[CHAPTER 2 2](#_Toc530585852)

[LITERATURE REVIEW 2](#_Toc530585853)

[2.0 OVERVIEW 2](#_Toc530585854)

[2.1 DEFINITION OF CONCEPTS 2](#_Toc530585855)

[2.1.1 Data 2](#_Toc530585856)

[2.1.2 Information 2](#_Toc530585857)

[2.1.3 Knowledge 2](#_Toc530585858)

[2.1.4 Differences Between Data, Information And Knowledge 2](#_Toc530585859)

[2.1.5 System 2](#_Toc530585860)

[2.1.6 Information System 2](#_Toc530585861)

[2.1.7 Health Care 2](#_Toc530585862)

[2.1.8 Medical Consultation 2](#_Toc530585863)

[2.1.9 Case Notes 2](#_Toc530585864)

[2.1.10 Health Record 2](#_Toc530585865)

[2.1.11 Electronic Health Record (Ehr) 2](#_Toc530585866)

[2.2 ARTIFICIAL INTELLIGENCE 2](#_Toc530585867)

[2.2.1 Branches of Artificial Intelligence 2](#_Toc530585868)

[2.2.1.1 Neural Networks 2](#_Toc530585869)

[2.2.1.2 Evolutionary Computation 2](#_Toc530585870)

[2.2.1.3 Imaging processing/recognition and computer vision 2](#_Toc530585871)

[2.2.1.4 Robotics 2](#_Toc530585872)

[2.2.1.5 Expert Systems 2](#_Toc530585873)

[2.2.1.6 Machine Learning 2](#_Toc530585874)

[2.2.1.7 Data Mining 2](#_Toc530585875)

[2.2.1.8 Text Mining 2](#_Toc530585876)

[2.2.1.9 NLP 2](#_Toc530585877)

[2.2.1.10 Data Visualization 2](#_Toc530585878)

[2.2.2 The Role Of Artificial Intelligence In Healthcare 2](#_Toc530585879)

[2.3 NATURAL LANGUAGE PROCESSING 2](#_Toc530585880)

[2.3.1 What Is NLP? 2](#_Toc530585881)

[2.3.2 History And Evolution Of NLP 2](#_Toc530585882)

[2.3.3 Why NLP? 2](#_Toc530585883)

[2.3.4 Applications Of NLP Systems 2](#_Toc530585884)

[2.3.5 Implementation Of NLP Systems 2](#_Toc530585885)

[2.3.6 NLP In Health Care 2](#_Toc530585886)

[2.3.6.1 Role Of NLP In Healthcare 2](#_Toc530585887)

[2.3.6.2 Challenges Of Integrating NLP Into Healthcare 2](#_Toc530585888)

[2.4 REVIEW OF RELATED WORKs 2](#_Toc530585889)

[2.4.1 Data-To-Text Summarisation Of Patient Records: Using Computer-Generated Summaries To Access Patient Histories (Scott, Hallett And Fettiplace, 2013) 2](#_Toc530585890)

[2.4.2 Mining Physicians’ Notes For Medical Insights (Hardesty, 2012) 2](#_Toc530585891)

[2.4.3 Automated Identification And Predictive Tools To Help Identify High-Risk Heart Failure Patients: Pilot Evaluation. (Evans, Benuzillo, Horne, Lloyd, Bradshaw, Budge, Rasmusson, Roberts, Buckway, Geer, Garrett, Lappé, 2016) 2](#_Toc530585892)

[2.4.4 Natural Language Processing To Detect Risk Patterns Related To Hospital Acquired Infections (Denys Proux, Marchal, Segond, Kergourlay, Darmoni, Pereira, Gicquel, Metzger, 2010) 2](#_Toc530585893)

[2.4.5. Text Mining And Medicine: Usefulness In Respiratory Diseases (Piedra, Ferrer, Gea, 2014) 2](#_Toc530585894)

# CHAPTER 2

# LITERATURE REVIEW

## 2.0 OVERVIEW

The aim of this project is to develop a Natural Language Processing based visualization system for the improvement of medical consultation. This chapter gives detailed descriptions of the theories and concepts involved with this project. Section 2.1 explains the main concepts involved in this project. Sections 2.2 and 2.3 presents the main research area which is Artificial Intelligence, Natural Language Processing and their application in the healthcare industry. Section 2.4 gives a summary of past works highlighting contributions and limitations.

## 2.1 DEFINITION OF CONCEPTS

This section highlights all the relevant terms that would be used in the course of this project and then explains them to give clarity to their usage.

### 2.1.1 DATA

Data can be defined as a set of characters that has been gathered for some purpose, usually analysis. It can be any character, including text and numbers, pictures, sound, or video. Data has no meaning and it is the lowest and rawest level of knowledge. To obtain data, observations and recordings have to be done. Raw data is used to refer to the information obtained during an experiment, before the information has been analysed or statistically manipulated. If data is not put into context, it doesn't do anything to a human or computer. Data requires interpretation to become information. Once data has been translated to information, it becomes easy for movement and processing. (Christensson, 2006, Hunter, 2007)

Examples of data include: Odunlade, Favour, 50, Female, 67.5kg, 1.52m, 39˚C, 130mg/Dl

In medicine, data includes;

* General numerical information, such as vital signs like heart rate, respiratory rate, and temperature.
* Diagnostic-related information, like laboratory test results from blood tests, genetic tests, culture results, and so on. It can also include imagery like x-rays.

Qualitative data or unstructured data are data that cannot be measured in terms of numbers. These data usually contain information like colour, texture and text. Quantitative data or structured data are data that can be measured easily.

**Structured Data**

[Structured data](https://www.datamation.com/big-data/structured-data.html) usually resides in relational databases (RDBMS). Fields store length-delineated data phone numbers, Social Security numbers, or ZIP codes. Even text strings of variable length like names are contained in records, making it a simple matter to search. Data may be human- or machine-generated as long as the data is created within an RDBMS structure. This format is eminently searchable both with human generated queries and via algorithms using type of data and field names, such as alphabetical or numeric, currency or date. (Taylor, 2018)

**Unstructured Data**

[Unstructured data](https://www.datamation.com/big-data/unstructured-data.html) is essentially everything else. It is more like human language. Unstructured data has internal structure but is not structured via pre-defined data models or schema. It may be textual or non-textual, and human- or machine-generated. It doesn’t fit nicely into relational databases like SQL, and searching it based on the old algorithms ranges from difficult to completely impossible. Examples include emails, text documents (Word docs, PDFs, etc.), social media posts, videos, audio files, and images. (Taylor, 2018)

### 2.1.2 INFORMATION

Information is simply data that has been processed, analysed and classified and is now meaningful to the receiver. Information depends on data and can be seen as data in context, which provides knowledge or insight on a certain matter.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S/N** | **PATIENT NO.** | **LAST NAME** | **FIRST NAME** | **AGE** | **SEX** | **HEIGHT (m)** | **WEIGHT (kg)** | **TEMPERATURE**  **(˚C)** | **BLOOD SUGAR LEVEL (mg/dL)** |
| 1 | 1238 | Odunlade | Favour | 50 | F | 1.52 | 67.5kg | 39 | 130 |
| 2 | 1678 | Chukwuma | Godswill | 10 | M | 1.40 | 45 | 37 | 135 |

Table 2.1 - Table showing patient's basic information

The table above shows data that has been processed and classified. From this, one can infer that Favour is a 50 year old female patient who is 1.52m tall and weighs 67.5kg. She has a temperature of 39˚C and a fasting blood sugar level of 130mg/dL.

### 2.1.3 KNOWLEDGE

Knowledge is simply information that has been put in context. It is the merge of information, experience and insight. Using experience to examine patterns within a given set of information which brings about insight is called Knowledge. So knowledge is broader, richer and deeper than information.

From Table 1, a 50 year old woman with a fasting blood sugar level of 130mg/dL can be diagnosed with Type 2 diabetes. This is because blood sugar levels greater than 126mg/dL are classified as diabetes and being that the woman is already 50 years of age, it can be concluded that she might be suffering from Type 2 diabetes and this affects adults. (Grout, 2014)

Whereas, Chukwuma Godswill who is a 10 year old male patient would be diagnosed with Type 1 diabetes as this one is more common in children.

