (average) value: Average of a cluster is going to be the sum of all unique terms in all complaints divided by the multiplication of the complaints in the cluster and the sum of unique terms that exist in the cluster.

$$A = \frac{\sum_1^i (t_i)}{n \cdot t} \ , \ i = complaint\ in\ cluster \tag{5.2}$$

Cluster Number	n	t	Σ (ti)	A
1	100	103	592	0.057
2	14	40	127	0.227
3	43	70	360	0.120
4	24	48	228	0.198
5	36	60	249	0.115
6	13	46	136	0.227
7	12	44	124	0.235
8	7	20	68	0.486
9	92	74	199	0.029
10	10	16	75	0.469
11	29	56	206	0.127
12	10	7	70	1.000

Table 5.1: Values for each cluster from first iteration.

The Table 5.1 above provides all kinds of results related to the identified clusters. Taking a closer look over the values there are significant differences in the marking of A (average). Going back to the Figure of Dendrogram can be seen that the sizes for the clusters that have been identified differ from the suggested sizes. Cluster 12 has only seven vocabulary terms that are present in all of the complaints in it. Taking a closer look in this cluster the text field is all the same and it can be said that it is not sufficient to take into consideration its result. Clusters 8 and 10 have around the same value of A (average) and as discussed previously it can be validated that their content refers to similar complaints topics and was handled the same way.

The clusters 2, 4, 6 and 7 after reviewing their information and be identified to have on average four common terms and from these the three to be the relative common one that has been identified in the first cluster from that suggested split and the fourth one to be different for each cluster. Clusters 3, 5 and 11 have exactly the same characteristics with the previous ones but contain a bigger amount of complaints that make a drop in the A value. Lastly, clusters 1 and 9 have the same characteristics with clusters 1 and 2 respectively from results that complaints where separated in only two clusters.

5.2.3 Second Iteration

The results of the First Iteration provided some insightful clusters but the majority of the clusters can not provide any information and the strict majority of them are bound with the three vocabulary terms that are present in 70% of the complaints and their formation is based on that terms. At this point, a second iteration of the technique with twelve clusters was applied and the theory of TF-IDF [34] was taken into consideration and the three discussed terms were filtered out from the data.

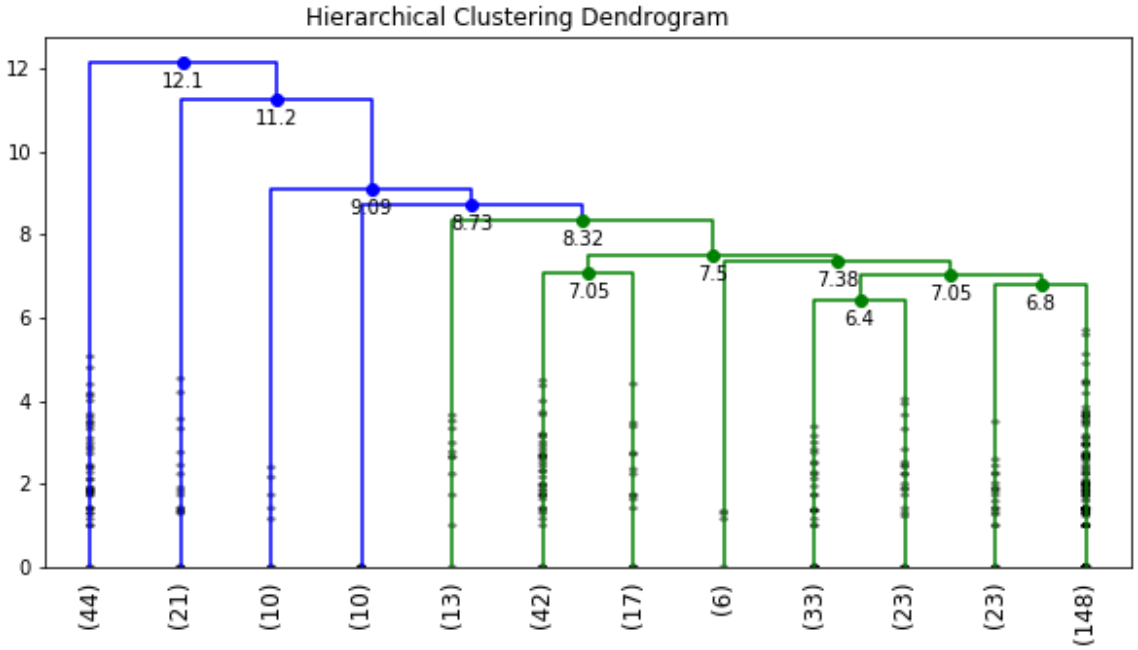

Figure 5.5: Dendrogram for 12 clusters from second iteration.

Second iteration dendrogram in Figure 5.5 has lower values for the distances between clusters and a very closer distribution for high and low values (1st H:18.5 L:6.35, 2nd H:12.1 L:6.4). This information is an indicator of the consistency between the data. Since it is lower the vocabulary term connection between clusters is reduced and more accurate.

Comparing the results of the clustering in Table 5.2 from the second iteration there is a major change that has to be discussed. Cluster 8 is the same with the 8th cluster of the first iteration but it misses not a complaint that should be there since it is validated from the secondary information that it belongs with the rest complaints. Secondly, the values of the A are lower which is caused by the filtering of the three common vocabulary terms and could be predicted from the average lower values of the distances in the dendrogram.

Cluster Number	n	t	Σ (ti)	A
1	148	108	362	0.023
2	44	72	274	0.086
3	17	30	86	0.169
4	23	26	75	0.125
5	23	26	75	0.125
6	33	38	86	0.069
7	21	34	137	0.192
8	6	8	34	0.708
9	10	13	58	0.446
10	10	4	40	1.000
11	23	35	95	0.118
12	13	34	85	0.192

Table 5.2: Values for each cluster from second iteration.

Even though one for the cluster misses one of its complaints the rest of the clusters has better consistency and they have visible topic information because of the frequently common vocabulary terms. The measure of A (average) still has the same characteristics in its behavior and cluster that tend to include related complaints have values above the threshold of 0.40. The following step would be to identify all possible clusters that have A above that threshold and investigate their content.

5.2.4 Repetitive approach for clustering

Since it is unknown how many valuable clusters may exist in the acquired sample of customer complaints it needs to investigate and algorithm that is going to find them. Since it is unknown the number of clusters it is going to be efficient to start with two clusters of complaints and make iterations that each time the number of clusters is increased by one. During that procedure, every identified candidate cluster should be investigated for their content and make a comparison of their A to see if it is above the threshold. If a candidate cluster meets the requirements it is exported and excluded from the data. The next iteration after a cluster is exported should not increase the clusters to avoid missing a possible cluster that was part of a different one in the previous step.

