THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

DATABASE SYSTEMS JOURNAL

Vol. XII/2021

LISTED IN

RePEc, EBSCO, DOAJ, Open J-Gate,
Cabell's Directories of Publishing Opportunities,
Index Copernicus, Google Scholar,
Directory of Science, Cite Factor,
Electronic Journals Library

BIG DATA

DATA SCIENCE

NoSQL

BUSINESS INTELLIGENCE

CLOUD COMPUTING

DATA MINING

DATA WAREHOUSES

DATABASES

ISSN: 2069 – 3230
dbjournal.ro

Database Systems Journal BOARD

Contact
CaleaDorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: https://dbjournal.ro/
E-mail: editordbjournal@gmail.com

CONTENTS

The Improvement of Decision-making Process using Business Intelligence Solutions

Adelina TĂNĂSESCU
The Bucharest University of Economic Studies, Romania
tanasescu.adelina@gmail.com

The aim of this paper is to present the benefits of using Business Intelligence solutions for the improvement of decision-making process. Data and information are extremely important in the development of a business and in its process of becoming the leader of the market. Business Intelligence solutions offer the opportunity of using descriptive analysis in order to make informed decisions for businesses. These solutions offer the possibility of data visualization, which helps to obtain various benefits.

Key words: *Business Intelligence, data, information, descriptive analysis, data visualization, decision-making process*

1 Introduction

The remarkable development of the IT field has produced visible changes in all branches of the economy. Currently, globalization is constantly growing and its main effect is that there is stronger competition between economic agents than in the past.

Because we speak about rapid changes in the market, the management of companies has to take into consideration two aspects. The first is that they must make decisions based on information from real data about the business. The second thing is about the period of time in which they make these decisions. The success of a business depends on how quickly the business adapts to changes: in the market, in the customers behavior, and changes made by its other competitors. In order to make decisions, the management must obtain the information it needs in a timely manner. Finding good solutions to these changes leads to outperforming competitors and gaining an advantage over them.

The evolution of information systems began in the late 1970s when management information systems (MIS) appeared. Later, in 1980, decision support systems appeared, and in 1985, expert systems appeared. The 1990s are the starting point for the Business Intelligence systems.

The competitive advantages represent the elements that help differentiate the competitors in the market. This type of advantage is the element that makes the difference between businesses in the same market, the difference between a business that stays afloat and another that is constantly evolving [1]. The use of BI solutions offers the possibility of gaining these advantages, because all the information brought by these solutions helps in making well-informed and rapid decisions, which are beneficial for the company.

Business Intelligence solutions offer the possibility to analyze historical data and allow data presentation in a visual form. The analyzed data is related both to the actual activity of the company and to the business environment in which it operates. This is also the reason why BI solutions offer competitive advantages. Thus, the improvement of the decision-making process takes place. The main reason is the increase of the spectrum in the analysis of the company's performance.

1. Data – the source of information

People often confuse data with information, but these are completely different things. Data is the source of information, the raw material, and is not processed. The information is processed data, or we could say it is data with a certain meaning.

Businesses produce a large volume of data through the activity they carry on. By data processing, they obtain information of great interest and use. Businesses should do that in order to obtain new strategies, or to modify the existing ones so that the company's performance is at the level imposed on the market in which it operates. The analysis of information must be carried out in not a very short period of time, so that the decision is not erroneous due to the lack of aspects taken into consideration, but not too long, because another competitor might gain an advantage by making the same decision, but faster.

Data collection and storage should be a necessity for any company. This data is the basis for information that adds value to a company that processes and analyzes it in a timely manner. The management of a company can better understand the causes that led to a certain effect through performing data analysis. For example, businesses use data analysis to see the reasons for sales being lower in a certain unit of the company or a certain period. In this way, the company can find answers to many problems that arise and thus make decisions that help to develop the business.

In the absence of data, the decisions made do not have a basis on anything concrete, and these would be general decisions, not company-specific ones. Several factors fluctuate from one company to another, even if they have the same field of activity. Therefore, it is extremely important that the decisions made by the managers are based on the analysis of their own company's data.

2. Business Intelligence

This chapter is going to present the Business Intelligence process, data warehouses as support in the development of specific BI activities, and descriptive analysis.

2.1 The process of Business Intelligence

Business Intelligence represents the process of analyzing the data within a company in order to improve the decision-making process. BI systems combine data collection and storage with knowledge management and data analysis. By using Business Intelligence, companies receive answers to topics such as: business performance, profit, expenses, and the reasons that led to these results.

The next figure (**Fig. 1**) shows schematically the Business Intelligence process. The first element is the data sources. They are processed using BI tools. In this way, the users, usually the managers, get information that helps them make decisions. These decisions apply in the business environment, which then provides new data that is collected and stored in the data sources [2].

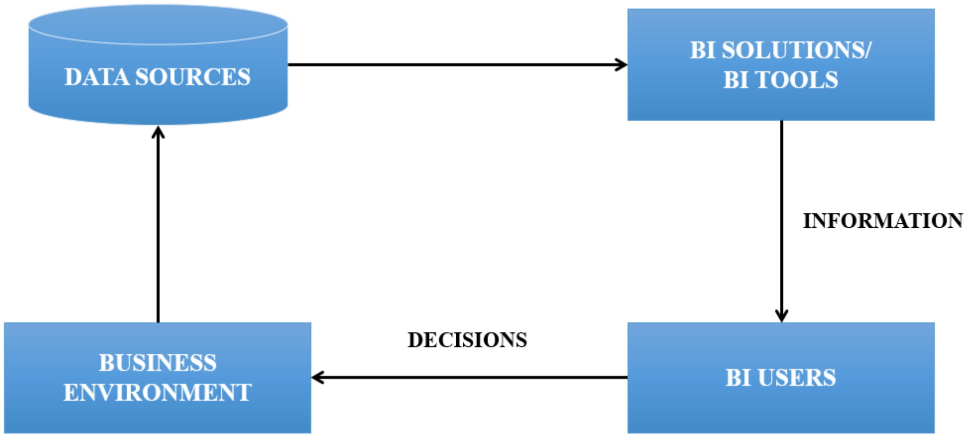


Fig. 1. Business Intelligence process
Source: Ana-Ramona Bologa, Mihaela Muntean – "*Business Intelligence. Teorie și practică*", Bucharest: Editura ASE, 2015

2.2 Data warehouse

If we talk about Business Intelligence, we also talk about the company's historical data. Business Intelligence focuses specifically on the analysis of past data. As a result of the company's activity, the raw data collection is carried out. Data is transformed and processed and then stored in data warehouses. A data warehouse is a data management system used for enabling and supporting specific Business Intelligence activities. Compared to a relational database where we talk about transactional processing – Online Transactional Processing (OLTP), for data warehouses we talk about analytical processing – Online Analytical Processing (OLAP).

Because we consider data warehouses to be a support for Business Intelligence, we must understand that in data warehouses the most frequent operations are those of reporting and analysis. These operations are the opposite of the ones in relational databases, (in relational databases the most frequent operation is the actualization). The frequency of operations performed in these two cases is also different. For relational databases, we talk about daily operations compared to data warehouses where the operations performed assist the decisions, so they are much rarer.

The data sources that help to compile the data warehouse are diverse and different. For this reason, it is necessary to filter and transform the data. Data sources are made of both internal data, from the operational process, and from external data (data about the market evolution, competitors, business environment, etc.).

Business Intelligence helps process this data in order to obtain the necessary information in the decision-making process. The visual representation of information is more likely to be easier to understand than visualization of data in a much narrow context.

2.3 Descriptive analysis

Business Intelligence utilizes descriptive analysis as a method of interpretation of historical data. The main reason for using this analysis is to make comparisons. The most often used key metrics in this type of analysis are those for prices, sales, products, and customers. This data helps with creating an overview of a company at a moment in time [3].