### 2.1.4 DIFFERENCES BETWEEN DATA, INFORMATION AND KNOWLEDGE

The table below highlights the difference between the three concepts

|  |  |  |  |
| --- | --- | --- | --- |
| S/N | DATA | INFORMATION | KNOWLEDGE |
| 1 | Data has no meaning. It is raw and unprocessed | Processed data with meaning. | Information which is broader, richer and deeper. |
| 2 | Observations and recordings are done to obtain data | Analysis is done to data to obtain information | Uses experience to examine patterns and draw meaning from information |
| 3 | Data is the lowest and rawest level of knowledge | Information is the next level above data. Definition is important to classify data and then transform it to information | Cognitive abilities help to transform data to information as we interpret data that is ever-present. |
| 4 | Odunlade, Favour, 50, Female, 67.5kg, 1.52m, 39˚C, 130mg/dL | Favour is a 50 year old female patient who is 1.52m tall and weighs 67.5kg. She has a temperature of 39˚C and a fasting blood sugar level of 130mg/dL. | A 50 year old woman with a fasting blood sugar level of 130mg/dL can be diagnosed with Type 2 diabetes. This is because blood sugar levels greater than 126mg/dL are classified as diabetes and being that the woman is already 50 years of age, it can be concluded that she might be suffering from Type 2 diabetes and this affects adults. |

Table 2.2- Difference between data, information and knowledge

### 2.1.5 SYSTEM

A system can be defined a group of units that interact and are interdependent on each other, and together, form a coherent body. A computer system is a basic, complete and functional computer, including all the hardware and software required to make it functional for a user. It should have the ability to receive user input, process data, and with the processed data, create information for storage and/or output.

### 2.1.6 INFORMATION SYSTEM

An Information System is a set of components which have been integrated for organizing and analysing data and for providing [information](https://www.britannica.com/science/information-science), knowledge, and digital products. (Vladimir, 2017).

An information system is a system used by business organizations to turn raw data into useful information, handles the flow and maintenance of such information which is then used for decision making within the organisation.

The main components of information systems are computer [hardware](https://www.britannica.com/technology/hardware-computing) and [software](https://www.britannica.com/technology/software), telecommunications, [databases](https://www.britannica.com/technology/database) and data warehouses, human resources, and procedures. The hardware, software, and telecommunications [constitute](https://www.merriam-webster.com/dictionary/constitute) Information [Technology](https://www.britannica.com/technology/technology) (IT), which is now the heart of operations and management of organizations.

Businesses operate more efficiently by using varied information systems to interact with customers and partners, curtail costs and generate revenues.

Types of Information Systems include:

* Transaction Processing Systems
* Management Information Systems
* Decision Support Systems
* Executive Support Systems (Cress, 2009)

1. **Transaction Processing Systems**

In order to perform their daily business operations, every firm needs to process transactions as a transaction refers to any event or activity that affects the organization. Transactions differ from organisation to organisation but there are some transactions which are common to all organisations. They include; placing orders, billing customers, hiring employees, employee record keeping, etc. To support the processing of business transactions, the transaction processing systems (TPS) are used in the organizations

A transaction processing system provides a way to collect, process, store, display modify or cancel transactions. Most of these systems allow multiple transactions to take place simultaneously. The data that this system collects is usually stored in databases which can be used to produce reports such as billing, wages, inventory summaries, manufacturing schedules, or check registers.

Transaction Processing System are operational-level systems at the bottom of the pyramid. They are usually operated directly by shop floor workers or front line staff, which provide the key data required to support the management of operations. This data is usually obtained through the automated or semi-automated tracking of low-level activities and basic transactions. (Cress, 2009 & Vladimir, 2017)

1. **Management Information Systems**

A management information system is a management-level system that condenses and converts the data collected by the transaction processing system to create reports in a way that managers can use it to monitor the performance of an organization and then make routine business decisions in response to problems. Some of the reports that this information system creates are summary, exception and ad hoc reports. All this is done to increase the efficiency of managerial activity.Example of an MIS output is the budget report.

Usually, management information systems are used to produce reports on monthly, quarterly, or yearly basis. However, if managers want to view the daily or hourly data, MIS enables them to do so. In addition, they provide managers online access to the current performance as well as past records of the organization. (Cress, 2009 & Vladimir, 2017)

1. **Decision Support Systems**

A Decision Support System is a knowledge based system, used at the management level to facilitate the creation of knowledge and allow its integration into the organization. These systems have more analytical power as compared to other information systems and as a result, are often used to analyse existing structured information and allow managers to project the potential effects of their decisions into the future.

They employ a wide variety of decision models to analyse data or summarize vast amount of data into a form (usually form of tables or charts) that make the comparison and analysis of data easier for managers.

Decision support systems provide interactive environment so that the users could work with them directly, add or change data as per their requirements, and ask new questions. They may support the exchange of information within the organization. A decision support system in a bank, for example, enable a manager to analyse the changing trends in deposits and loans in order to ascertain the yearly targets. (Cress, 2009 & Vladimir, 2017)

1. **Executive Support System**

Executive Support Systems (ESS) are strategic level information systems that provide top-level executive of a readily accessible, interactive format to get the overview of the entire organizations performance. They help executives and senior managers analyse the environment in which the organization operates, to identify long-term trends, and to plan appropriate courses of action. The decisions taken with the help of executive support system are non-routine decisions that affect the entire organization. Executive Information System are designed to be operated directly by executives without the need for intermediaries and easily tailored to the preferences of the individual using them.

As compared to DSSs, ESSs offer more general computing capabilities, better telecommunications and efficient display options. They use the advanced graphics software to display the critical information in the form of charts or graphs that help senior executives to solve a wide range of problems. To make effective decisions, they use summarized internal data from MIS and DSS as well as data from external sources about events like new tax laws, new competitors, etc. They filter, compress, and track data of high importance and make it available to the strategic-level managers.

Executive support systems help to monitor performance, track activities of competitors, identify opportunities, and forecast trends. They also assist senior managers in answering the following question:

* What business should we do?
* How are our competitors doing the business?
* Which units can be sold and which new units are to be bought?

### 2.1.7 health care

Health care describes all the efforts made to maintain or improve physical, mental and emotional well-being via the prevention, diagnosis, and treatment of disease, illness, injury, and other physical and mental impairments in human beings. This may be done with surgery, the administering of medicine, or other alterations to a person’s lifestyle. Healthcare is delivered by health professionals (providers or practitioners) in allied health fields.

Primary Health Care (PHC) is a conceptual model which refers to both processes and beliefs about the ways in which health care is structured. PHC encompasses primary care, disease prevention, health promotion, population health, and community development within a holistic framework, with the aim of providing essential community-focused health care (Thomas-MacLean, Tarlier, Ackroyd-Stolarz, Fortin, and Stewart, 2008).

According to (Toon, 1994), primary healthcare has three elements:

1. Curing diseases or alleviating suffering by biomedical intervention
2. Preventing diseases by biomedical intervention
3. Helping people understand and make sense of their illnesses – the interpretative function.

People use health care services for many reasons: to cure illnesses and health conditions, to mend breaks and tears, to prevent or delay future health care problems, to reduce pain and increase quality of life, and sometimes merely to obtain information about their health status and prognosis. (Bernstein, Hing, Moss, Allen, Siller and Tiggle, 2003)

### 2.1.8 MEDICAL CONSULTATION

The medical consultation is best understood as a two-way social interaction wherein the doctor reviews a patient's medical history, elicits information from the patient, examines the patient and makes recommendations as to care and treatment. The patient can choose what information to disclose and how to present it, can ask questions that influence the doctor’s perception of the problem, can make explicit requests and, above all, can choose how to respond to the advice offered or the treatment prescribed. The outcome of the consultation is affected by the actions and choices of both participants.

The best definition of the consultation was given by Sir James Spence Professor of paediatrics, Newcastleupon-Tyne in 1960. He said “The essential unit of medical practice is the occasion when, in the intimacy of the consulting room or sick room, a person who is ill, or believes himself to be ill, seeks the advice of a doctor whom he trusts. This is a consultation and all else in the practice of medicine derives from it.” (Agius, 2014)

Another kind of consultation organises around the development of a complex shared understanding: the creation of new knowledge for both participants, unique to that patient; doctor and patient exploring together what it means for that person individually in their life to have a particular illness such as diabetes, and to live with it. (Fischer and Ereaut, 2012)

### 2.1.9 CASE NOTES

Case notes are a record about a patient’s treatment. They may be electronic or paper-based and, by law, must be kept secure and private. Case notes allow for the documentation of what treatment has been given to the patient and the response of the patient to the treatment which provides accountability for the clinicians.

Case notes have many benefits for clinical work:

* They provide a record of treatment goals and can be used to monitor progress
* They provide a space to reflect on treatment and outcomes.

### 2.1.10 HEALTH RECORD

A health record is a comprehensive collection of a patient’s health information (like dietary habits, fitness, physical abilities/status, hormonal balance, age, weight, blood pressure, etc.) and medical records, those files that physicians keep and update whenever patients visit them for a check-up, or go into the emergency room.

### 2.1.11 ELECTRONIC HEALTH RECORD (EHR)

According to WHO in 2006, The Electronic Health Record is a systemized collection of all personal health information belonging to an individual. The information is entered and accessed electronically by healthcare providers over the person’s lifetime; and it extends beyond acute inpatient situations including all ambulatory care settings at which the patient receives care. Ideally it should reflect the entire health history of an individual across his or her lifetime including data from multiple providers from a variety of healthcare settings.

The EHR enables clinicians treating people in a variety of settings to exchange and continuously update a patient's clinical data and then present that information in logical clinical groupings that clinicians in other health organizations can access easily - such as laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and school and workplace clinics

The key functions of an electronic health record system outlined by an Institute of Medicine committee are health information and data storage, results management, order entry and management, decision support, electronic communication connectivity, patient support, administrative processing, and reporting and population health management (Institute of Medicine, 2003).