The operation should follow some restrictions in order to avoid the possibility of noise that may exist in data and affect the results. This can be achieved by restricting the process to accept as cluster only if there are at least three complaints ($n \geqslant 3$) included and the vocabulary terms to

be three and above also ($t \geqslant 3$). As a threshold, it is needed to be $A \geqslant 0.40$ which is introduced from the found values in the first and second iteration of the process.

```
while NumComplaints > NumClusters:

    newCluster = False
    #reset of temporary data
    TEMPcomplaints = pd.DataFrame(columns=DATAcomplaints.columns)
    TEMPresults = pd.DataFrame(columns=DATAresults.columns)

    # apply Agglomerative Clustering
    valueCluster = fc.performClustering(DATAresults,NumClusters)

    for i in range(NumClusters):

        #get cluster data according to results

        cc, cr = fc.getCluster(DATAcomplaints, DATAresults, valueCluster, i)
        cw = cr.sum(axis =0, skipna = True)

        if cc.shape[0] < 3:

            UNIcomplaints = UNIcomplaints.append(cc)
            UNIresults = UNIresults.append(cr)

        else:

            #remove 0 from cw (word count)
            cw = cw.to_frame()
            cw.rename(index=str)
            cf = pd.DataFrame(columns=cw.columns)
            for y in range(cw.shape[0]):
                if cw.iloc[y][0] > 0:
                    cf = cf.append(cw.iloc[y])
            cw = cf
            del cf

            #check if there are matches
            if cw.shape[0] < 4:

                UNIcomplaints = UNIcomplaints.append(cc)
                UNIresults = UNIresults.append(cr)

            #calculation of aveCluster
            else:
                temp = 0
                for w in range(cw.shape[0]):
                    temp = temp + cw.iloc[w][0]

                aveCluster = temp / (cw.shape[0] * cc.shape[0])

                if aveCluster > 0.40:

                    newCluster = True
                    NoCluster += 1
                    df = df.append({'number':NoCluster,
                                    'n':cc.shape[0],'t':cw.shape[0],
                                    'A':aveCluster}, ignore_index=True)

                    #export found cluster
                    with pd.ExcelWriter('generated/cluster'+
                                        str(NoCluster)+'.xlsx') as writer:
                        cw.to_excel(writer, sheet_name='cWords')
                        cr.to_excel(writer, sheet_name='cResult')
                        cc.to_excel(writer, sheet_name='cComplaints')

                else:

                    #copy to tempComplaintList
                    TEMPcomplaints = TEMPcomplaints.append(cc)
                    TEMPresults = TEMPresults.append(cr)

    if not(newCluster):

        NumClusters +=1
```

Listing 5.1: Process to identify candidate clusters.

The above part of code in Listing 5.1 present the process followed to identify candidate clusters of customer complaints. The choice of the loop structure was made in order to investigate every possible complaint that may be part of a candidate cluster. Code includes multiple attributes that are going to be explained:

- complaints, results: data of complaints and results from text analytics.
  - DATA: complete set of data.
  - TEMP: temporary data to be used in the next iteration of the loop.
  - UNI: complaints and their text analytics classified as unique.
- cc, cr: complaints and results from a single cluster.
- cw, cf: word count summations.
- aveCluster: the value A

There are also used two functions in Listing 5.2 and in Listing 5.3 that perform specific tasks. The performClustering() that generates the number of the cluster that each complaint is part of by using as input the text analytics results and the desired number of clusters. The getCluster() function makes the operations selects the complaints and results of the selected cluster.

```
def performClustering(a,b):
    """
    a=table_with_results b=Nm_of_Clusters
    """
    cluster = AgglomerativeClustering(n_clusters=b, affinity='euclidean', linkage='ward'
    )
    clusterValue = cluster.fit_predict(a)
    clusterValue = pd.DataFrame(data=NmCluster)
    return clusterValue
```

Listing 5.2: performClustering().

```
def getCluster(a,b,c,d):
    """
    a=complants b=table_with_results c=cluster_values d=selected_Cluster
    """
    cComplaints = pd.DataFrame(columns=a.columns)
    cResults = pd.DataFrame(columns=b.columns)
    for i in range(c.shape[0]):
        if c.loc[i][0] == d:
            cComplaints = cComplaints.append(a.loc[i])
            cResults = cResults.append(b.loc[i])
    return cComplaints, cResults
```

Listing 5.3: getCluster().

5.2.5 Discussion of approach results

The previous section has described a way to identify candidate clusters. Using the customers' complaints from the case of Azurion and their text analytics results there are going to be found forty-two candidate clusters of complaints. These are being presented with their related values in the following Table 5.3 and colored depending if there are actually cluster or not. The process also introduces a set of complaints to be considered unique. In the unique set, are included complaints that did not meet the requirements to be clustered.

Number	n	t	A
1	10	13	0.45
2	10	4	1.00
3	6	8	0.71
4	4	13	0.60
5	4	18	0.47
6	4	7	0.82
7	3	25	0.49
8	5	13	0.58
9	9	8	0.44
10	3	9	0.67
11	4	9	0.67
12	6	20	0.43
13	3	16	0.48
14	3	15	0.49
15	3	10	0.53
16	3	8	0.50
17	4	11	0.41
18	5	9	0.49
19	3	6	0.56
20	5	10	0.48
21	4	5	0.45

Number	n	t	A
22	3	10	0.50
23	3	7	0.62
24	3	8	0.58
25	3	7	0.48
26	4	8	0.44
27	3	9	0.56
28	3	4	0.50
29	5	7	0.46
30	5	7	0.43
31	3	6	0.61
32	3	7	0.52
33	3	4	0.58
34	3	7	0.62
35	5	9	0.51
36	3	5	0.67
37	3	4	0.50
38	4	4	0.44
39	3	4	0.67
40	4	4	0.44
41	3	4	0.50
42	3	4	0.50

Table 5.3: Identified clusters and results.

Cluster marked with the green color where investigated and found data to confirm that the complaints are found to describe the same topic. The first three clusters are the same as the ones found after performing the two attempts to cluster the complaints without taking into consideration the introduced value of A. The candidate clusters are a product from the text analytics results that excluded the vocabulary terms that were present in more than 70% of the data. The third cluster misses one complaint that was placed in a different cluster (the seventh) as it was discussed in second iteration section. Another six clusters that have been found and validated. The candidate cluster with the yellow color has been validated partially. At least half of the complaints can be validated to be related to the same topic. The rest candidate cluster which is twenty-nine after performing an investigation on their content it was not possible to validate if they are correctly placed in the same cluster or it was clear that they are not related to the same topic. Finally, the set complaints that were not included in any candidate cluster had 217 complaints which represent 56% for the available customer complaints from the case of Azurion. Taking into consideration that many of the candidate clusters are not validated as clusters it can be stated:

- 56% of data are identified not to be part of a candidate cluster
- 30% of data are placed in candidate clusters but are validated (yellow and red clusters)
- 14% of data are placed in candidate clusters and validated (green clusters)

5.2.6 Validation information

Performing the validation process included investigation over the secondary information from the complaints as presented in the section of data preparation. This process included many difficulties that have influenced the results a lot. The first challenge was that the available codes which provide the validation were not placed in the same data value of the complaints and sometimes they were mixed with other codes and the investigation had to be performed manually to ensure the quality of the outcome. The second factor the greatly influenced the validation was that only 164 complaints (32% of the data) had available secondary information that could be used for validation.