The descriptive analysis involves analyzing raw data to draw useful and understandable conclusions by decision-makers: managers, investors, and other stakeholders. A report showing a certain value of sales lacks context. In fact, this is the one that helps to give meaning to data. The context gives a view of the elements that led to the result achieved by the company in sales. A larger context helps to obtain an informed overview of the company's performance regarding different aspects [3].

There are two main methods of data collection in order to perform descriptive analysis, data aggregation and data mining. In order to obtain information from the data, it must be gathered and analyzed. The descriptive analysis is an important component of performance analysis. The information provides strong support for decision makers to analyze the performance achieved by the business [3].

3. Data visualization

The Business Intelligence process includes data management, but more precisely: data collection, data cleaning and storage, data analysis, and presentation. Data visualization is made according to the user preferences, in different shapes and formats.

People have a better ability to understand information when its presentation is visual, not just in writing. Visualization involves the use of images made of various visual elements of different shapes and colors. A visual analysis of a business's data has the aim to achieve the following objectives [4]:

- Visualization of key metrics for an easier and faster understanding of the data, in order to facilitate the decision-making process;
- Providing an interactive visual way for data exploration;

The benefits of visual data analysis are easy to notice. The main objective of data visualization is the possibility of improving the decision-making process. The means by which data visualization helps to make strategic decisions are the ones below [5]:

- Using a visual representation, we can obtain other information that we could not have noticed by visualizing the data in its raw form. Thus, by data visualization, you can see many of the information hidden behind the numbers, so the data analysis has a broader context.
- Data visualization is an important factor in identifying insights that lead to a better decision-making process. It gives the decision-makers information on different aspects of business's performance, in order to make decisions that are more effective.
- With help from data visualization, the decision-making process takes place in full knowledge of the facts, due to the clear perspective it brings on the values that establish the business's performance.
- Data visualization also helps to view business's progress and notice different trends. In this way, it is easier to notice any slip from the ideal situation and immediately fix it.

4. The benefits of using Business Intelligence

The economy is constantly changing, so decision-makers need to take into consideration many aspects in order to make the right decisions for the development of their businesses and for achieving their desired performance goals, such as making a substantial profit. The aspects that need to be a part of this analysis have as sources data from the operational process and data about the business environment. Among the benefits brought by the use of Business Intelligence solutions, we can mention the following [6]:

- Performing faster data analysis using multiple data sources;
- Increasing organizational efficiency by reducing the time for data analysis;
- Making data-driven decisions by using up-to-date data;
- Improving the experience and satisfaction of customers by analyzing the data that came from them (for example, the customer's reviews);
- Improving the satisfaction of all employees by giving access to their own data, so they will not have to require data from the IT department;
- Using internal and external data sources for a rapid answer to questions;
- Increasing the competitive advantages by analyzing the market and the business's performance within it;

5. Microsoft Power BI

This chapter presents the Microsoft solution for Business Intelligence – Power BI Desktop and an example of application made with this tool.

In February 2021, Gartner considered Microsoft and its Business Intelligence tool – Power BI – among the leaders in the Analytics and BI platforms category. Alongside Microsoft, there are also solutions from Tableau and Qlik, as seen in **Fig. 2**.

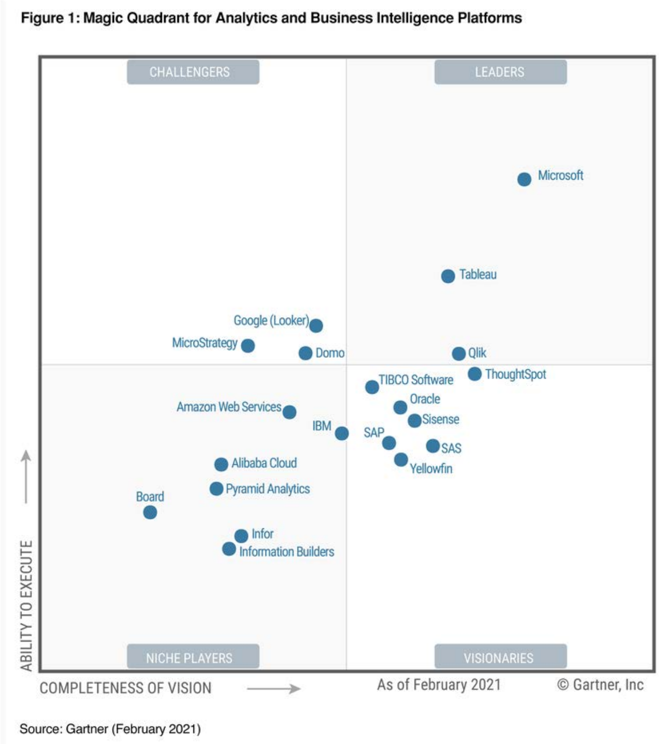


Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms

Source: Gartner (February 2021)

Fig. 2. Magic quadrant for Analytics and Business Intelligence Platforms
Source: https://info.microsoft.com/ww-Landing-2021-Gartner-MQ-for-Analytics-and-Business-Intelligence-Power-BI.html?LCID=EN-US

5.1 Power BI Desktop

Power BI Desktop is a tool for data visualization that allows data transformation and modeling. Reports and dashboards are the final product of data visualization. Among the strengths of this Business Intelligence tool, we can include the reports, dashboards, and various types of data sources. The reports consist of several pages that contain various visual elements, such as graphs, diagrams, tables, etc. The dashboards are pages with multiple data visualizations. As the dashboards consist of only one page, they will contain the most important elements of the business's history [7].

The advantages of using Power BI Desktop are the ability to aggregate data sources, using the relationships between datasets, the interactive interface that facilitates data visualization and allows easier learning compared to other BI tools (such as Tableau and Spotfire); this is also the reason that the collaboration between departments for data analysis is encouraged [8].

5.2 Business Intelligence solution using Power BI Desktop

In order to explain the usefulness of Business Intelligence solutions in improving the decision-making process by using descriptive analysis and data visualization, in the following, we will see a solution made with Power BI Desktop, the BI tool of Microsoft.
Power BI offers the possibility to use various data sources, as seen in **Fig. 3**. Thus, the use of this tool provides flexibility for the company that uses it because it can put together data from various file types, databases, cloud, online services, and others.

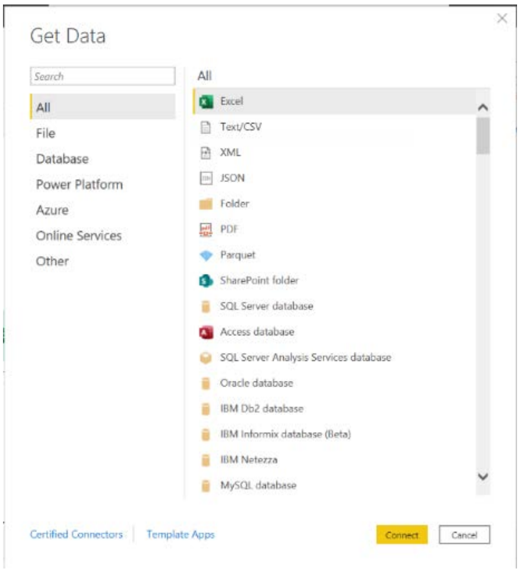


Fig. 3. Data source types used in Power BI Desktop

For this solution, we are using public datasets, two Excel files. These contain data about a company's customers, their orders, the company's profit, discounts, information on the economic regions that customers are from (EU – also presented as UE, EEA – also presented as SEE), and other data.