Such an integrated system allows a physician to enter a patient's record number and view a menu showing his or her current medications, problem list, history of recent visits to health providers with submenus for notes from those visits, images and reports of diagnostic procedures, a functional status assessment and social service eligibility report, schedule of preventive services, allergies, contact information for all persons caring for the patient, names and contact information for family caregivers, guidelines for appropriate care, and clinical decision support. (Burton, Anderson, and Kues, 2004).

EHR systems are designed to store data accurately and to capture the state of a patient across time. It eliminates the need to track down a patient's previous paper medical records and assists in ensuring data is accurate and legible. It can reduce risk of data replication as there is only one modifiable file, which means the file is more likely up to date, and decreases risk of lost paperwork. Due to the digital information being searchable and in a single file, EMRs are more effective when extracting medical data for the examination of possible trends and long term changes in a patient.

## 2.2 ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is the ability of a digital [computer](https://www.britannica.com/technology/computer) or computer-controlled [robot](https://www.britannica.com/technology/robot-technology) or machine or system to demonstrate intelligence by performing tasks such as finding solutions to complex problems, in a more human-like fashion. It is the extension of human intelligence using computers by developing systems endowed with the [intellectual](https://www.merriam-webster.com/dictionary/intellectual) processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience. Some systems have attained the performance levels of human experts and professionals in performing certain specific tasks, so that artificial intelligence is then found in applications as [diverse](https://www.merriam-webster.com/dictionary/diverse) as medical [diagnosis](https://www.merriam-webster.com/dictionary/diagnosis), game technology, computer [search engines](https://www.britannica.com/technology/search-engine), and voice or handwriting recognition. (Copeland, 2018, Kok, Boers, Kosters, Van der Putten, & Poel 2009)

### 2.2.1 Branches of Artificial Intelligence

The different branches of Artificial Intelligence include:

1. Neural Networks – e.g. brain modelling, time series prediction, classification
2. Evolutionary Computation – e.g. genetic algorithms, genetic programming
3. Imaging processing/recognition and computer vision – e.g. object recognition, image understanding
4. Robotics – e.g. intelligent control, autonomous exploration
5. Expert Systems – e.g. decision support systems, teaching systems
6. Speech Processing– e.g. speech recognition and production
7. Natural Language Processing – e.g. machine translation
8. Planning – e.g. scheduling, game playing
9. Machine Learning – e.g. decision tree learning, version space learning

#### 2.2.1.1 Neural Networks

Neural networks are brain-inspired systems which are intended to replicate the way that we humans learn. The basic idea behind a neural network is to simulate (copy in a simplified but reasonably faithful way) lots of densely interconnected brain cells inside a computer so you can get it to learn things, recognize patterns, and make decisions in a humanlike way. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize. In other words, it is not programmed to learn explicitly, it learns all by itself, just like a brain. (Woodford, 2018, Dormehl, 2018)

#### 2.2.1.2 Evolutionary Computation

Evolutionary computation is a subfield of artificial intelligence which solves problems using techniques inspired by biology. Its principles are based on the theory of biological evolution, such as genetic inheritance and natural selection, which are used to create optimization procedures or methodologies, usually implemented on computers, to solve problems. (Yao, 1999)

#### 2.2.1.3 Imaging processing/recognition and computer vision

Image processing is simply a systematic transformation of an input image into an output image. In between the processes involved, information is usually extracted from the image to assist in the transformation. The tasks performed during the image processing prepares it for Computer Vision. Basic image processing includes rotation, colour scale changes, crop, filter effects, smoothing, sharpening, contrasting, stretching etc.

Computer Vision tries to do what a human brain does with the retinal input, it includes understandingand predicting the visual input. Computer vision involves studying an image or a group of images, use image processing and machine learning techniques, to mine information from image other than its properties. E.g. for computer vision can be differentiating the chairs from the tables in the image of a classroom. A Computer Vision algorithm can then be taught to decipher meaning from images as a result of existing information it was fed earlier. Like the human brain can recognise someone by their face, similarly computer vision algorithms can be taught to recognise patterns, distinguish between objects etc. (Wu, 2003)

#### 2.2.1.4 Robotics

Robotics is a branch of technology that concerns itself strictly with the different branches and applications of robots. A robot is a programmable machine that carries out a set of tasks autonomously in some way. They are not computers nor are they strictly artificially intelligent.

Artificial intelligence comes into play when these robots no longer just carry out repetitive tasks, they actually start thinking on their own and performing tasks based on results from previous tasks or information. Overtime, these robots tend to perfect what they were made for.

Artificial Intelligent robots that are controlled by inbuilt AI programs. They need this in order to perform increasingly more complex tasks. (Owen-Hill, 2017, Patidar 2018)

#### 2.2.1.5 Expert Systems

Expert systems are systems which are capable of making reasoned judgements and offering solutions to specific problems in a given domain or which are able to give advice, both in a way and at a level comparable to that of human experts in the field (Lucas, & Van Der Gaag, 1991)

#### 2.2.1.6 Machine Learning

Machine learning (ML) is a category of algorithm that uses statistical techniques to give software applications the ability to "learn" with data and to become more accurate in predicting outcomes over time, without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. (Burns, 2016)

The processes involved in machine learning involves searching through data to look for patterns and adjusting program actions accordingly. Personalized marketing is one application of machine learning as recommendation engines use machine learning to send personalized ads to their customers. Other common machine learning use cases include fraud detection, spam filtering, network security threat detection, predictive maintenance and building news feeds.

#### 2.2.1.7 **Data Mining**

The enormous amount of data stored in files, databases, the Web and other information repositories gives rise to the need to develop a means to analyse and interpret such data, in order to extract interesting knowledge that can therefore be used in decision-making. (Chamatka & Butey, 2014)

Data Mining can therefore be defined as the process of discovering patterns and relationships amongst large amounts of data, from the different data sources. (Han, Pei and Kamber, 2011)

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Data mining tools can predict behaviours and future trends, allowing businesses to make positive, knowledge based decisions such as cut costs or increase revenue.

Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, etc.

**Who uses Data Mining?**

Data mining is primarily used by industries that cater to the consumer, like retail, financial and marketing companies. Online retail stores use data mining to analyse customers’ purchase history to find out what products they have been buying and what promotions they might be interested in. Netflix uses data mining to recommend movies to its customers, Google uses mining to tailor advertisements to internet users and Walmart uses data mining to manage inventory and identify areas where new products are likely to be successful. Mining is more likely to be used by larger companies, as enormous computers are required to sift through data. (Andale, 2015)

In healthcare, data mining has been effective in a number of areas such as predictive medicine, customer relationship management, detection of fraud and abuse, management of healthcare and measuring the effectiveness of certain treatments.

#### 2.2.1.8 Text Mining

Text Mining or Text Analytics is an extension of data mining. It tries to find textual patterns from large non-structured sources, as opposed to data stored in relational databases. Text Analytics, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting non-trivial information (hidden relationships and trends) and knowledge from unstructured text (Bonani, 2018). Text Analytics is similar to data mining, except that data mining tools are designed to handle structured data from databases, either stored as such or as a result from pre-processing unstructured data. Text Analytics can cover unstructured or semi-structured data sets such as emails, full-text documents and HTML files, blogs, newspaper articles, academic papers, etc. Text Analytics is an interdisciplinary field which draws on information extraction, data mining, machine learning, statistics and computational linguistics. (Moreno & Redondo, 2016)

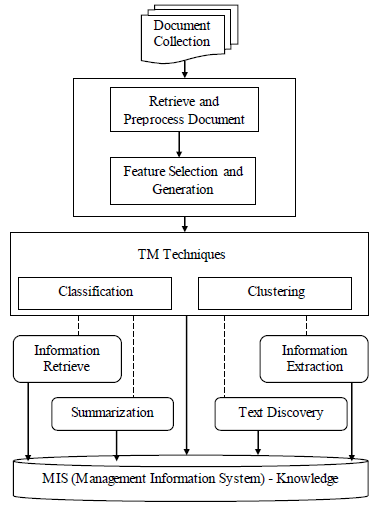


Figure 2.1 – Architecture or flow of the text mining process

##### 2.2.1.8.1 Advantages Of Text Mining

Text Mining has a number of advantages, some are highlighted below:

1. Text analytics helps to extract concepts from text and present it in a more simple way
2. Text Mining helps to summarize the documents
3. It saves time and resources and performs efficiently than human brains.
4. You can plug in any vocabularies to use the terminology in your area of interest
5. The text which is indexed using Text mining can be used in predictive analytics
6. It helps to track opinions over time

##### 2.2.1.8.2 Applications Of Text Mining

Text Mining can be applied in all industries and for various purposes. Some of them are highlighted below:

1. **Analysing open ended survey responses:** To know more about the customers’ opinions, open ended questions are used to allow respondents give their opinions without constraints, rather than relying on structured questionnaires. Text mining can be used to analyse such information in the form of text.
2. **Automatic processing of messages, emails:** Text Mining is also mainly used to classify the text and filter the unnecessary mail using certain words or phrases. Such mails will automatically discard such mails to spam. Text Mining will also send an alert to the email user to remove the mails with such offending words or content.
3. **Analysing warranty or insurance claims:** In most of the business organizations information is collected mainly in the form of text. For example in a hospital the patient interviews can be narrated briefly in text form and the reports are also in the form of text. These notes are nowadays collected electronically so that it can be easily transferred into text mining algorithms. These records can then be used to diagnose the actual situation.
4. **Investigating competitors by crawling their web sites:** Text Mining is also used in processing the contents of web pages in a particular domain. Through this way the text mining system will automatically find a list of terms which is used in the site, and from there one can find out the most important terms used in the website. By this way one can know the capabilities about the competitors which can help you to deliver business efficiently.