The 164 complaints were connected to 127 unique code values. Their distribution is:

- 106: one complaint each.
- 13: two complaints each.
- 4: three complaints each.
- 4: four complaints each.
- 1: seven complaints.

After applying the above-discussed approach there have been identified the candidate clusters presented in the following Table 5.4:

Number	n	t	A
1	6	8	0.71
2	4	13	0.60
3	4	7	0.82
4	3	14	0.57
5	3	9	0.67
6	4	12	0.46
7	3	11	0.48
8	3	14	0.62
9	5	18	0.43
10	3	10	0.60
11	3	10	0.50
12	3	8	0.54

Number	n	t	A
13	3	6	0.50
14	5	9	0.47
15	5	4	0.45
16	3	9	0.52
17	3	6	0.50
18	3	6	0.56
19	3	7	0.43
20	3	9	0.41
21	3	4	0.50
22	3	6	0.72
Unique	86	95	0.04

Table 5.4: Candidate clusters found from validation sample.

Having found candidate clusters and the validation code it is time to make a comparison between them. There are only nine codes that include at least three complaints and there are going to be discussed and presented in the following Table 5.5.

Engineering ID Code	Complaints		Major group	amount		Minor Group	amount
CVPRJ00379444	7		Cluster 1	6		Unique	1
CVPRJ00346039	4		Cluster 16	3		Cluster 22	1
CVPRJ00362065	4		Cluster 3	4			
CVPRJ00364461, CVPRJ00381241	4		Cluster 2	4			
CVPRJ00353368	4		Unique	4			
CVPRJ00374126	3		Cluster 17	3			
CVPRJ00284793, ECR-058927	3		Cluster 15	2		Unique	1
CVDEV00025275	3		Cluster 11	2		Cluster 19	1
CVPRJ00381727	3		Unique	2		Cluster 19	1

Table 5.5: Validation results.

Formal validation iteration included a sample of 164 customer complaints. The above engineering code is the only one the occurred at least three times in different complaints. From the clustering process, only the complaint of three codes (green) was placed in a cluster correctly. One code (yellow) had six of the seven complaints placed in the same cluster and the last one was marked as a unique complaint. Another code (red) had all four of its complaints marked as

unique. The rest four codes (white) even though a number of their complaints were placed in a cluster with some irrelevant complaints at least one was placed in a different cluster or marked as unique. Every other identified cluster could not be validated from formal codes.

The code CVPRJ00379444 is related to seven customer complaints. Only six of these ended to be in the same cluster. Taking a closer look at their data from the vocabulary terms there are some interesting facts. In the cluster, there are identified eight unique vocabulary terms and only four of them are present in all six complaints. Similarly, in the missing complaint, are noted eleven vocabulary terms and between them, there are also the four that are common in the cluster. The number and the terms are presented in the following Table 5.6.

Cluster 1:	
exposure	2
fluoroscopy	5
foot_end	2
foot_switch	6
log_files	1
software	6
x_ray	6
x_ray_generator	6

7th Complaint:	
bolus_chase	1
foot_switch	1
images	1
moving	1
patient_table	1
regulatory_information	1
software	1
study	1
table	1
x_ray	1
x_ray_generator	1

Table 5.6: Vocabulary term in Cluster 1 and 7th complaint. (common terms with green)

There are similarities but the algorithm did not place them in the same cluster. This is the fault of the difference in the number of unique terms in the missing complaint. Taking a closer look over the short descriptions of the seven complaints it is clear that the focus of the text is different.

Short Description of complaints in cluster:

- Fluroscopy failed Error.
- No X-ray when footswitch is pressed slowly.
- Exposure Not Possible - error.
- No X-ray when footswitch is pressed slowly.
- Fluoroscopy fail occured pressing foot switch slowely.
- Fluoro/Exp are not possible intermittently.

Short Description of missing complaint:

- Customer states that two injection were given to the patient during a bolus chase study with no images.

The code CVPRJ00353368 is another interesting case that should have been in a cluster and it is not. All four complaints with the code were sent in the set of unique complaints. Inspecting the complaints manually it is found that two of them are the same and most probably it is the same complaint record two times. The short description clearly shows that there is a connection somehow.

Short Description of complaints with code CVPRJ00353368:

- Due to color and texture the mouse will not function on the new white surface of mouse tray.
- Azurion Mouse Tray.
- Wireless mouse not working when used on mouse table.
- Wireless mouse not working when used on mouse table.

Looking at the text analytics results in Table 5.7 of these four complaints there is a different impression. There is only one vocabulary term which is common in all four of them and three that are present in three of the complaints. Taking into consideration that two of the complaints are copies and contain exactly the same terms there is one assumption:

- The chosen vocabulary terms set is not sufficient to cover this case.

accessory_rail	2	mouse_table	3
control_room	2	patient_table	3
examination_room	2	settings	1
interventional_tools	2	system_settings	1
monitors	2	table	3
mouse	4	training	1

Table 5.7: Vocabulary terms in the four complaints that should have been in a cluster.

5.3 Data Graphs

Customer complaints clusters are going to increase the efficiency of the complaints handling process but each cluster can provide additional information that can impact the business plan. The following part is going to present an approach that will introduce the generation of graphs clusters of customer complaints. The representation will be able to provide a dynamic view of the source that triggered the complaints.

5.3.1 Graph preparation

The data graph is going to represent the vocabulary terms from a cluster. These are going to be connected depending on the frequency of their appearances and connected according to the times that have been found in the same complaints. Taking advantage the text analytics that has been already generated and the define function getCluster() from the previous section there is no need of extra actions to get the results for a specific cluster.

The next step would be to make a calculation between the terms and their connections and identify two kinds of data. Firstly, the frequency of the vocabulary terms which can be identified by summing the equivalent column from the results. Secondly, it is needed to calculate the connection values between the term by checking the results table for common occurrences of the terms in a complaint. in order to calculate these the function of graphCalculations() that is included in the next Listing 5.4.

```
def graphCalculations(a):
    """
    a=table_with_results
    """
    connectionMatrix = pd.DataFrame(columns=a.columns,index=a.columns)
    # term frequency
    termsSum = pd.DataFrame(index = a.columns)
    termsSum = PDr.sum(0)

    # connection values
    for i in range(a.shape[1]):
        for y in range(i,a.shape[1]):
            if not(i==y):
                w = 0
                for u in range(a.shape[0]):
                    # common occurrence check
                    if a.iloc[u][i]==1 and a.iloc[u][y]==1:
                        w += 1
                connectionMatrix.iloc[i][y] = w

    return connectionMatrix, termsSum
```

Listing 5.4: graphCalculations().