In order to be able to utilize the data, it must first be cleaned and shaped. Shaping data means that transformation is needed to be able to perform operations like changing formats, removing rows or columns, and renaming columns. In addition, we established the connections between the tables resulting from the datasets. **Fig. 4** shows the connections between the dataset tables.

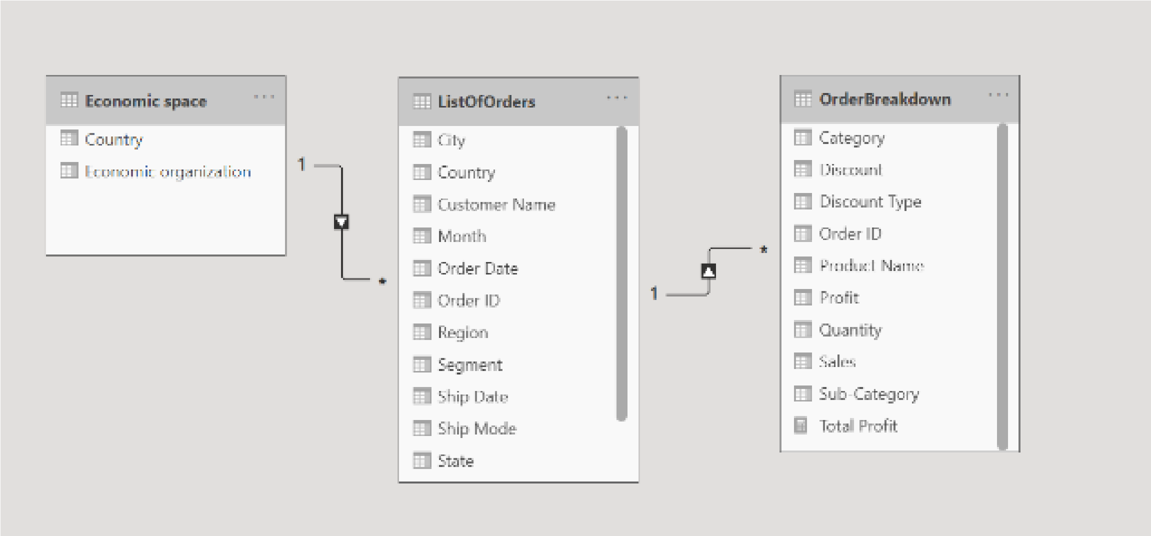
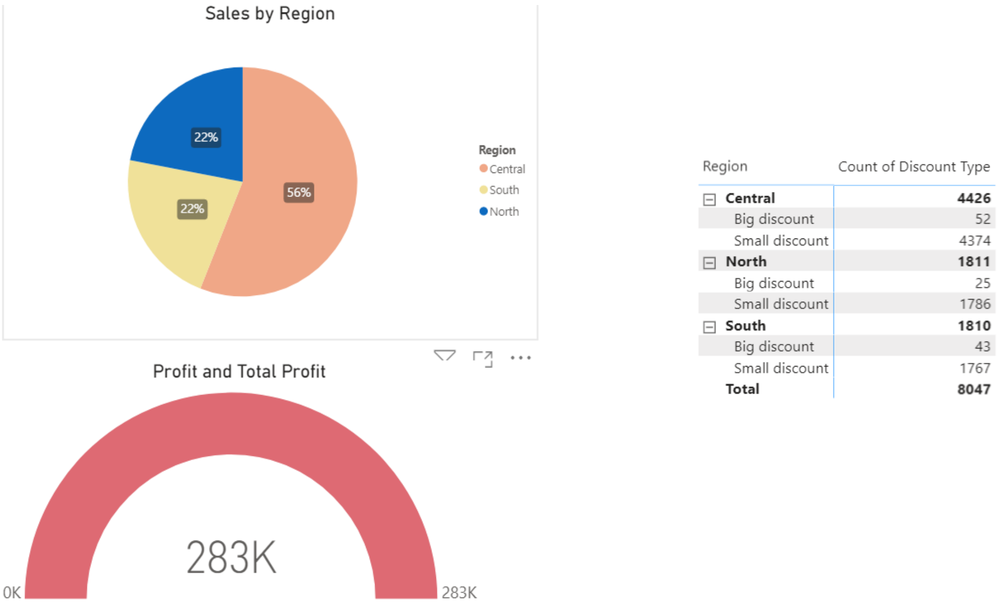


Fig. 4. The connections between the dataset tables

For the descriptive analysis of data, we made some reports and a dashboard that contains different visual elements.

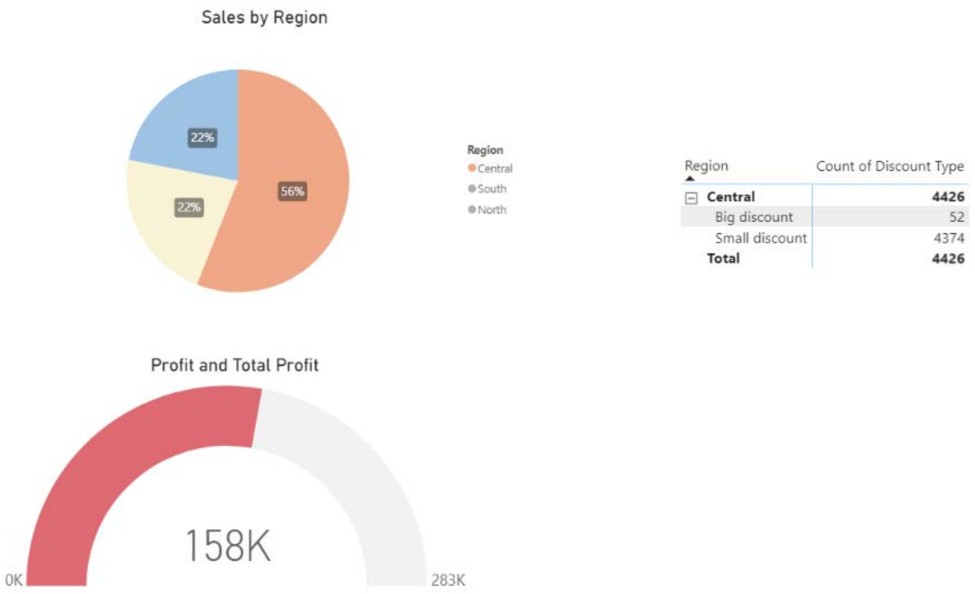
In **Fig. 5**, you can see a report page, which contains a pie chart that shows the sales by region, a gauge for the profit made in every region and the total profit, and a matrix that shows the count of discount types given in every region.



Region	Count of Discount Type
⊟ **Central**	**4426**
Big discount	52
Small discount	4374
⊟ **North**	**1811**
Big discount	25
Small discount	1786
⊟ **South**	**1810**
Big discount	43
Small discount	1767
Total	**8047**