The other applications of Text Mining includes the following

1. Text summarization
2. Business Intelligence
3. Social Media Monitoring
4. National Security or Intelligence works
5. Records Management
6. Bioinformatics

##### 2.2.1.8.3 **Techniques in Text Mining**

There are five basic techniques used in Text Mining system (Moreno & Redondo, 2016). They are discussed in detail below

1. **Information Extraction**

Information Extraction (IE) is a technique that extract meaningful information from large amount of text. This is used to analyse the unstructured text by finding out the key phrases and the relationships between them. (Talib, Hanif, Ayesha, & Fatima, 2016). In this technique the process of pattern matching is used to find out predefined sequences in text. It helps in transforming the unstructured text into structured form. The Information extraction technique involves language processing modules. This is mostly used where there is large amount of data.

1. **Categorization**

Categorization involves identifying the main themes of a document by placing the text document under one or more category. Categorization often relies on relationships identified by looking for broad terms, narrower terms, synonyms, and related terms. The categorization process includes pre-processing, indexing, dimensional reduction and classification. The text can be categorized using techniques like Naive Bayesian classifier, Decision tree, Nearest Neighbour classifier and Support Vector Machines. (Gaikwad et al, 2014)

1. **Clustering**

Clustering method is used to group text documents which has similar contents (Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez, & Kochut, 2017), but it differs from categorization in that documents are clustered without the use of predefined topics. In other words, while categorization implies supervised (machine) learning in the sense that previous knowledge is used to assign a given document to a given category, clustering is unsupervised learning: there are no previously defined topics or categories. It has partitions called clusters and each partition will have a number of documents with similar contents. Clustering makes sure that no document will be omitted from the search and it derives all the documents which has similar contents. K-means is the frequently used clustering technique. This technique also compares each cluster and finds how well the document are connected to each other. Companies use this technique to create a database with thousands of similar documents.

1. **Visualization**

Visualization technique is used to simplify the process of finding relevant information. This technique uses text flags to represent documents or group of documents and uses colours to indicate the compactness. Visualization technique helps to display textual information in a more attractive way. (Gaikwad et al, 2014)

1. **Summarization**

The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. It makes the document work reading for the users and understand the content at a glance. Summarization replaces the entire set of documents as it helps to highlight major points in a document. (Gaikwad et al, 2014) Pre-processing and processing operations are performed on the raw text for summarization, to obtain representation of the original text. Tokenization, stop word removal, and stemming methods are applied for pre-processing. Lexicon lists are generated at processing stage of text summarization. (Talib et al., 2016)

##### 2.2.1.8.5 Methods And Models Used In Text Mining

Based on the information retrieval technique, text mining has four main methods (Gaikwad, Chaugule, & Patil, 2014)

1. **Term Based Method (TBM)**

Term in a document means a word which has semantic meaning. In this method the entire set of documents is analysed on the basis of term. The advantages of this method include efficient computational performance as well as mature theories for term weighting. One main disadvantage of this method is the problem of synonymy and polysemy. Synonymy is where multiple words having the same meaning. Polysemy is where a single word has more meanings.

1. **Phrase Based Method (PBM)**

In this method the document is analysed based on the phrases which are less obvious to more meanings and more discriminative. The disadvantages of this method includes

* They have inferior statistical properties to terms
* They have low frequency of occurrence
* They have large number of noisy an redundant phrases

1. **Concept Based Method (CBM)**

In this method the document is analysed based on sentence and document level. In this method there are three main components. The first component examines the meaningful part of the sentences. The second component produces a conceptual ontological graph to explain the structures. The third component extracts top concepts based on the first two components. This method can differentiate between the important and unimportant words.

1. **Pattern Taxonomy Method (PTM)**

In this method the document is analysed based on the patterns. Patterns in a document can be found out using data mining techniques like association rule mining, sequential pattern mining, frequent item set mining and closed pattern mining. This method uses two processes – pattern deploying and pattern evolving. This method is proved to perform better than all the other models or methods.

##### 2.2.1.8.6 Text Mining Algorithms

Text mining involves applying very computationally-intensive algorithms to large document collections, which can speed up the analysis considerably. The most common ones are highlighted below.

1. **K-Nearest Neighbour (K-NN)**

K-NN is a non-parametric,lazylearning algorithm used for classification and regression. The KNN classifier categorizes an unlabelled test example using the label of the majority of examples among its k-nearest (most similar) neighbours in the training set. The similarity depends on a specific distance metric, therefore, the performance of the classifier depends significantly on the distance metric used. (Weinberger & Saul, 2009)

Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. The k-NN algorithm is among the simplest of all [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms. When we say a technique is non-parametric, it means that it does not make any assumptions on the underlying data distribution. In other words, the model structure is determined from the data. Therefore, k-NN could and probably should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution data. K-NN is also a lazy algorithm (as opposed to an eager algorithm). What this means is that it does not use the training data points to do any generalization. In other words, there is no explicit training phase or it is very minimal. This also means that the training phase is pretty fast.

A major problem inherited form the design of KNN algorithm is that its classification performance depends on choosing the optimal number of neighbours (k), which is different from one data sample to another. (Hassanat, Abbadi, Altarawneh, & Alhasanat, 2014)

1. **Support Vector Machines**

The support vector machine (SVM) is a supervised learning method that generates input-output mapping functions from a set of labelled training data. The mapping function can be either a classification function, i.e., the category of the input data, or a regression function. (Wang, 2005)

This approach is one of the most effective and accurate classification algorithm. In this approach, hyper-planes and dimension estimation-based techniques are used to discover or classify the data. The main advantage of this algorithm is to achieve high accurate classification results. However, it is quite complex to implement.

In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyper-plane which categorizes new examples. In two dimensional space this hyper-plane is a line dividing a plane in two parts where in each class lay in either side.

1. **Bayesian Classifier**

Bayesian classifiers are statistical classifiers based on Bayes’ Theorem of probability to predict the class membership probabilities, such as the probability that a given sample belongs to a particular class. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. (Leung, 2007)

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these attributes depend on each other or upon the existence of the other attributes, all of these properties independently contribute to the probability that this fruit is an apple. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered “naive”.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

1. **K-Mean**

K-means is an unsupervised learning algorithm that solves the ‘clustering’ problem, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

* The centroids of the K clusters, which can be used to label new data
* Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyse the groups that have formed organically. The main disadvantage of *K*-means clustering is that it is indeed very sensitive to the initial choice of the number of *K*. (Allahyari et al, 2017)

1. **Decision Trees**

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data). The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label. (Saxena, 2017)

Decision Tree algorithm belongs to the family of supervised learning algorithms which can be used for solving regression and classification problems too. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making).

Decision tress often mimic the human level thinking so it is so simple to understand the data and make some good interpretations.

Decision trees actually make you see the logic for the data to interpret (not like black box algorithms like SVM, NN, etc.)

#### 2.2.1.9 NLP

**Natural Language Processing (NLP)** is the process of using of computer algorithms to identify key elements in everyday language and extract meaning from unstructured spoken or written input.  NLP is a discipline of computer science that requires skills in artificial intelligence, computational linguistics, and other machine learning disciplines.

Some NLP efforts are focused on mimicking human-like responses to queries or conversations.  Others try to understand human speech through voice recognition technology, such as the automated customer service applications used by many large companies.

Many NLP systems “learn” over time, reabsorbing the results of previous interactions as feedback about which results were accurate and which did not meet expectations.  These machine learning programs can operate based on statistical probabilities, which weigh the likelihood that a given piece of data is actually what the user has requested.  Based on whether or not that answer meets approval, the probabilities can be adjusted in the future to meet the evolving needs of the end-user (Bresnick, 2016).

#### 2.2.1.10 Data Visualization

Data Visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context (pictorial or graphical format). Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software (Rouse, 2012).

### 2.2.2 THE ROLE OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE

The main aim of Artificial Intelligence is to mimic human cognitive functions. This is definitely relevant in the healthcare industry and even better because the increasing availability of healthcare data and rapid progress of analytics techniques would serve as catalysts to its growth and impact in the industry. Guided by relevant clinical questions, powerful AI techniques can unlock clinically relevant information hidden in the massive amount of data, which in turn can assist clinical decision making.