The connection matrix is going to include values that represent the frequency each term was connected. Because of that only the half upper diagonal part of the matrix is going to have a number that represents the strength of the connections.Taking as an example Cluster 9 from second iteration its matrix is in Table 5.8:

	about_the_system	ceiling_rails	examination_light	examination_room	flexvision	monitor_ceiling_suspension	moving	nurse_side	patient_table	radiation	stand	table	tilting
about_the_system	1	1	0	1	1	1	1	1	1	0	1	1	
ceiling_rails		10	1	3	9	10	1	2	5	1	2	3	
examination_light			1	3	9	10	1	2	5	1	2	3	
examination_room				1	0	1	0	0	0	0	0	0	
flexvision					2	3	1	2	2	0	2	1	
monitor_ceiling_suspension						9	1	2	5	1	2	3	
moving							1	2	5	1	2	3	
nurse_side								1	1	0	1	1	
patient_table									2	0	2	1	
radiation										1	2	1	
stand											0	0	
table												1	
tilting													

Table 5.8: Cluster 9 connection matrix.

Graph creation

The graph is going to include two types of elements. The first element has come to be the nodes and second the edges. The nodes are going to represent all found vocabulary terms of the selected complaint cluster. On the other hand, edges are going to connect the nodes/terms and show their connection frequency. To include edge values in the graph it is needed to change them in a relative format and their value is going to be the width of the edge in the graph. The following Figure 5.6 provides the connection between five different nodes. They are different connection strength

because of the different width size of each edge. The connection from A to B or E is significantly higher than the one between A and D.

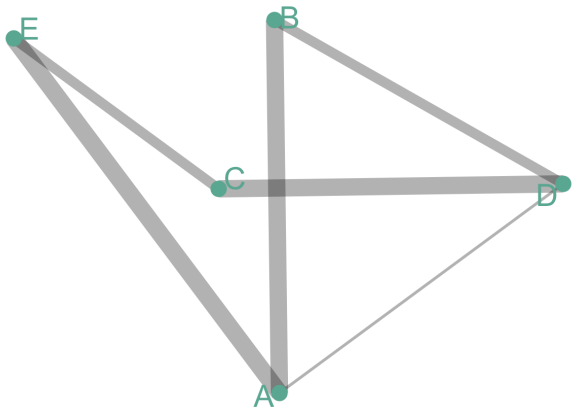

Figure 5.6: Example of weighted matrix [15].

5.3.2 Usability of the graph

After applying all the previous steps the graphs for a cluster can be introduced. The following figures will be based on a cluster that includes ten complaints. It is going to be the discussed cluster 9 from the second iteration. Four variations of graphs are generated in order to show the possibilities of usages of data network graphs.

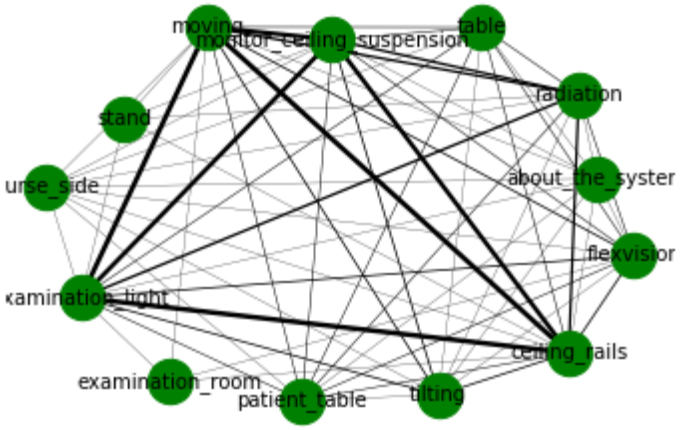

Figure 5.7: Cluster 9 graph with all connected terms.

The first graph in Figure 5.7 provides an all connected description of the topic. Every vocabulary term is connected with different weights. Through this one, it is possible to see the impact on different parts of the system and understand the big picture of the case. The complexity which is introduced can help to understand what made customers create the complaint and from their point of view which parts of the system have different behavior from the expected in case of a malfunction. In case of a cluster that describes an enhancement request the nodes are going to show what is going to be the impact of the feature that is asked.

Graphs made after:

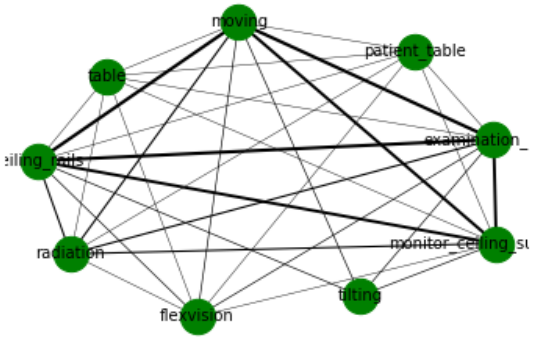

Figure 5.8: Mean = 1.8.

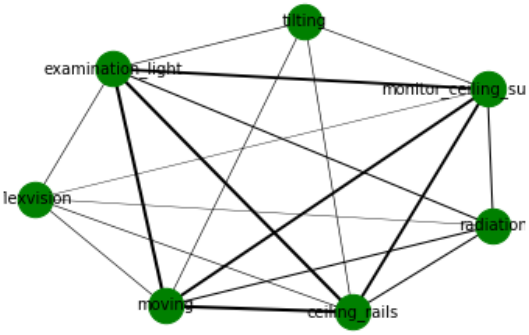

Figure 5.9: Median = 2.0.

The above Figure 5.8 present the generated graph that takes into consideration only connection values above the mean and similarly the Figure 5.9 only those above median. There are more clear graphs that will be readable as the 'common ground' of the case. These are going to be information included in the majority of the complaints and will provide the business an impact that can be used to make analysis in the case and include this information in their development process.

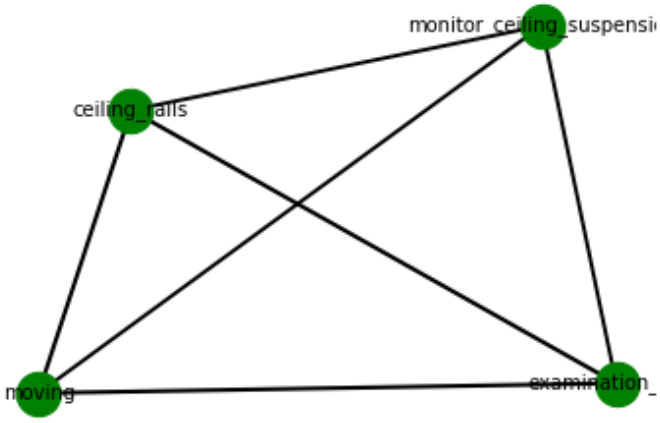

Figure 5.10: Cluster 9 graph with common vocabulary terms.