Fig. 5. Power BI report – example 1

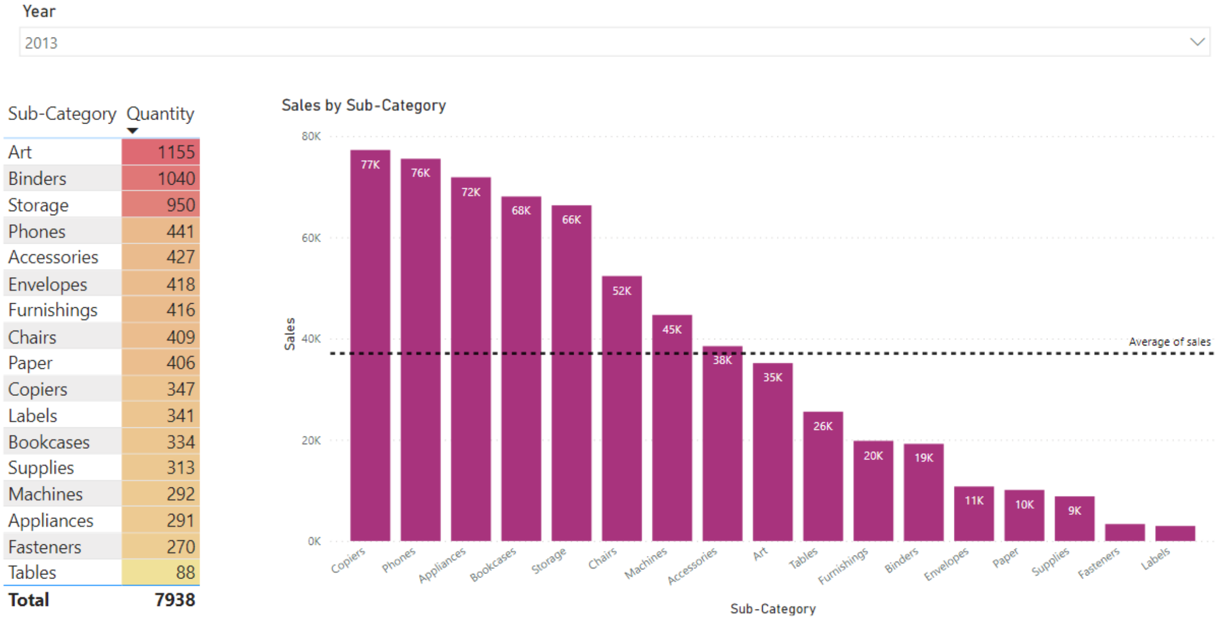
In the report shown previously in **Fig. 5**, Power BI allows users to select a geographic region (Central, South, or North). In this case, the BI tool displays only data about the selected region in all the visual elements, as shown in **Fig. 6**.



Region	Count of Discount Type
⊟ **Central**	**4426**
Big discount	52
Small discount	4374
Total	**4426**

Fig. 6. Power BI report - region selection

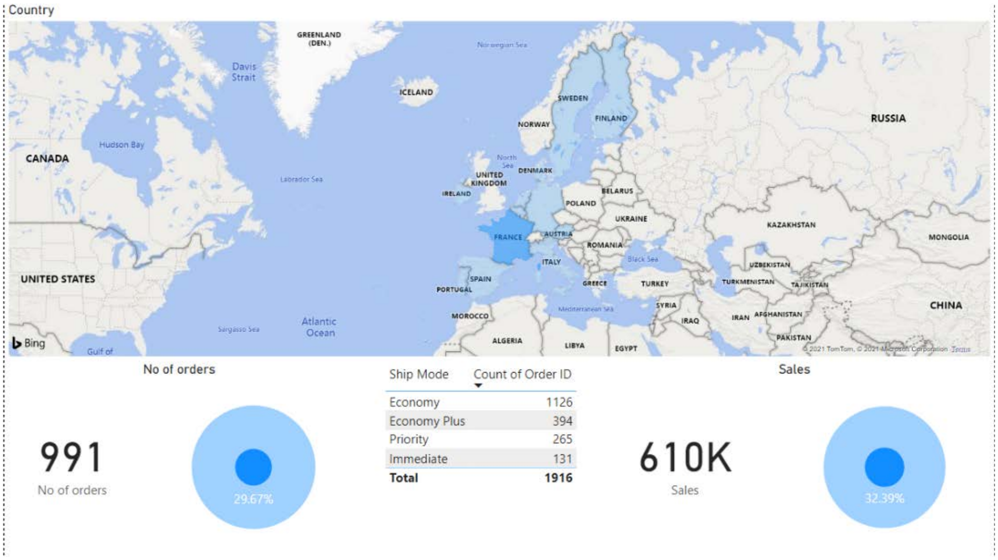
In **Fig. 7**, we can see other aspects of this Business Intelligence solution, which helps in improving the decision-making process. This report includes data on product subcategories. The user can apply filters to the data. In this example, we can apply a filter on the year of order (in **Fig. 7**, we chose the year 2013). The decision-makers receive information about the subcategories of products with the highest quantities sold. They can also get information from the comparison between the sales by subcategory and the average of sales.


Year
2013

Sub-Category	Quantity
Art	1155
Binders	1040
Storage	950
Phones	441
Accessories	427
Envelopes	418
Furnishings	416
Chairs	409
Paper	406
Copiers	347
Labels	341
Bookcases	334
Supplies	313
Machines	292
Appliances	291
Fasteners	270
Tables	88
Total	**7938**

Fig. 7. Power BI report – example 2 – product subcategories

The Filled Map visuals are also easy to use and allow the users to understand the information much better. **Fig. 8** shows a Filled Map visual. In this case, through the map provided, the decision-makers can select an EU country for which they want to see more information regarding aspects such as, number of orders, the revenue from sales, and also what is the preferred delivery type. In **Fig. 8**, we have information about France, which has a number of 991 orders, the sales revenue is 610 thousand euros, and the most popular delivery type is Economy. By selecting other countries on the Filled Map visual, users can see comparisons between the countries, so they can make decisions to improve their company's activity. For example, if managers see a barely used delivery type, they could give up on it.



No of orders

991
No of orders

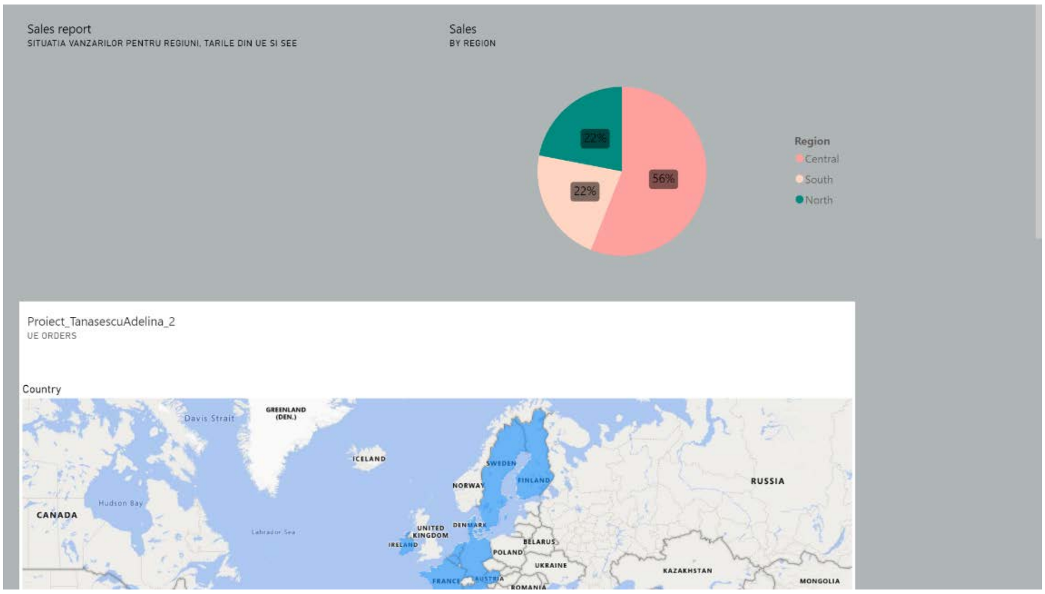
Ship Mode	Count of Order ID
Economy	1126
Economy Plus	394
Priority	265
Immediate	131
Total	**1916**

Sales

610K
Sales

Fig. 8. Power BI report - example 3- orders and sales in France

As we have previously seen, Power BI Desktop offers the possibility of making reports, but with this BI tool, you can also create dashboards. In **Fig. 9**, we have presented a dashboard for this company. The dashboard's theme is sales by geographic and economic regions and contains visuals from the previously mentioned reports.