Before deploying AI systems in healthcare applications, there is a need for such systems to be ‘trained’ using data generated from clinical activities, such as screening, diagnosis, treatment assignment and so on, so that they can learn similar groups of subjects, associations between subject features and outcomes of interest.

AI devices fall into two major categories. The first category includes machine learning (ML) techniques that analyse structured data such as imaging, genetic and EP data. In the medical applications, the ML procedures attempt to cluster patients’ traits, or infer the probability of the disease outcomes. The second category includes natural language processing (NLP) methods that extract information from unstructured data such as clinical notes/medical journals to supplement and enrich structured medical data. The NLP procedures target at turning texts to machine-readable structured data, which can then be analysed by ML techniques. (Jiang, Jiang, Zhi, Dong, Li, Ma, Yilong, Wang, Dong, Shen & Wang, 2017)

The following highlight some of the roles that AI has plays in the healthcare industry:

1. **Virtual nurses**

AI can be applied in the form of a personal assistant who monitors and assists patients with some of their needs when clinical personnel are not available. Since virtual nurses are available 24/7, they are able to monitor patients and provide quick answers to questions. For example, Sensely’s “Molly” is an AI-powered nurse avatar being used by UCSF and the UK’s NHS to interact with patients, ask questions about their health, assess their symptoms, and direct them to the most effective care setting. AI-powered nurse assistants could save 20% of the time nurses spend on patient maintenance tasks. Another example is Care Angel's virtual nurse assistant can even provide wellness checks through voice and AI. (Kalis, Collier & Fu, 2018)

1. **AI-assisted robotic surgery**

Robots are able to analyse data from a patient’s pre-operation medical records to guide a surgeon's instrument during surgery. This can generate a 21% reduction in patients’ length of stay in the hospital following surgery, as a result of fewer complications and errors. The incisions made during Robot-assisted surgery are considered "minimally invasive" so patients won't need to heal from large incisions. AI enables robots use data from past operations to inform new surgical techniques. A study of 379 orthopaedic patients found that AI-assisted robotic procedure resulted in five times fewer complications compared to surgeons operating alone. (Schroerlucke, Wang, Cannestra, Good, Lim, Hsu & Zahrawi, 2017). At Oxford’s John Radcliffe hospital, a robot was used on an eye surgery for the first time. (Edwards, Xue, Meenink, Beelen, Naus, Simunovic, Latasiewicz, Farmery, de Smet, & MacLaren, 2018). The most advanced surgical robot, the Da Vinci allows doctors to perform complex procedures with greater control than conventional approaches. Heartlander, a miniature robot developed by the Robotics Institute at the Carnegie Mellon University, assists heart surgeons as it enters a small incision on the chest to perform mapping and therapy over the surface of the heart. (Marr, 2018, Kalis et al, 2018)

1. **Decision support**

The use of AI to aid clinical judgement or diagnosis is undoubtedly in its infancy, however, recent results are showing the possibility. In 2017, a group at Stanford University tested an AI algorithm against 21 dermatologists on its ability to identify skin cancers. The results prove that it is capable of classifying skin cancer with a level of competence comparable to dermatologists. (Esteva, Kuprel, Novoa, Ko, Swetter, Blau & Thrun, 2017) A Danish AI software company tested its deep-learning program by having a computer eavesdrop while human dispatchers took emergency calls. The algorithm then analyzed what a person says, together with the tone of voice and background noise and detected cardiac arrests with a 93% success rate compared to 73% for humans (Marr, 2018, Kalis et al, 2018, Kahn, 2018)

1. **Fraud detection**

Errors and fraud pose an expensive problem for healthcare organizations and also for insurers. Fraud detection has traditionally relied on a combination of computerized (rules-based) and manual reviews of medical claims which is a time-consuming process. Health insurers are experimenting with AI-supported data mining, coupled with AI-based neural networks (which mimic the processes of the human brain, but much more quickly) to search Medicare claims for patterns associated with medical reimbursement fraud. AI could be used to monitor and detect abnormal interactions with proprietary data and this would reduce health record breaches. (Kalis et al, 2018)

## 2.3 NAtural LANGUAGE PROCESSING

### 2.3.1 WHAT IS NLP?

Natural Language Processing (NLP), is simply the ability of a machine or piece of software to understand, analyse and process human language within the context in which it is spoken or written. NLP, otherwise known as computational linguistics, is a constituent of Artificial Intelligence, and with the combination of machine learning and linguistics, it is concerned with the processing of significant amounts of natural language data and therefore allows humans to interact seamlessly with machines as if they were human. (Khurana, Koli, Khatter & Singh, 2017)

### 2.3.2 HISTORY AND EVOLUTION OF NLP

The very first NLP systems were based on difficult and compounded sets of handwritten rules. Later on, as the technology advanced, these handwritten rules were replaced by machine learning algorithms which were told what words and phrases to look for in texts and given specific responses when those phrases appeared. They were implemented in the form of if-then-else decision trees and other statistical models.

The current approaches to NLP are based on ‘Deep Learning’ which is a breed of artificial intelligence which analyses different data patterns to consistently amend, improve and make better the program’s understanding. Deep Learning models require large amounts of both labelled as well as unlabelled data to train on and identify relevant correlations. They can even learn different grades or tiers of representation of concepts. (Burns, 2017 and Gall, 2018)

### 2.3.3 WHY NLP?

Of the data available on computer networks, about 70%-80% comprises of unstructured data, which is also known as free-form text. The amount of information and insights available within this unstructured data is hidden and therefore unavailable to those who need them, such governments, public services, businesses, and individuals. Humans have to read these texts or devise some other means to derive information value from them. Natural language processing can be applied to characterize, interpret, or understand the information content of free-form text.

The amount of data generated each day on the internet is about [2.5 quintillion bytes of data](https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1) (Marr, 2018). This is undoubtedly, a massive amount of data. Business managers all over the world have a Big Data problem as they are drowning in streams of unstructured data, dashboards and spreadsheets and are trying to compile all of it together in order to generate meaningful information and draw insights from such data.

Google, Netflix, data companies, video games and more all use AI to comb through large amounts of data. The end result is insights and analysis that would otherwise either be impossible or take far too long.

Presently, natural language processing aims to interpret text in a number of ways.

* Most aim to interpret text according to arbitrary notions of affective content or similarity or as arising from emotion or feeling (as in sentiment analysis, text clustering, and document classification). This engineering approach to text processing currently enjoys commercial success.
* Some efforts today aim to interpret free-form text to extract information with which to answer directed questions or populate databases (as in information extraction, question-answering, and bioinformatics). This work requires processing with a comparatively more refined sensitivity for intended meaning.
* Various other efforts attempt to understand the meaning of texts by examining morphemes, words, phrases, clauses, sentences, and other elements of linguistic structure to reveal the complicated details of conceptual content encoded there.
* The most ambitious of these efforts require integrated theories of conceptual semantics and human cognition, which is the domain where ground truth meaning resides.

The particular allure of NLP systems is that they “learn” over time, reabsorbing the results of previous interactions as feedback about which results were accurate and which did not meet expectations. They therefore deliver better and better insights as more interactions occur

### 2.3.4 APPLICATIONS OF NLP SYSTEMS

The following highlight the applications of NLP Systems:

1. **Machine Translation**:This involves the automatic translation of one human language into another human language.
2. **Named entity extraction:**In data mining, a named entity definition is a phrase or word that clearly identifies one item from a set of other items that have similar attributes. Examples include first and last names, age, geographic locations, addresses, phone numbers, email addresses, company names, etc. Named entity extraction, sometimes also called named entity recognition, makes it easier to mine data.
3. **Optical Character Recognition (OCR):**This involves turning images, like PDF documents or scans of care summaries and imaging reports, into text files that can then be parsed and analysed
4. **Question-Answer Session:**Determine the answer to a question asked in human language.
5. **Topic Segmentation:**Under this, a given piece of text is separated into segments or parts with each part being related to a specific topic. This process also involves determining the topic to which the corresponding segment is related.
6. **Speech Recognition:**When a person speaks in his/ her language; the machine should be able to recognize the speech and convert it to its corresponding textual representation.
7. **Sentiment Analysis:**Through this emotion of the comments can be judged to be positive, negative or neutral. It serves as a big help to brands that need to judge the performance of their business by analysing consumer comments.
8. **Sentence segmentation, part-of-speech tagging, and parsing:**Natural language processing can be used to analyse parts of a sentence to better understand the grammatical construction of the sentence.
9. **Deep analytics:**Deep analytics involves the application of advanced data processing techniques in order to extract specific information from large or multi-source data sets. Deep analytics is particularly useful when dealing with precisely targeted or highly complex queries with unstructured and semi-structured data. Deep analytics is often used in the financial sector, the scientific community, the pharmaceutical sector, and biomedical industries. Increasingly, however, deep analysis is also being used by organizations and companies interested in mining data of business value from expansive sets of consumer data.
10. **Co-reference resolution:**In a chunk of text, co-reference resolution can be used to determine which words are used to refer to the same objects.
11. **Automatic summarization:**Natural language processing can be used to produce a readable summary from a large chunk of text. For example, one might us automatic summarization to produce a short summary of a dense academic article, by identifying key concepts or phrases present in the source material.
12. Mapping data elements present in unstructured text to structured fields in an electronic health record in order to improve clinical data integrity
13. Converting data in the other direction from machine-readable formats into natural language for reporting and educational purposes
14. Answering unique free-text queries that require the synthesis of multiple data sources

### 2.3.5 IMPLEMENTATION OF NLP SYSTEMS

The first step in NLP depends on the application of the system. For voice-based systems such as Alexa or Google Assistant, voice needs to be translated into text. That is usually done using the Hidden Markov Models system (HMM). The HMM uses math models to determine what has been said and translate that into text usable by the NLP system. Simply put, the HMM listens to 10- to 20-millisecond clips of the speech and looks for phonemes (the smallest unit of speech) to compare with pre-recorded speech. (Mills, 2018)

Next is the actual understanding of the language and context. Each NLP system uses slightly different techniques, but on the whole, they're fairly similar. The systems try to break each word down into its part of speech (noun, verb, etc.). This happens through a series of coded grammar rules that rely on algorithms that incorporate statistical machine learning to help determine the context of what you said.