The final Figure 5.10 present only the common vocabulary terms between all complaints of the cluster. The terms are: 'ceiling rails', 'moving', 'examination light' and 'monitor ceiling suspension'. The usage of the final graph can easily provide a step to get closer to the classification of the complaints. Making a discussion with employees of Philips who know and understand Azurion but they do not know the content of the complaints by seeing this graph and term they could understand the topic that complaints were referring to. The topic is referring to a problem with the movement of the system components that are hanged for the ceiling of the examination room.

Going back to the connection matrix in Table 5.8, the term 'monitor ceiling suspension' is present only in 9 of 10 complaints. Taking a closer look at the content it is possible to find the abbreviation of MCS in all complaints. MCS stands for 'monitor ceiling suspension'. This is an example of an important limitation that is discussed in the conclusions section.

Chapter 6

Discussion

This chapter introduces the approach description. The approach is described in steps that can be applied in other cases of customer complaints handling. The second part presents similar approaches and discusses their differences.

6.1 Approach Description

The project used a case to identify the steps that can help in the process of customers' complaint handling. The identified approach can be applied in different cases by following it and use different complaints data. This section is going to list and describe the approach in a way that the reader will be able to re-apply it.

Figure 6.1: Approach sequence of activities.

The Figure 6.1 provide the sequence of activities that should be followed in order to reach follow the approach.

Data Preparation

It is common that complaint data do not have all the needed information available. It is important to ensure that complaints text data and other useful values are available and take actions to remove incomplete complaints that may not be usable. Analyzing the text information of the complaints; it is not possible to complete missing texts with any of the commonly used technique since it will corrupt the content of a complaint.

Another action that can be included in this activity is the language proofing of the text. The text may contain all kinds of mistakes but the spelling of words is a type of mistake that can significantly disrupt the outcome of the approach. Complaints text should have a spelling check before it is used in the next part of the approach.

Data Understanding

The second activity should focus on the understanding of the content of the complaints. This can be achieved in multiple ways. Firstly, it is going to be efficient to read the available documents that refer to the product that customer complaints are referring to. Secondly, it is going to help if there is interaction with the product and know why, how and where it is used. Lastly, communicating with people who use it and take advantage of the benefits that the product's functionality provides.

The above activities that assist to understand the content are going to assist in the latter activities of the approach. Without the understanding of the domain that complaints are referring to it is not possible to select limits and clustering settings.

Identify Index

The approach required a set of vocabulary terms related to the product. The set should be found by a research that is going to identify what is needed to be included and what is redundant. To perform it is needed to master the domain of product. From the current research, there are no strict suggestions on how to achieve making this set of terms because it was not part of the research scope.

Text Analytics

The activity of text analytics required the set of vocabulary terms from the previous step. This is also the activity that customer complaints will be transformed in a comparable format. Each complaint text is going to be checked for the existence of terms. Terms found in the text are going to be represented in a vector as ones and the others as zeros.

This simple activity will take advantage of the dedicated sets of vocabulary terms to mine information from the content of the complaint. By recording a vector with zeros and ones each complaint will be in a format that can be used from common clustering techniques.

Data Clustering

Data clustering should be performed in a semi-supervised way. The hierarchical clustering is the method that can provide an efficient outcome. Depending on the number of available customer complaints there should perform some iterations of clustering to identify the threshold for the Average value. The equation of Average and its description can be found in 5.2.2 First Iteration.

After the first iterations, the generated trees will provide information for the distances between clusters. At this point, text analytics results should be review in case adjustments are needed. Adjustments are most probably going to be related to the frequency of terms compared to the set of complaints. If some of the terms are present in too many complaints they can be considered as noise and similarly if terms are not used in at least two or three complaints (depending on the minimum size of acceptable cluster) they should also be filtered.

Setting a minimum size of a cluster and the threshold for the average value is important to get results of clustering in the hole set of complaints. An iteration that finds clusters that meet the defined requirements and extracts the complaints from the set. This iteration should have multiple runs until there are no other clusters which meet the requirements.

Data Graphs

After finding clusters of complaints the next and final activity of the approach will provide a way to create a classification of a cluster. Firstly is needed to select the text analytics results of the cluster that is going to be processed. The graph of the cluster needs nodes and edges. Nodes are going to be each vocabulary terms that have been found at least once in the cluster. The edges are going to be calculated and record in a connection matrix. The values of the matrix will represent the number of times that two terms were present in the same complaint. These values are going to be used in order to set the weight/size of the edges. All previous action will resolve an all connected data graph of the cluster.

The mean and the median of the values in the connection matrix should be calculated. These metrics should be placed as a threshold of the nodes and edges and generate different graphs. The filtered graphs will be more clear and are going to provide a view of that was common in the most complaints of the cluster. The final graph that can be generated should include only the terms that are present in all complaints of the cluster. This one will be even more simple from the previous ones and most probably it will give a clear idea to the content in the complaints.

To conclude this section the last two activities provide solutions in the research goals of the project. Clustering in groups the complaints and graph provide the impact of a cluster and insights for the content.

6.2 Related Work

The following section is going to introduce and discuss other approaches for handling customer complaints and why the current approach introduces different functionality. There are already plenty of approaches to complaints handling. There are going to be discussed differences and similarities between them. Each case has positive and negative elements that are affecting the complaints handling process. It also differs between different industries and services field. For each approach that will be discussed, there are going to be added some positive and negative facts concerning the suggested approach.

The approach presented by Mitchell [31] describes a system the successfully work of the sector of the computer services. Its functionality provides an automated system the receives complaints and classifies them. It is generating information for the trends that are being followed for the complaints of their customers. This provides an instant increase of satisfaction to their client that the business care for their problem and is going to handle them. Because it was introduced some year with when the technology evolution was completely different there are limitations on the performance which is it not needed to be taken into account at the current document.

Identified difference:

- **Pro:** More immediate response to events and unforeseen issues.
- **Con:** Need for personal presence because of the semi-supervised nature of the approach.

A completely automated system maybe be beneficial for that kind of business but there are some negative aspects to state. Firstly, the current approach is being introduced and researched over the medical field. In the medical field, a complete automated system may have a negative impact on the customers. Medical devices need in many occasions install handling of the case to overcome a situation. For example, the system Azurion is being used to perform medical procedures on patients. If the hospital technicians are unable to handle a situation they are going to contact the business and file a complaint even in a different format of the ones used in the

research. The current approach provides the possibilities for topic recognition that may assist the domain expert from the content of previous cases. On the other hand, it requires the human aspect to be present that may be more time consuming and have increased the cost for the business.