Sales report
SITUATIA VANZARILOR PENTRU REGIUNI, TARILE DIN UE SI SEE

Sales
BY REGION

Proiect_TanasescuAdelina_2
UE ORDERS

Fig. 9. Power BI dashboard

6. Conclusions

Firstly, through this assessment, we can certainly say that, at the moment, data has an impactful role in the economic process because, through their processing, the decision-makers obtain valuable information. Business Intelligence solutions aim to use data to improve the decision-making process.

Secondly, descriptive analysis is a part of the Business Intelligence process. According to the example given in Chapter 6.2, descriptive analysis plays a very significant role in the decision-making process because it helps a lot with historical data interpretation and comparisons between data at different moments. At the same time, the descriptive analysis creates a broader context for data interpretation and analysis of various factors.

Furthermore, data visualization probably has a major impact on improving the decision-making process. A Business Intelligence solution helps in this matter, and therefore it gives the support to present data from different sources, as they gather in one place. In this way, performing the analysis is much more efficient than in the case of a separate data analysis for every data source.

Finally, Power BI Desktop is a tool that helps to develop Business Intelligence solutions. The most important features are the ability to use many types of data sources and the interactive interface, which is extremely easy to use and allows for easy learning. By using this tool, the decision-makers will receive support for making correct decisions.

To conclude, Business Intelligence solutions help to improve the decision-making process. They provide the space to gather the data from various sources and to analyze it in a broader context. In this way, the company's management can make informed decisions that help in the company's development.

References

[1] "Avantajul Competitiv Al Afacerii: Cum Te Diferențiezi De Concurență," [Online]. Available: https://antreprenoriat101.ro/avantajul-competitiv/.

[2] A.-R. B. Mihaela Muntean, Business Intelligence. Teorie și practică, Bucharest: Editura ASE, 2015.

[3] J. Frankenfield, „Descriptive Analytics," [Interactiv]. Available: https://www.investopedia.com/terms/d/descriptive-analytics.asp.

[4] J. G. Zheng, „Data visualization for Business Intelligence," [Interactiv]. Available: https:// www.researchgate.net/profile/Jack-Zheng-4/publication/321804138_Data_Visualization_for_Business_Intelligence/links/5a3290a9a6fdcc9b2d169738/Data-Visualization-for-Business-Intelligence.pdf.

[5] „What is data visualization?," [Interactiv]. Available: https://powerbi.microsoft.com/en-us/data-visualization/.

[6] „The Top 7 Benefits of Business Intelligence," [Interactiv]. Available: https://www.tableau.com/learn/articles/business-intelligence/enterprise-business-intelligence/benefits.

[7] „Introduction to dashboards for Power BI designers," [Interactiv]. Available: https://docs.microsoft.com/en-us/power-bi/create-reports/service-dashboards.

[8] „Introduction to Power BI for Data Visualization," [Interactiv]. Available: https://www.syntelli.com/introduction-to-power-bi-for-data-visualization.

Adelina TĂNĂSESCU is a graduate of the Faculty of Economic Cybernetics, Statistics and Informatics at the Bucharest University of Economic Studies. She is currently a student at the Databases – Support for Business Master program at the Bucharest University of Economic Studies. Her main fields of interest are databases, data analysis, data warehouses, and Business Intelligence.

Churn Prediction in Telecommunications Sector using Machine Learning

Andreea-Maria COPĂCEANU
The Bucharest University of Economic Studies, Romania
andreea.copaceanu@csie.ase.ro

In these days, due to the increasing competition, churn prediction has gathered greater interest in business, especially in the telecom industry, since gaining new customers is more expensive than retaining the existing ones. The primary objective in telecom churn analysis is to accurately estimate the churn behavior by identifying the customers who are at risk of churning. Another objective is to identify the main reasons for customer churn. This paper focuses on various machine learning algorithms for predicting customer churn, though which we build classification models such as Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine. Prediction performance of the classifiers is evaluated and compared through measures such as Area Under the Curve (AUC), accuracy, and recall rate. Such predictive models have the potential to be used in the telecom industry for making better decisions in customer management.

Key words: *Churn Prediction, Machine Learning, Retention, Telecommunication, Decision Tree*

1 Introduction

In a highly competitive environment, companies must constantly innovate and focus on improving the quality of their services. Customer loyalty represents more than just keeping the right customers, but it represents the key to maximizing profits. Acquiring a new customer is from 5 to 25 times more expensive than retaining an existing one. According to Frederick Reichheld of Bain & Company, inventor of the net promoter score (NPS), increasing customer retention rates by 5% increases profits by 25% to 95%.[1] Considering this, customer retention is one of the main challenges of companies, especially in the telecommunication industry, where customers have multiple options in terms of better and less expensive services. The most often cause of customers churn is due to non-satisfaction in the service offered by a provider or due to more enhanced affordable service by another service provider. [2] In an almost saturated market, companies need a method to identify the customers who are most likely to churn, so that they can build proactive retention campaigns. [3] Thus, an appropriate churn prediction model is essential to predict the customer churn. The proposed model should have the capability to accurately identify customers at risk to churn and then find the reasons behind churning, so as to avoid loss of customers and also propose measures to retain them. The effectiveness of a churn prediction model depends on the learning achieved from the data set provided. An appropriately preprocessed data set gives high performance to the classifiers. Therefore, proper preprocessing is required to remove any redundant or useless features that do not have any relation to the target feature. [4] In general, machine learning techniques are greatly introduced as churn prediction methods. [2] These techniques can help building prediction models in order to discover behaviors and future trends and allow companies to make smart decisions, based on the knowledge extracted from the data. The objective of this study is to investigate existing techniques in machine learning, to evaluate the classifier models for customer churn predictions, and to identify the churn key factors, using data from a telecom public data set. From the experiments, we observed that Random Forest produced better accuracy compared to other machine learning algorithms, while Decision Tree produced a better recall rate compared to other algorithms. We identified the factors behind the customers churning by using the Feature Selection technique. The rest of the paper is structured as follows. In Section 2 we present related work. In Section 3 we present the telecom churn case study. In Section 4 we expose the results. Finally, in Section 4 we conclude the paper.

2. Related work

Churn prediction in telecom companies has been addressed in the literature using various techniques, including machine learning or data mining. These techniques are aimed to assist companies to identify, predict, and retain churning customers.

In [3], the authors presented an advanced data mining methodology which predicts customer churn using a call record data set for 3333 customers with 21 features. The author applied the principal component analysis (PCA) technique to reduce the data dimensionality. Three machine learning algorithms were used for classification: Neural Networks, Support Vector Machine and Bayes Network. The overall accuracy values were 99.10%, 99.55%, and 99.70% for Bayes Networks, Neural networks, and Support Vector Machine, respectively.

Different machine learning algorithms were proposed in [5], through which different models were employed, such as Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting Trees, which are used to predict churn customers. The authors used AUC to measure the performance of the models. According to the AUC values, the method that gave the most accurate model was Gradient Boosting with AUC value of 84.57%.

The author presented in [6] a comparison study using three classifiers K-NN, Random Forest, and XG boost, respectively. The XG boost classifier performed the best, compared to the KNN and RF classifiers, in terms of accuracy score and F score.

A churn prediction model was proposed in [7], as well as clustering techniques to identify the churn customers. The proposed model used classification algorithms, in which the Random Forest algorithm performed the best with 88.63% correctly classified instances. Furthermore, the study also provided factors behind customer churn using the Attribute Selected Classifier algorithm.

In [8], the authors studied the problem of customer churn in a big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the churn depending on the volume, variety, and velocity of the data. The Random Forest algorithm was used and evaluated using AUC.