If the system is not a speech-to-text NLP, the system just skips the first step and moves directly into analysing the words using the algorithms and grammar rules. By the end of the process, the computer should understand the meaning of what you said. There are several challenges in accomplishing this when considering problems such as words having several meanings (polysemy) or different words having similar meanings (synonymy), but developers encode rules into their NLU systems and train them to learn to apply the rules correctly. (Mills, 2018)

**Semantic Analysis**

Semantic analysis cannot be left out when explaining NLP. One could even argue that semantic analysis helps form the backbone of natural language processing as it is closely related to NLP.

Semantic analysis is how NLP AI interprets human sentences logically. Semantic analysis helps to add content after the HMM method breaks down the sentence into their basic structure. For instance, if an NLP program looks at the word "discharge" it needs context to determine if the text refers to “discharge” as regards to fluid that comes from a person’s body or “discharge” which is used when a person is allowed to go patient is deemed fit to go home by the doctor.

If the HMM method breaks down text and NLP allows for human-to-computer communication, then semantic analysis allows everything to make sense contextually. (Mills, 2018)

### 2.3.6 NLP IN HEALTH CARE

There is an explosion of all sorts of data in the healthcare space and most healthcare providers still lack access to coherent external data and find it hard to extract actionable insights from all the unstructured data available within the EHR/EMR systems. The industry needs to find the best ways to extract what’s relevant and bring it together to help clinicians make the best decisions for their patients. NLP algorithms have already proven valuable in this venture, largely showing potential in simplifying clinical documentation and enabling voice-to-text dictation.

#### 2.3.6.1 ROLE OF NLP IN HEALTHCARE

An NLP-powered context engine can mine unstructured documents at high speed, uncover valuable insights from all the data at hand and deliver actionable suggestions around clinical decision support. NLP has played a vital role in the healthcare industry in the following ways:

1. **Identifying patients in need of improved care coordination**

NLP helps organizations quickly identify patients who may be eligible for immunotherapies, clinical trials and medical research. Automated case finding ensures patients are informed about all possible care options early in the disease process versus post discharge thereby making clinical trials more efficient and effective. (Bonney, 2015)

Machine learning and NLP tools have also shown potential for detecting complex patients who may benefit from enhanced care coordination by extracting meaningful information from large datasets, these tools can provide clinicians with the information they need to detect complex patients.

1. **Improving the healthcare provider’s interactions with patients and the EHR**

NLP tools, such as voice recognition, may offer a viable solution to EHR problem. Many clinicians already make use of this technology as an alternative to typing or handwriting clinical notes. This helps improve patient-provider interactions and reduce EHR frustration.

Providers have described it as a more cooperative and efficient approach as they can talk to both the record and the patient at the same time. It increases the amount of time spent in the patient’s presence.”

By allowing providers to dictate their notes, this approach makes documentation requirements easier also improves the quality of the documentation, which may make it more useful for analytics downstream.

In the future, voice recognition tools may go beyond clinical dictation to receive and carry out directions from providers.

1. **Improving patient health literacy**

With more organizations using patient portals, patients can now access their health data, make more informed medical decisions, and keep their health on track. However, the data presented is often taken directly from sources intended primarily for physician use with little consumer-friendly interpretation. Therefore, the benefits of patient data access are however lessened if patients can’t make sense of what their data means.

Physicians must often spend extra time defining terms for patients and soothing the anxieties of those who may have misread a diagnosis or lab test result.

By applying NLP to EHR data and integrating the results into the patient portal, providers could improve patients’ understanding of their health information. Confusing and complex medical jargon translated into everyday speech may help patients feel more in control of the information presented to them (Kent, 2018)

1. **Provide real-time data about patients.**

Hospitals use NLP concurrently to monitor patients while they are in the hospital receiving treatment. For example, as physicians document certain diagnoses or procedures (e.g., insertion of urinary catheter), NLP generates alerts and reminders so providers monitor these patients frequently to mitigate the risk of urinary tract infection—one of many hospital-acquired conditions identified by CMS. Another example is using NLP to track sepsis patients. NLP helps uncover the symptoms immediately rather than retrospectively. Physicians can then direct their attention toward these patients and reduce the likelihood that they will be readmitted. (Bonney, 2015)

#### 2.3.6.2 CHALLENGES OF INTEGRATING NLP INTO HEALTHCARE

The major challenge involved in NLP is that human communicate in a speech that is often ambiguous, uncertain and unclear, having different tones and incorporating different slangs, regional phrases and different contexts. It is then difficult to make a machine recognize human speech, understand the natural language and be able to generate the natural language.

Unstructured clinical notes and narrative text still present a major problem for computer scientists.

“[Clinical text]… is often ungrammatical, consists of ‘bullet point’ telegraphic phrases with limited context, and lacks complete sentences. Clinical notes make heavy use of acronyms and abbreviations, making them highly ambiguous.” (Townsend, 2013)

Up to a third of clinical abbreviations in the Unified Medical Language System (UMLS) Metathesaurus have multiple meanings, and more than half of terms, acronyms, or abbreviations typically used in clinical notes are puzzlingly ambiguous, Townsend added.

“For example, ‘discharge’ can signify either bodily excretion or release from a hospital; ‘cold’ can refer to a disease, a temperature sensation, or an environmental condition,” she explained. “Similarly, the abbreviation ‘MD’ can be interpreted as the credential for ‘Doctor of Medicine’ or as an abbreviation for ‘mental disorder.’”

While the human brain can usually decipher these types of differences by relying on the context of the surrounding words for clues, NLP technology still has a long way to go before it can reach the same reliability threshold as the typical human reader.

Using NLP to fill in the gaps of structured data on the back end is also a challenge.  Poor standardization of data elements, insufficient data governance policies, and infinite variation in the design and programming of electronic health records have left NLP experts with a big job to do. (Bresnick, 2016)

## 2.4 REVIEW OF RELATED WORKs

### 2.4.1 DATA-TO-TEXT SUMMARISATION OF PATIENT RECORDS: USING COMPUTER-GENERATED SUMMARIES TO ACCESS PATIENT HISTORIES (Scott, Hallett and Fettiplace, 2013)

**Problem:** Current studies have shown that the quality of healthcare outcomes increases when doctors are able to spend sufficient time with their patients to explore their symptoms, explain their condition and negotiate their treatment plan. However, there is limited time for clinicians to interact with their patients and so there is need for them to receive information in an easy-to-digest format. The use of textual summaries was adopted as the clinicians are able to get a fast overview of a patient’s medical record.

**Objective:** To access the efficiency and accuracy of automatically generated textual summaries of patients’ medical histories at the point of care.

**Methodology:** A computational system (a Report Generator) was developed to produce a range of summarised reports of patient records from patient histories derived from a repository of medical records of cancer patients and composed of narrative documents (e.g., letters, discharge reports, etc.) and structured data (e.g., test results, prescriptions, etc.)

The input to the Report Generator is a Chronicle which is a more structured version of the Electronic Health Record. This Chronicle facilitates better and easier text generation and allows for a higher degree of flexibility of the generated text.

The output of the Report Generator is a range of textual summaries of the information contained in the Chronology. These range in length from short paragraphs to many pages. In the current implementation, the generator produces two main types of report. The first is a longitudinal report, which is intended to provide a quick historical overview of the patient's illness, whilst preserving the main events (such as diagnoses, investigations and interventions). It presents the events in the patient's history ordered chronologically and grouped according to type. In this type of report, events are fully described (i.e., an event description includes all the attributes of the event)

The second type of report focuses on a given type of event in a patient's history, such as the history of diagnoses, interventions, investigations or drug prescription. This allows us to provide a range of reports that are presented from different perspectives.