The ontology-based approach [24] investigates a complaint handling systems based on predefined cased that have been classified by customer complaint ontology. Received complaints are classified hierarchically in one of the existing cases. The classification searches for enough similar content and not only those that are referring to exactly the same situation. It tries to reduce uniqueness overall and each case type includes many sub-cases that its content is related.

Identified difference:

- **Pro:** Universal classification for the customer complaints.
- **Con:** It can not identify unforeseen situations that are not following predefined cases.

A business that develops a product performs many activities before the product reaches the market to ensure that there are handled any known cases that can provoke unwanted behavior of the system. Having a complaints handling process that records and classifies every possible known topic for customer complaints has a great weakness. The identified weakness is the fact that if something is not part of the defined topic can lead to inefficient handling of the case. On the other hand, the presented approach uses for similarity measure the content of the data and not predefined instructions the may not cover a customer complaint.

A similar way of working is presented from the cognitive computing approach [12]. With the use of a machine learning system unstructured data are being classified. To make the classification the initial set of data had been annotated manually to be the training set of the data. It follows the suggested step from the UIMA [11] to complement the sequence of used techniques. It is mentioned that the Bag-of-Words is an important technique that is important to be used.

Identified difference:

- **Pro:** After its establishment it is unsupervised.
- **Con:** It requires a big set of data to exist and annotated manually.

The results of the cognitive approach are very promising but some obstacles cannot be overcome in some cases. It required a significantly big set of data to be available to create machine learning training. In cases of new products, a business does not have available complaint data to follow that approach. Main different from the presented approach is the applicability in every available case even though it would end up to use more resources as the data increase because of its nature to be semi-supervised.

Chapter 7

Conclusions

The project was driven from the need to take advantage of the information that can be found in the customer complaints. A data-driven approach to understand and retrieve information from customer complaints is presented in this thesis. Creating a different format of customer complaints that include only information relevant to the product that can be further processed and used. There are given steps that if there are followed there is an outcome of collective complaints insights and information that could be identified if they have been handled separately.

The major impact for the project was the step followed in order to apply techniques that can reach answering the research goals. Firstly, data were filtered and to overcome cases of missing values since there were text data it was not possible to complete missing values without impacting the content of the data set. Following up action was to find a relevant index with the investigated case. The index was used to transform complaints data into the value of zero and one that represented terms for the index. These actions provided a format of the data that would be processed easier even further.

Research goal 1 was reached by using a recursive approach of hierarchical clustering. During that process, every time the defined metric meets the requirement which has been previously identified a new candidate cluster was created. As a process, it needed to follow after an investigation that would introduce the metric which was used. For research goal 2 all previous actions were taken into consideration and used to provide data graphs that introduce the insights of a cluster. As a process, it needed to create a matrix of connection value from the text analytics related to each cluster.

To conclude the thesis provides an approach for handling customer complaints. The approach uses multiple techniques that process data found in the complaints. It focuses on the use of the domain-related with the product which can be beneficial for the business. The used case during the project implementation was a product of Philips.

7.1 Limitations

There are identified several limitations that can be related to this project. These limitations could not be managed in the short amount of time that the project lasted.

The selected index of vocabulary terms of the case was not made for the presented approach because it was out of the scope of the project; it is possible to state that the content of the complaints shown a lack in the domain of the product. Many of the complaints did not refer to even a single term related to the product and many of the complaints described exactly the same content

in a completely different way. This approach is based on the data from the available text of the customer complaints and in case that the content was described with different terms it could not be used.

Another limitation that was faced is related to the validation of the results. Only a small part of the available data included information that could be used to validate. This reduced a lot the option to investigate accurately for clusters that may exist but there are no data to validate them.

Many customer complaints are trying to describe a situation without knowing it. This approach uses the available text to mine information for the complaint. In case that the creator of the complaint does not fully understand the related product the provided description is also lacking in accurate information and it will influence negatively the results.

7.2 Further development

The work presented in the current document has created a lot of space for further development. In the following line, there are suggested some potential ideas that could be studied in the future.

The approach is using a set of vocabulary terms in order to make the text analytics from the complaints. As additional research, it could be investigated by a usability study what would be the optimal set of terms to be used and create guidelines on how to do this process. In the presented case candidate set of complaints that have not been in a cluster because their topic was not represented in the used index.

A cluster of customer complaints provides information for specific topics related to the product there are referring to. Taking advantage of this information with the use of patterns and anti-patterns it would be possible to identify content in the event log of the products. This would allow investigating other similar cases that have not been recorded in a complaint. Also, it would benefit the business to gain a better understanding of the topic discussed in a complaint cluster.

Taking advantage of the generated graphs as future work could be achieved by finding a new classification approach for customer complaints. This could provide a great impact on the business for the generic content of the customer complaints. Making an automated classification with the use of clustering would reduce dramatically the workload of the complaints handling the process and provide faster input in the development of the product.

Bibliography

[1] Comparing different clustering algorithms on toy datasets. ix, 10

[2] Hierarchical clustering, Feb 2016. ix, 9

[3] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012. 7

[4] Ernst Althaus, Andreas Hildebrandt, and Anna Katharina Hildebrandt. A greedy algorithm for hierarchical complete linkage clustering. In *International Conference on Algorithms for Computational Biology*, pages 25–34. Springer, 2014. 8

[5] Eugene W. Anderson. Customer satisfaction and word of mouth. *Journal of Service Research*, 1(1):5–17, 1998. 5

[6] Howard Anton and C Rorres. Elementary linear algebra. john willey & sons. *Inc, New York, USA*, 1994. 9

[7] Sam Coates-Stephens. The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities*, 26(5-6):441–456, 1992. 19

[8] Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. *SIGKDD explorations*, 4(1):65–75, 2002. 8

[9] Alireza Farhangfar, Lukasz Kurgan, and Jennifer Dy. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12):3692–3705, 2008. 15

[10] R. Feldman and H. Hirsh. Finding associations in collections of text. in machine learning and data mining: method and applications. pages 223–240, 1997. 8

[11] David Ferrucci, Adam Lally, Daniel Gruhl, Edward Epstein, Marshall Schor, J William Murdock, Andy Frenkiel, Eric W Brown, Thomas Hampp, Yurdaer Doganata, et al. Towards an interoperability standard for text and multi-modal analytics. *IBM Res. Rep*, 2006. 44

[12] J Forster and B Entrup. A cognitive computing approach for classification of complaints in the insurance industry. In *IOP Conference Series: Materials Science and Engineering*, volume 261, page 012016. IOP Publishing, 2017. 44