The authors proposed in [9] machine learning algorithms on a big data platform in order to predict the customer churn. The performance of the model was measured by Area Under Curve (AUC), obtaining a value of 93.3%. The model was built and tested through Spark environment, using a large data set provided by SyriaTel telecom company. The model experimented four algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree “GBM” and Extreme Gradient Boosting “XGBOOST”. The best results were obtained for the XGBOOST algorithm.

A classification model for churn prediction based on the Rough Set Theory in telecom was proposed in [10]. The proposed Rough Set classification model outperformed the other models like Linear Regression, Decision Tree, and Voted Perception.

The author implemented in [11] a classification model based on Logistic Regression and used statistical methods for feature selection. The experimental results showed that by increasing the threshold values and selecting the right features with different combinations, the model will deliver better results in the churn prediction process.

3. Churn prediction case study using Machine Learning

Based on these concepts, a customer churn prediction case study will be further presented. A telecom data set is going to be used so that to predict churn using machine learning algorithms and detect the main factors that may lead the customer to switch to another telecom provider.

3.1 Data preprocessing

We are going to use a public telecom data set on churn. This data set contains various information about customers including customer care service details, customer value-added services, customer personal details, customer usage pattern and customer bill and payment details.

The data set contains 58 attributes and 51,047 observations, which indicate whether the customer had left the telecom provider or not.

Part of these features that are not relevant for churn prediction, will be removed from the data set. In this regard, the feature CustomerID, which contains numeric identifier of a customer, is not relevant in the analysis, and therefore it is removed.

In the same manner, feature NotNewCellphoneUser is removed, as it contains the opposite values of feature NewCellphoneUser, the latter will be retained in the data set.

Feature ServiceArea is also removed from the data set, as it contains alphanumeric values representing the service area of the telecom company and it is irrelevant for churn prediction. Also, the feature Homeownership is removed from data set.

There are 15 features in the data set containing missing values, as shown in Table 1. The feature HandsetPrice is dropped from the data set, as it contains more than 50% of missing values.

For the rest of the features in the table, missing values are replaced with the mean value of the entire feature column, using the fillna() method.

Table 1. Features having missing or “Unknown” value

No.	Variable	Missing Values
1	MonthlyRevenue	156
2	MonthlyMinutes	156
3	TotalRecurringCharge	156
4	DirectorAssistedCalls	156
5	OverageMinutes	156
6	RoamingCalls	156
7	PercChangeMinutes	367
8	PercChangeRevenues	367
9	ServiceArea	24
10	Handsets	1
11	HandsetModels	1
12	CurrentEquipmentDays	1
13	AgeHH1	909
14	AgeHH2	909
15	HandsetPrice	28982

Categorical variables can hide important information in the data set. Variable values with 2 categories will be converted to numbers using Python map() function, and the other 4 variables with more than 2 categories CreditRating, PrizmCode, Occupation and MaritalStatus, will be transformed using OneHotEncoder function of SciKit package, which maps a column of category indices to a column of binary vectors.

3.2 Data visualization

When a feature is analyzed independently, we are usually mostly interested in the distribution of its values. Churn is the target variable, and it is binary: Yes indicates that that the company lost this customer, and No indicates that the customer was retained. In the data set, 28.6% (14,257) customers are churners, whereas 71.5% (35,519) customers are non-churners, as shown in **Fig. 1.**

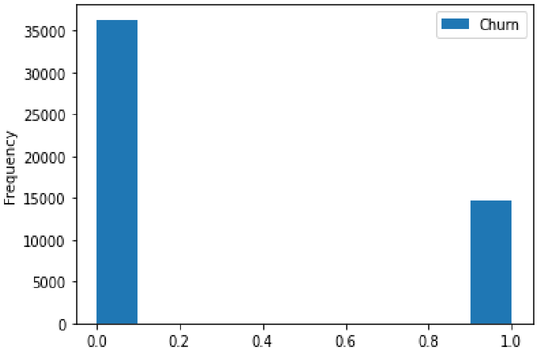


Fig. 1. Churn distribution

Bivariate analysis finds out the relationship between two variables. In this analysis, the distribution of categorical variables according to the Churn variable is plotted, as shown in **Fig. 2**. This shows us a skewed distribution for most part of the variables. Out of 21 variables, 17 variables have two values, and the remaining variables have three or more values.

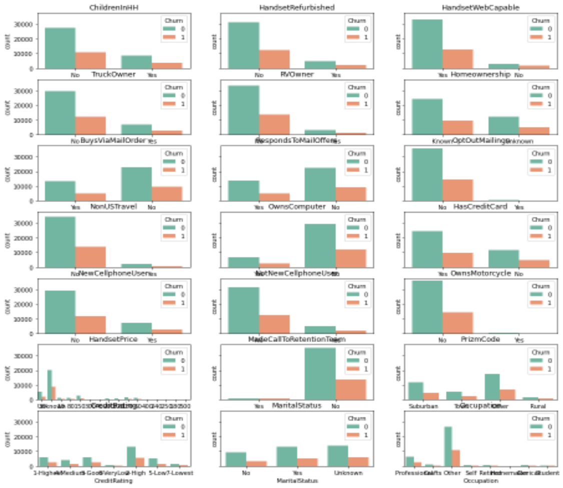


Fig. 2. Categorical variables and Churn Variable

Pearson correlation coefficients of numerical variables are shown in the correlation matrix in **Fig. 3**.

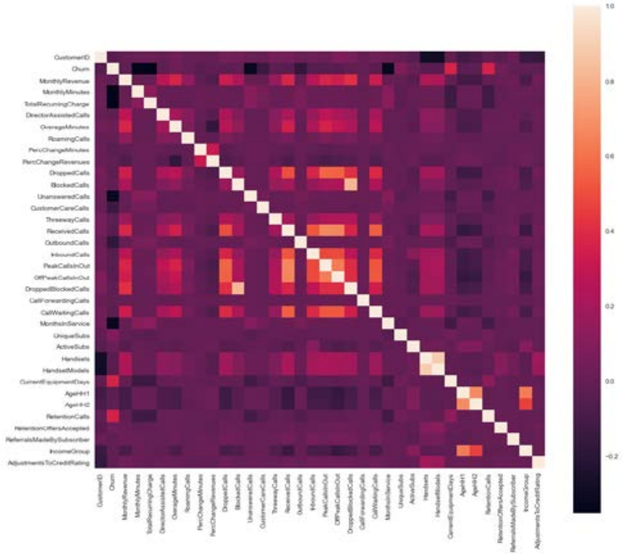


Fig. 3. Correlation matrix

Variables CurrentEquipmentDays, RetentionCalls, MonthlyMinutes, TotalRecurringCharge, UnansweredCalls, OutboundCalls, and MonthsInService are highly correlated with Churn predictor variable. (**Fig. 3**) Using the sklearn feature selection method, the most important variables to predict the churn are obtained and shown in **Table 2**.

Table 2. Feature importance score obtained using Feature Selection

No.	Feature name	F Scores
1	CurrentEquipmentDays	9,936.10
2	TotalRecurringCharge	7,251.70
3	MonthlyMinutes	7,027.36
4	MonthsInService	5,982.62
5	UnansweredCalls	5,768.01
6	RetentionCalls	5,233.20
7	OutboundCalls	1,998.18
8	CustomerCareCalls	882.28
9	MadeCallToRetentionTeam	625.62
10	UniqueSubs	512.07
11	CreditRating_5-Low	164.18
12	PercChangeMinutes	101.38
13	OffPeakCallsInOut	82.09
14	PeakCallsInOut	75.66
15	InboundCalls	72.94

The most highly correlated feature that predicts customer churn based on feature selection (**Table 2**) is CurrentEquipmentDays, which means that the longer a customer has the current equipment for, the more likely he is going to churn. This makes sense when we consider that a customer with an outdated equipment is likely to be looking for a new one. When looking for a new equipment, a customer without brand loyalty will consider several brands for it, thus increasing the chances that he will switch to another provider. Also, the fact that a customer has the same equipment for a long time without upgrading might be a signal that he is not using the service frequently, in which case it would also make sense that the customer churns.