**Result:** Twenty-one clinicians were presented with information about two cancer patients and asked to answer key questions. For each clinician, the information on one of the patients comprised their official hospital records, and for the other patient it comprised summaries that were computer-generated by a natural language generation system from data extracted from the official records. We measured the accuracy of the clinicians' responses to the questions, the time they took to complete them, and recorded their attitude to the computer-generated summaries.

Accuracy: The results show that clinicians are slightly better at answering the set of key questions when using the automatically-generated record summaries than the (traditional) full records. They provide the correct answers 80% of the time when using the summaries, and only 75% of the time when using the full records.

Efficiency: The results show that use of the summaries reduced significantly the time taken to respond to the set of questions for each patient. Overall, using the summaries allowed the clinicians to shave off just over 50% of the time taken to answer all the questions compared to using the records.

**Limitation:** According to the doctors that were used in the observation, many recorded that the summaries did not contain enough details and that the longitudinal summary, with more detail about presenting complaint would be more useful.

**Conclusion:** AI-based computer-generated textual summaries of patient histories can be as accurate as, and more efficient than, human-produced patient records for clinicians seeking to accurately identify key information about a patients overall history.

### 2.4.2 MINING PHYSICIANS’ NOTES FOR MEDICAL INSIGHTS (Hardesty, 2012)

**Problem:** Electronic records could contain a wealth of medically useful data: hidden correlations between symptoms, treatments and outcomes, for instance, or indications that patients are promising candidates for trials of new drugs.

However, much of that data is buried in physicians’ freeform notes. Word-sense disambiguation is a term used by computer scientists to describe the difficulty in extracting data from unstructured text. The ability to infer words’ intended meanings makes it much easier for computers to find useful patterns in mountains of data.

**Objective:** To present a new system for disambiguating the senses of words used in doctors’ clinical notes.

**Methodology:** The inspiration for this system was taken from a branch of research known as topic modelling. Topic modelling seeks to automatically identify the topics of documents by inferring relationships among prominently featured words. Topic modelling assigns a mathematical weight to each theme in a text, using an algorithm to determine which sense of a word is likely meant based on the surrounding language.

This “fundamentally new approach” will allow much more accurate systems to function without human supervision, reducing the time and cost of manual transcription efforts.  The more data that is processed by the system, the more accurate it becomes as it learns where it was correct and where its inferences failed.  The thesaurus of medical terms known as the Unified Medical Language System (UMLS), is included to widen the system’s available knowledge base and improve its word association capabilities.

**Result:** The system is 75% accurate in deciphering words with multiple meanings in the freehand portion of a physician’s medical notes contained in an electronic health record (EHR).

### 2.4.3 AUTOMATED IDENTIFICATION AND PREDICTIVE TOOLS TO HELP IDENTIFY HIGH-RISK HEART FAILURE PATIENTS: PILOT EVALUATION. (Evans, Benuzillo, Horne, Lloyd, Bradshaw, Budge, Rasmusson, Roberts, Buckway, Geer, Garrett, Lappé, 2016)

**Objective:** Develop and evaluate an automated identification and predictive risk report for hospitalized heart failure (HF) patients.

**Methodology:** Natural Language Processing was used to analyse dictated free-text reports from the previous 24 hours each day to help improve the early identification of patients hospitalized with Heart Failure. Also, another application was developed using an Intermountain Healthcare-developed predictive score to determine each HF patient's risk for 30-day hospital readmission and 30-day mortality period. That information was then included in an identification and predictive risk report, with further evaluation carried out at a hospital that treats high-risk HF patients.

**Result:** Using NLP to identify HF patients increased the identification score's sensitivity from 82.6% to 95.3% and its specificity from 82.7% to 97.5%, and the model's positive predictive value is 97.45%. The insight provided by the HF identification and predictive report is now used in daily multidisciplinary discharge planning meetings. Compared to the previously used manual methodology, less time is taken by the clinician to review potential HF admissions (10 vs 40 min). A significant reduction in 30-day mortality as well as a significant increase in patient discharges to home care instead of to a specialized nursing facility have been identified after evaluating the use of the HF predictive report.

**Conclusion:** Using clinical decision support to help identify HF patients and automatically calculating their 30-day all-cause readmission and 30-day mortality risks, coupled with a multidisciplinary care process pathway, was found to be an effective process to improve HF patient identification, significantly reduce 30-day mortality, and significantly increase patient discharges to home care.

### 2.4.4 NATURAL LANGUAGE PROCESSING TO DETECT RISK PATTERNS RELATED TO HOSPITAL ACQUIRED INFECTIONS (Denys Proux, Marchal, Segond, Kergourlay, Darmoni, Pereira, Gicquel, Metzger, 2010)

Hospital Acquired Infections (HAI) has a major impact on public health and on related healthcare cost. HAI experts are fighting against this issue but they are struggling to access data. Information systems in hospitals are complex, highly heterogeneous, and generally not convenient to perform a real time surveillance. Developing a tool able to parse patient records in order to automatically detect signs of a possible issue would be a tremendous help for these experts and could allow them to react more rapidly and as a consequence to reduce the impact of such infections. Recent advances in Computational Intelligence Techniques such as Information Extraction, Risk Patterns Detection in documents and Decision Support Systems now allow to develop such systems.

**Problem:** Performing hospital-wide surveillance to monitor, detect and prevent Hospital Acquired Infections among patients is a complex and time consuming task.

**Objective:** To detect Hospital Acquired Infections by using risk patterns identification methods in patient records in order to reduce the number of unnoticed cases and time for reaction.

**Methodology:** Natural Language Processing (NLP) techniques will be applied to parse Patient Discharge Summaries and identify specific terms and sequences of facts.

**Result:** The NLP algorithm developed monitors and prevents Hospital Acquired Infections by scanning patients’ records and automatically detects signs of a possible issue. This helps experts to react more rapidly and as a result, reduce the impact of such infections.

### 2.4.5. Text Mining and Medicine: Usefulness in Respiratory Diseases (Piedra, Ferrer, Gea, 2014)

The review was done to provide an overview of text and data mining in the exercise of care in respiratory medicine and research in the same field.

The initial problem was the high level of financial and social challenges that common respiratory diseases were presenting to health and healthcare services. There was a need to build systems that analyse large volumes of clinical data relating to respiratory diseases and then use the data for research and to improve medical practices.

In this study, it was discovered that through text mining, research pathways that have been unexplored can then be developed so as to bring about new therapeutic alternatives.

It also became clear that the use of text mining to analyse knowledge from various basic sciences can shed new light on respiratory diseases.

In the area of diagnosis and clinical management of the respiratory patient, it has been established that the automatic analysis of data allows new associations between variables to be established, thereby generating hypothesis that can lead to alternative or complementary diagnostic systems.

In the care of semicritical and critical respiratory patients, data mining has proved particularly fertile. Data collection systems have been set up over the last decades in units catering for these patients. As a result, two major applications based on text mining have been developed. One is the generation of short- and mid-term predictive models that allow the creation of intelligent alerts and decision-making systems. One of particular importance is the use of data mining and neural networks to develop a clinical decision support model, based on the identification of key physiological variables for determining when mechanical ventilation should be initiated.

It was concluded that the most obvious areas for the medical application of text mining (both in research and in the clinic) are the integration and transfer of advances made in the most basic sciences, and a better understanding of the diagnostic processes, severity classifications and determination of disease prognosis.

**REFERENCES**

Agius, M. (2014). The medical consultation and the human person. *Psychiatr Danub*, *26*(Suppl 1), 15-18.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Andale, S. (2015) Data Mining: Simple Definition, Uses & Techniques. *Statistics How To.* Retrieved on November 5 2018 from https://www.statisticshowto.datasciencecentral.com/data-mining/

Balamurugan, R. and Pushpa, S. (2015). A Review on Various Text Mining Techniques and Algorithms.

Bernstein, A. B, Hing, E., Moss, A.J., Allen, K.F., Siller, A.B., Tiggle, R.B. (2003). Health care in America: Trends in utilization. Hyattsville, Maryland: National Center for Health Statistics.

Bonani, B., (2018) Techniques and applications of text mining. *Digital Vidya.* Retrieved on November 21 2018 from https://www.digitalvidya.com/blog/techniques-applications-text-mining/

Bresnick, J. (2016*)* What is the Role of Natural Language Processing in Healthcare*? HealthITAnalytics.* https://healthitanalytics.com/features/what-is-the-role-of-natural-language-processing-in-healthcare

Bresnick, J. (2017) Natural Language Processing Could Translate EHR Jargon for Patients*. HealthITAnalytics.* <https://healthitanalytics.com/news/natural-language-processing-could-translate-ehr-jargon-for-patients>

Burns, E. (2016) Machine Learning (ML) *TechTarget.* Retrieved on November 5 2018 from <https://searchenterpriseai.techtarget.com/definition/machine-learning-ML>

Burns, E. (2017) Natural Language Processing. *TechTarget.* Retrieved on November 21 2018 from https://searchbusinessanalytics.techtarget.com/definition/natural-language-processing-NLP

Burton, L. C., Anderson, G. F., & Kues, I. W. (2004). Using Electronic Health Records to Help Coordinate Care. *The Milbank Quarterly*, *82*(3), 457–481. <http://doi.org/10.1111/j.0887-378X.2004.00318.x>

Chamatkar, M. A. J., & Butey, P. K. (2014). Importance of Data Mining with Different Types of Data Applications and Challenging Areas. *Journal of Engineering Research and Applications*, *4*(5), 38-41.