[13] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975. 10, 23

[14] Shreya Ghelani. From word embeddings to pretrained language models-a new age in nlp-part 1, May 2019. ix, 7

[15] Yan Holtz and Conor Healy. Network diagram. ix, 38

[16] Christian Homburg and Andreas Fürst. How organizational complaint handling drives customer loyalty: An analysis of the mechanistic and the organic approach. *Journal of Marketing*, 69(3):95–114, 2005. 5

[17] Rima Houari, Ahcène Bounceur, Abdelkamel Tari, and M Tahar Kecha. Handling missing data problems with sampling methods. pages 99–104, 06 2014. 17

[18] Saam Iranmanesh and Esther Rodriguez-Villegas. A 950 nw analog-based data reduction chip for wearable eeg systems in epilepsy. *IEEE Journal of Solid-State Circuits*, 52(9):2362–2373, 2017. 16

[19] Ashwin Ittoo, Le Minh Nguyen, and Antal van den Bosch. Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry*, 78:96–107, 2016. 7, 8

[20] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science, 1996. 16

[21] Kurt M Joseph, Robert R Bushey, Benjamin A Knott, and John M Martin. System and method for processing complaints, September 18 2007. US Patent 7,272,222. 5

[22] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. 8

[23] Hyunsoo Kim, Peg Howland, and Haesun Park. Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6(Jan):37–53, 2005. 16

[24] Ching-Hung Lee, Yu-Hui Wang, and Amy JC Trappey. Ontology-based reasoning for the intelligent handling of customer complaints. *Computers & Industrial Engineering*, 84:144–155, 2015. 5, 6, 44

[25] Robert F Ling. On the theory and construction of k-clusters. *The computer journal*, 15(4):326–332, 1972. 10, 23

[26] J. Lyons. Natural language and universal grammar. 1991. ISBN 978-0521246965. 19

[27] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967. 10, 23

[28] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010. 7

[29] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999. 6

[30] Stephen K. Markham, Michael Kowolenko, and Timothy L. Michaelis. Unstructured text analytics to support new product development decisions. *Research-Technology Management*, 58(2):30–39, 2015. 8

[31] V-W Mitchell. Handling consumer complaint information: why and how? *Management Decision*, 31(3), 1993. 5, 6, 43

[32] Carol M. Musil, Camille B. Warner, Piyanee Klainin Yobas, and Susan L. Jones. A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, 24(7):815–829, 2002. PMID: 12428897. 17

[33] Souraya Sidani Aurelio José Figueredo Patrick E. McKnight, Katherine M. McKnight. *Missing Data: A Gentle Introduction*. The Guilford Press: A Division of Guildford Publications Inc., 72 Spring Street, New York, NY 10012, 2007. ISBN: 9781593853938. 16

[34] Anand Rajaraman and Jeffrey David Ullman. *Data Mining*, page 1–17. Cambridge University Press, 2011. 7, 22, 29

[35] Heejung Ro and June Wong. Customer opportunistic complaints management: A critical incident approach. *International Journal of Hospitality Management*, 31(2):419–427, 2012. 6

[36] Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. mcgraw-hill, 1983. 16

[37] Paul Semaan. Natural language generation: An overview. *Journal of Computer Science & Research (JCSCR)-ISSN*, pages 50–57, 2012. 7

[38] C. Shearer. A Survey of Sequential Pattern Mining. *The CRISP-DM Model: The New Blueprint for Data Mining*, 5:13–22, 2000. ix, 3

[39] Robin Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973. 25, 27

[40] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):591–606, 2008. 7

[41] Richard A. Spreng, Gilbert D. Harrell, and Robert D. Mackoy. Service recovery: Impact on satisfaction and intentions. *Journal of Services Marketing*, 9(1):15–23, 1995. 5

[42] Gabor J Szekely and Maria L Rizzo. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, 22(2):151–183, 2005. 9

[43] Philips Team. *Azurion Release 1.2 Instructions for Use*. Philips Medical Systems Nederland B.V., Veenpluis 4-6, 5684 PC Best, The Netherlands, 2017. Philips Healthcare 4522 203 52421. 11, 19, 20, 21

[44] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963. 9

[45] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994. 24

[46] Frazer Williams. Consumer response system, information manual. 1992. 6

[47] Mohamed Zairi. Managing customer dissatisfaction through effective complaints management systems. *The TQM Magazine*, 12(5):331–337, 2000. 6

Appendix A

Azurion Use Tests

The current Appendix is going to focus on the demo session that has happened in Philips. Their goal was to get a better understanding of the Azurion. Analyzing the content from the customer complaints about that specific system would also need to be familiar with its functionality and components. There have happened three sessions of interaction with Azurion during the period of the project. Each one has happened to achieve different goals. The first session had focused to introduce the system and its functionality. The second session happened in order to reproduce some usage scenarios. The last session had the goal to reproduce usability tests.

First Session

The main goal of this session was to get familiar with the system. An employee of Philips briefly explained how to interact with the systems and show how to use the available means for interacting with the system. Next part of the session there were introduced many innovative features of the system and replayed as there were going to be used by Philips customers. Last part of the session provided information about the functionality of the system documentation and managing operations that are being used from the clients.

Second Session

Having already from the previous session a good idea for the functionality of the system the second session was focused on using the system to generate event log for specific scenarios. Tasks were performed as like as there are being used in procedures at hospitals or clinics. A brief and summarized log is going to follow to describe better the activities some of the actions that have been performed.

```
10:17   Start procedure
10:18   Move scanner
10:18   Pedal press
10:20   X-ray disabled
10:21   Select operation
10:21   Move scanner
10:22   Move C-arm
10:23   Exposure single-press
10:23   Exposure with both C-arms
10:25   Move patient table
10:26   Fluoroscopy single-press
10:27   Fluoroscopy long-press
10:30   Move wedges
10:31   Move shutters
10:33   Fluoroscopy long-press
10:35   Store last capture
10:36   Reset positioning settings
10:37   Finish procedure
```

Listing A.1: Actions from use test of Azurion.

Third Session

The last session had a completely different character from the others. During the third, there were perform usability tests. These happen every time something new is being prepared to be added in the system. Many tests performed and many different stakeholders interacted with the system and try out its functionality.

Appendix B

Hospital Visits

In this appendix, describes the activities that happened during the visits to a hospital that has and operates multiple Azurion systems. The time passed in the visits was sufficient to understand the opinion of customers who interact with the system and use its functionality.

All visits followed the same routine and schedule during the day as it is followed in the hospital:

- Preparation for entry to the site
- Brief introduction for upcoming surgeries
- Observe surgeries
- Break and discussions
- Brief introduction for upcoming surgeries
- Observe surgeries
- End of the day and cleanliness

There were multiple surgeries performed and had a different goal to achieve. Some of them were lasting only for a short time around 20 to 30 minutes and others could last up to 3 hours. It varied a lot because of the different types of intervention that needed to be performed. In most of the cases, the system has been used regularly and the patient and the doctor were exposed for a lot of time to radiation. In some minor case, the radiation exposure was happening only at the beginning of the surgery and afterward, there were use other systems which in many cases were using Azurion components.