3.3 Software tools

This study will use Python, which is one of the mainstream languages in data science and machine learning, providing libraries for data visualization, preparation, and machine learning tasks. The main packages in Python used for data analysis and machine learning are:

- **Pandas** is one of the main tools used by data analysts, for multiple phases of data science workflow, including data cleaning, visualization, and exploratory data analysis.
- **NumPy** is the fundamental package for numerical computation in Python and provides support for large multidimensional array objects.
- **Scikit-learn** is a machine learning library that implements a range of machine learning, preprocessing, cross-validation, and visualization algorithms.
- **SciPy** provides various numerical tools, such as interpolation, integration, optimization, image processing, statistics, special functions.
- **Matplotlib** is a data visualization library for creating static, animated, and interactive visualizations in Python.

3.4 Machine Learning algorithms for Churn analysis

For churn prediction case study, we are going to use Logistic Regression, Random Forest, Support vector machine, and Decision Tree machine learning algorithms. After applying the algorithms, the models will be evaluated in order to decide which model is the best to be used for prediction.

The models have been evaluated using the holdout method, which divided the given data into two independent data sets, 80% for training and 20% for testing. After applying the chosen algorithm to the data set, we will obtain a model that will be evaluated based on performance indicators.

The first algorithm is Logistic Regression and we have a model accuracy of 87,52%, with a precision of 80,82% and a recall of 73,41%. For the second algorithm, Random Forest, we got a classification rate of 95,12%, considered a very good accuracy, with a precision of 92,39% and a recall of 90,25%. The Decision Tree algorithm shows an accuracy of 94,36%, with a precision of 89,55%, and a recall of 90,7%. The last algorithm Support Vector Machine has an accuracy of 87,47%, with a precision of 80,25% and a recall of 74,04%.

The classification report for the four models is summarized in Table 3 and shows the main classification metrics accuracy, precision, and recall for each model. The report is used to measure the quality of predictions for a classification algorithm.

Table 3. Classification report results

Model	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.8752	0.8082	0.7341
Random Forest	0.9512	0.9239	0.9025
Decision Tree	0.9436	0.8955	0.9070
Support Vector Machine	0.8747	0.8025	0.7404

In order to choose the best model that will predict if the customer is going to churn, we want to maximize the recall, as it represents the ability of the classifier to correctly predict all truly positive cases.

Considering this, the Decision Tree model is the best model to be used in order to predict new values for possible customers that will churn. We can see below the confusion matrix for decision tree, which gives a holistic view of the performance of the model.

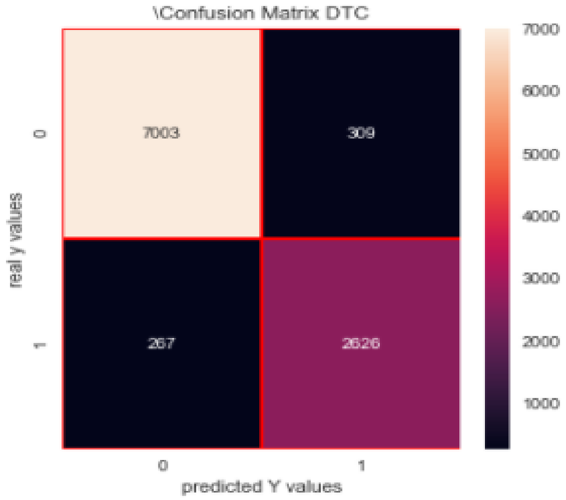
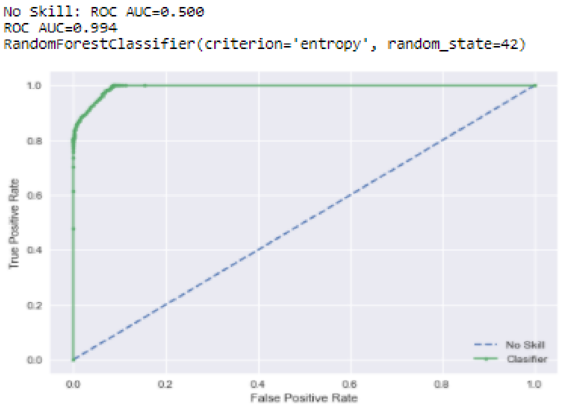


Fig. 4. Confusion Matrix Decision Tree

So, from the total number of customers, we can see that most of them were predicted correctly for churn. 7003 customers have no for churn and were also predicted with no, 2626 have yes for churn and were predicted with yes. Further, 309 were incorrectly classified as churners, when they are not and 267 were predicted as non-churners, when they are going to churn. (**Fig. 4**)

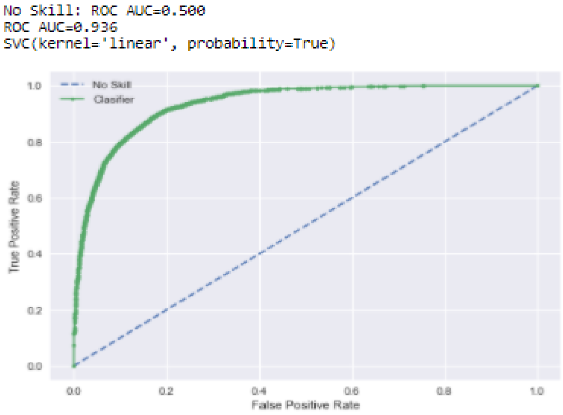
The ROC curve for each model is shown in figures 5-8 according to Table 3, for each model output. The ROC curve shows the trade-off between the true positive rate (TRP) and the false positive rate (FPR). Given a test set and a model, TPR is the proportion of positive (churned) tuples that are correctly labeled by the model. FPR is the proportion of negative (no churn) tuples that are mislabeled as positive.

The AUC values have been plotted for the models as shown in Figures 5-8, where the best result for AUC is 0,994, obtained for Random Forest. (**Fig. 5**)



```
No Skill: ROC AUC=0.500
ROC AUC=0.994
RandomForestClassifier(criterion='entropy', random_state=42)
```

Fig. 5. Random Forest AUC



```
No Skill: ROC AUC=0.500
ROC AUC=0.936
SVC(kernel='linear', probability=True)
```

Fig.7. Logistic Regression AUC

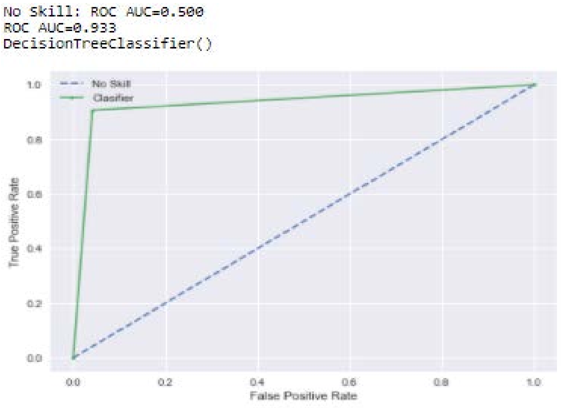
4. Results

In this case study, we used a public telecom data set, targeting churning customers. The data set contains details about customers including customer care service, customer value added services, customer personal details, and customer usage pattern. For these customers, we observed that those with lower values of total recurring charge, monthly minutes of use, months in service, number of unanswered calls, or number of outbound calls and high values for number of days of the current equipment, are very likely to churn.