Christensson, P. (2006). Data Definition. <https://techterms.com>

[Copeland](https://www.britannica.com/contributor/BJ-Copeland/4511), J. (2018)**.** Artificial intelligence. *Encyclopaedia Britannica, Inc.* <https://www.britannica.com/technology/artificial-intelligence>

Cress, K. (2009). 5 Types of Information Systems. *MindMeister.* Retrieved October 4 2018 from <https://www.mindmeister.com/37310006/5-types-of-information-systems>

Davoren, J. (2018). Types of Information Systems in an Organization. *Small Business - Chron.com*. Retrieved October 4 2018 from <http://smallbusiness.chron.com/types-information-systems-organization-43097.html>

Dormehl, L. (2018). What is an artificial neural network? Here’s everything you need to know. *DigitalTrends.* Retrieved on October 5 2018 from https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/

Edwards, T., Xue, K., Meenink, H., Beelen, M., Naus, G., Simunovic, M., Latasiewicz, M., Farmery, A., de Smet, M. & MacLaren, R. (2018) First-in-human study of the safety and viability of intraocular robotic surgery. *Nature Biomedical Engineering, Volume 2, pages 649–*656 (2018)

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H. & Thrun, S. (2017) Dermatologist-level classification of skin cancer with deep neural networks *Nature volume 542, pages115–118*

Evans, S., Benuzillo, J., Horne, B., Lloyd, J., Bradshaw, A., Budge, D., Rasmusson, K., Roberts, C., Buckway, J., Geer, N., Garrett, T., Lappé, D. (2016) Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation, Journal of the American Medical Informatics Association, Volume 23, Issue 5, 1 September 2016, Pages 872–878, <https://doi.org/10.1093/jamia/ocv197>

Fischer, M. and Ereaut, G. (2012). When Doctors and Patients Talk: Making sense of the consultation. *The Health Foundation.* pp 10

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17).

Gall, R. (2018) Natural Language Processing: An Essential Element of Artificial Intelligence. *SmartDataCollective.*

Grout, J. (2014) Understanding your Lab Report. *Diabetes Self-Management.* Retrieved October 3 2018 from https://www.diabetesselfmanagement.com/about-diabetes/diabetes-basics/understanding-your-lab-test-results/

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hardesty, L. (2012). Mining physicians’ notes for medical insights. *MIT News Office*

Hunter, W. (2007) How to define data, information and knowledge. *TechTarget.* Retrieved on November 5 2018 from <https://searchdatamanagement.techtarget.com/feature/Defining-data-information-and-knowledge>

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S, Yilong, Wang, Y., Dong, Q., Shen, H. & Wang, Y. (2017) Artificial intelligence in healthcare: past, present and future. *BMJ Journals, Stroke and Vascular Neurology*; Issue *4, Volume 2 doi:10.1136/svn-2017-000101*

Kahn, J. (2018) The AI That Spots a Stopped Heart. Retrieved on November 5 2018 from https://www.bloomberg.com/businessweek

Kalis, B., Collier, M., Fu, R. (2018) 10 Promising AI Applications in Health Care. Retrieved on November 8 2018 from <https://hbr.org/2018/05/10-promising-ai-applications-in-health-care>

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges. *arXiv preprint arXiv:1708.05148*.

Kivumbi (2011). "Difference between Knowledge and Information." DifferenceBetween.net. Accessed from http://www.differencebetween.net/language/difference-between-knowledge-and-information on September 6, 2011

Kok, J. N., Boers, E. J., Kosters, W. A., Van der Putten, P., & Poel, M. (2009). Artificial intelligence: definition, trends, techniques, and cases. *Encyclopedia of Life Support Systems (EOLSS)*

Lucas, P., & Van Der Gaag, L. (1991). *Principles of expert systems*. Wokingham: Addison-Wesley.

Marr, B. (2018) How Is AI Used In Healthcare - 5 Powerful Real-World Examples That Show The Latest Advances. Retrieved on November 8 2018 from https://www.forbes.com/sites/bernardmarr/2018/07/27/how-is-ai-used-in-healthcare-5-powerful-real-world-examples-that-show-the-latest-advances/#327a278b5dfb

[Mills](https://www.forbes.com/sites/forbestechcouncil/people/terencemills/), T. (2018) What Is Natural Language Processing And What Is It Used For? *Forbes Technology Council.* https://www.forbes.com/sites/forbestechcouncil/2018/07/02/what-is-natural-language-processing-and-what-is-it-used-for/#63baa7345d71

Moreno, A., & Redondo, T. (2016). Text analytics: the convergence of big data and artificial intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 3, No 6, 57-64.

[Owen-Hill](https://blog.robotiq.com/author/alex-owen-hill), A. (2017). What’s the difference between Robotics and Artificial Intelligence? Retrieved on November 5 2018 from <https://blog.robotiq.com/whats-the-difference-between-robotics-and-artificial-intelligence>

[Patidar](https://dzone.com/users/3306999/shailnap.html), S. (2018)AI Robot — Robotics and Artificial Intelligence. AI Zone. Retrieved on November 5 2018 from <https://dzone.com/articles/ai-robot-robotics-and-artificial-intelligence>

Raw data. (n.d.) Mosby's Medical Dictionary, 8th edition. (2009). Retrieved October 1 2018 from <https://medical-dictionary.thefreedictionary.com/raw+data>

Rouse, M, (2016). Machine Learning*. TechTarget.* Retrieved on September 10, 2018 from <https://searchenterpriseai.techtarget.com/definition/machine-learning-ML>

Samuel R. Schroerlucke; Michael Y. Wang; Andrew F. Cannestra; Christopher R. Good; Jae Lim; Victor W. Hsu; Faissal Zahrawi (2017) Complication Rate in Robotic-Guided vs Fluoro-Guided Minimally Invasive Spinal Fusion Surgery: Report from MIS Refresh Prospective Comparative Study. *The Spine Journal, ISSN: 1529-9430, Vol: 17, Issue: 10*

Scott, D., Hallett, C., & Fettiplace, R. (2013). Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories. *Patient Education and Counseling*, *92*(2), 153–159. <http://doi.org/10.1016/j.pec.2013.04.019>

Sunil, R. (2015) 6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R)https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science & Applications*, *1*(7), 414-418.

Thomas-MacLean, R., Tarlier, D., Ackroyd-Stolarz, S., Fortin, M., & Stewart, M. (2008). TUTOR-PHC 2003/2004 research trainees. “No cookie-cutter response” conceptualizing primary health care.

Toon, P. D. (1994). What is good general practice? A philosophical study of the concept of high quality medical care. *Occas Pap R Coll Gen Pract. 1994 Jul; (65): i-viii, 1-55.*

Trevino, A. (2016) [Introduction to K-means Clustering](https://www.datascience.com/blog/k-means-clustering). <https://www.datascience.com/blog/k-means-clustering>

Vaughan, J. (2017) Data. *SearchData Management.* Retrieved October 3 2018 from https://searchdatamanagement.techtarget.com/definition/data

Vladimir, Z. (2017). Information system, *Encyclopaedia Britannica, Inc.* <https://www.britannica.com/topic/information-system>

Woodford, C. (2018). Neural Networks

World Health Organization. (2006). Electronic health records: manual for developing countries. Manila: WHO Regional Office for the Western Pacific.

Wu, Y. (2003). An introduction to computer vision. *Northwestern University, Evanston*.

Yao, X. (1999). Evolutionary computation: Theory and applications. World scientific.

Zaïane, O. (1999) CMPUT690 Principles of Knowledge Discovery in Databases. Retrieved on October 5 2018 from <http://www.exinfm.com/pdffiles/intro_dm.pdf>

Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A., & Alhasanat, A. A. (2014). Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. *arXiv preprint arXiv:1409.0919*.

Weinberger, K. Q. & Saul, L. K. (2009) "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *Journal of Machine Learning Research, vol. 10, pp. 207-244*.

Wang, L. (Ed.). (2005). *Support vector machines: theory and applications* (Vol. 177). Springer Science & Business Media.

Leung, K. M. (2007). Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering.

Saxena, R. (2017) How Decision Tree Algorithm Works. *Data Aspirant.* Retrieved on November 21 2018 from <http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/>

Bonney, S. (2015) Practical Applications of NLP in Healthcare: Enhancing the Value of Electronic Data. *PorterResearch.* Retrieved on November 21 2018 from <https://porterresearch.com/practical-applications-of-nlp-in-healthcare-enhancing-the-value-of-electronic-data/>

Taylor, C. (2018) Structured vs. Unstructured Data. *Datamation.*  Retrieved on November 21 2018 from https://www.datamation.com/big-data/structured-vs-unstructured-data.html