During the discussion time and surgeries (if workload permitted) doctors, nurses and technicians provided their opinion for the system. These are stakeholders that use it daily and need to operate as intended without having any unwanted situations. The discussions are an informal method to file a complaint the may suggest a new enhancement which can improve the functionality of the system from their perspective or indicate behavior that is not expected but the user can not understand it.

In total have happened three visits to the hospital. At the end of each, the understanding of the customers' complaints content was significantly increased. The interaction with a completely different stakeholder of the system who focuses only the elements he is going to use from the product provide crucial information for to way complaints are written.

Appendix C

Term Distribution Results from First Iteration

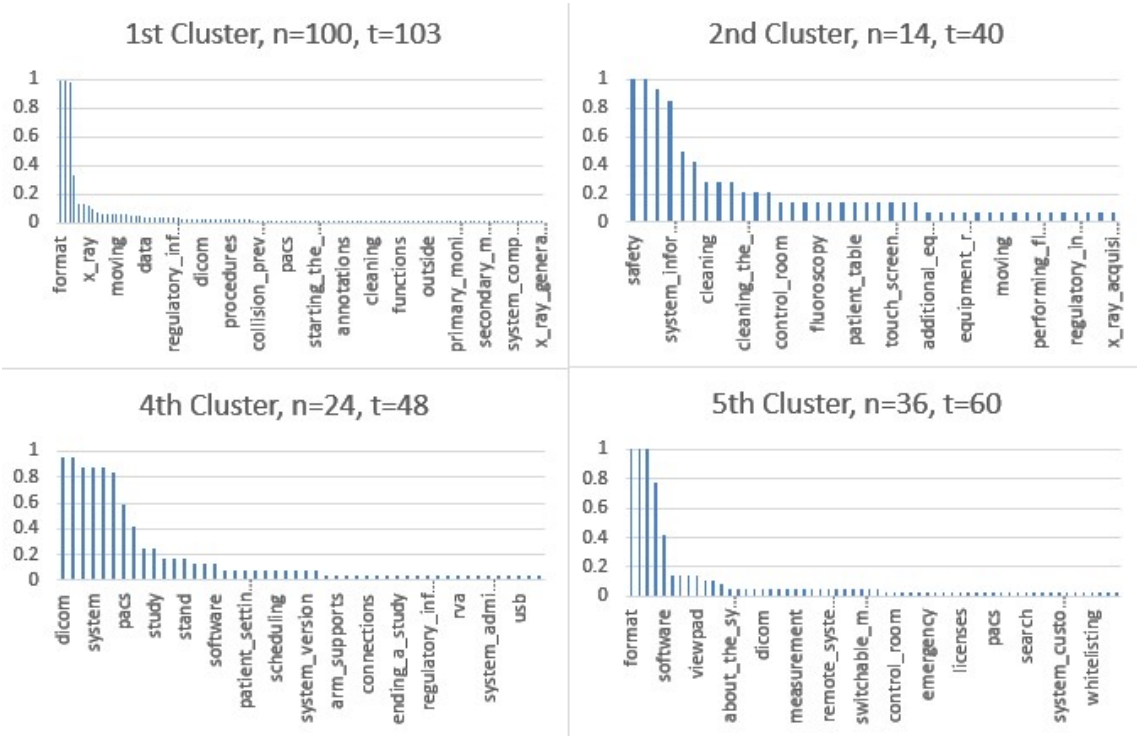

Figure C.1: Term distribution in 1st, 2nd, 4th and 5th clusters.

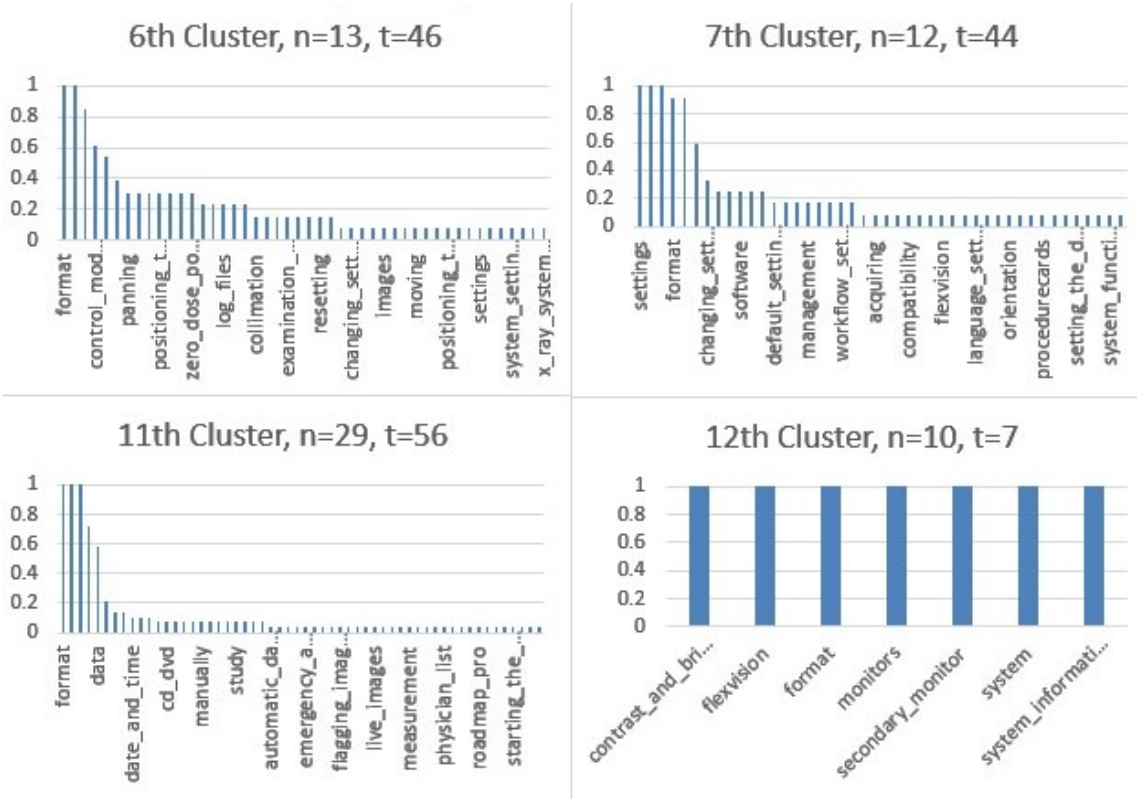

Figure C.2: Term distribution in 6th, 7th, 11th and 12th clusters.