In order to avoid this, a classification machine learning algorithm is being used, to predict the customers that are exposed to churn. We have tested four algorithms and have chosen the one that presented the minimal false negative rate, which is Decision Tree. Using this model, the company can predict those customers that are going to churn and take the necessary actions in order to prevent them from churning.

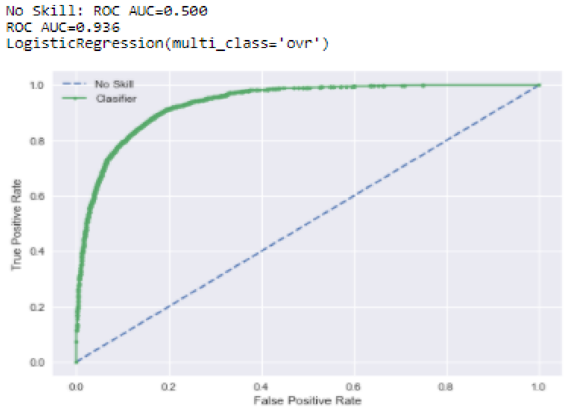
Based on our analysis, a solution might be to include an incentive plan to offer a discounted new equipment to those customers who have the equipment for a long time.

This is based on the high correlation of the number of days of the current equipment with churn, which suggests that churn is more likely when the customer has the same equipment for a long time.



```
No Skill: ROC AUC=0.500
ROC AUC=0.933
DecisionTreeClassifier()
```

Fig. 6. Decision Tree AUC



```
No Skill: ROC AUC=0.500
ROC AUC=0.936
LogisticRegression(multi_class='ovr')
```

Fig. 8. Support Vector Machine AUC

5. Conclusion

The paper aims to find the most accurate model for churn prediction in telecom and, at the same time, to detect the key factors that might lead customers to churn.

For this purpose, four classification algorithms, Logistic Regression, Random Forest, Support Vector Machine and Decision Tree, have been implemented and analyzed in Python.

The models' performance has been measured by recall value, since the goal in this case is to predict as accurate as possible the customers that are going to churn. The best recall value was 90,7%, obtained for the Decision Tree model. This aspect of the analysis suggest that Decision Tree is the best model to be used in the churn case, in order to predict those customers that are most probable to churn. Another evaluation metric in place, that can be used to measure the models' performance is area under the ROC curve, where the best AUC is 0.994, obtained for the Random Forest model.

The analysis leads to the conclusion that telecom operators can obtain the best predictive models to predict churn, by analyzing historical customer records and further understanding customers behavior. Based on this, decision-making employees can build different marketing approaches to retain churners based on the predictors that have higher importance in scoring the model performance.

References

[1] Gallo, A. (2014, October 29). TheValueof Keeping the Right Customers.
Retrieved from Harvard Business Review: https://hbr.org/2014/10/the-valueofkeeping-the-right-customers;

[2] Ahmed, A., & Linen, D. M. (2017). A review and analysis of churn prediction methods for customer retention in telecomindustries. International Conference onAdvanced Computing and CommunicationSystems (ICACCS). IEEE;

[3] Brandusoiu, I.(2016). Churn Predictionin the Telecommunications Sector UsingSupport Vector Machine. *Annals of theOradea University: Fascicle Managementand Technological Engineering*, pp. 97100;

[4] Mishra, K., & Rani, R. (2017). Churnprediction in telecommunication usingmachine learning. *International Conferenceon Energy, Communication, Data Analyticsand Soft Computing (ICECDS).* IEEE;

[5] Gaur, A., & Dubey, R. (2018). PredictingCustomer Churn Prediction In TelecomSector Using Various Machine LearningTechniques. *International Conference onAdvanced Computation andTelecommunication (ICACAT).* IEEE;

[6] Pamina, J. B. An effective classifier forpredicting churn in telecommunication. *Jourof Adv Research in Dynamical & ControlSystems, 11*;

[7] Ullah, I. R. (2019). A churn predictionmodel using random forest: analysis ofmachine learning techniques for churnprediction and factor identification intelecom sector. (pp. 60134-60149). IEEE;

[8] Huang, Y. Z. (2015). Telco churn prediction with big data. *ACM SIGMOD international conference on management of data*, (pp. 607-618);

[9] Ahmad, A. J. (2019). Customer churnprediction in telecom using machine learningin big data platform. *Journal of Big Data,6(1)*, pp.1-24;

[10] Makhtar, M. N. (2017). Churnclassification model for localtelecommunication company based on roughset theory. *Journal of Fundamental andApplied Sciences, 9(6S)*, pp. 854-868;

[11] Sai, B. a. (2019). Predictive Analysisand Modeling of Customer Churn inTelecom using Machine LearningTechnique. *3rd International Conference onTrends in Electronics and Informatics(ICOEI)* (pp. pp. 6-11). IEEE;

[12] Brandusoiu. (2016). Methods for churnprediction in the prepaid mobiletelecommunications industry. *InternationalConference on Communications (COMM)*(pp. 97-100). IEEE;

[13] Dalvi, P. K. (2016). Analysis ofcustomer churn prediction in telecomindustry using decision trees and logisticregression. *Symposium on Colossal DataAnalysis and Networking (CDAN)* (pp. 14). IEEE;

[14] Reichheld, F. (2011, October).*Prescription for Cutting Costs. Bain &Company.* Retrieved from Bain &Company:https://media.bain.com/Images/B_Prescription_cutting_costs.pdf;

[15] Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A. and Kanade, V. A. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. Symposium on Colossal Data Analysis and Networking (CDAN), 2016, pp. 1-4.

Andreea-Maria COPĂCEANU (b. August 7, 1992) has graduated the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies, in 2014. She followed a master's degree in Cybernetics and Quantitative Economics, within the same faculty. Currently, she is a PhD student, teaching assistant in the department of the Faculty of Cybernetics, Statistics and Economic Informatics and works in IT as Cloud Engineer in the Support Department. She has high interest in areas such as Data Science, Analytics, Machine Learning, Databases, Big Data, Business Architectures, and Strategic Marketing.

Differentially Private Data Release for Data Analytics - A Model Review

Peter N. MUTURI[1], Andrew M. KAHONGE[1], Christopher K. CHEPKEN[1]
[1]School of Computing and Informatics,
University of Nairobi,
Nairobi, Kenya.
pmuturi@mmu.ac.ke, andrew.mwaura@uonbi.ac.ke, chepken@uonbi.ac.ke

To leverage on the potential of data analytics, enabling private data release is needed. The challenge in achieving private data release has been balancing between privacy and analytical utility. Among the models that seek to solve the challenge, ε-differential privacy promises to achieve the balance by regulating the epsilon (ε) value. The choice of the appropriate epsilon value that achieves the balance has been a challenge, making the ε-differential privacy not practically applicable by many. A practical and heuristic method to estimate this privacy parameter needs formulation. The variable to estimate appropriate privacy parameter that is not provided in heuristic manner is the reidentification probability. Previous research has based that probability on released data sets and linkage data sets, with less focus on data analysts. This paper proposes a causal relationship model for estimating the reidentification probability, which adds the analyst's aspect to the model.

Keywords: *Privacy, Data Utility, Differential Privacy, Big Data, Private release, Anonymization*

1 Introduction

With the convergence and working together of smart devices, the Internet of Things, and Internet-based applications, massive data can be produced, collected, processed, and stored effectively. However, storing large volumes of data without making value from it is not helpful, and, indeed, it is a waste of computing resources [1]. The value of the data sets held is achieved through data analytics [2], which may be undertaken by the owners of these data sets (curators), or the curators may release the data sets, which are then used by third parties for various analytical purposes. One concern that needs to be addressed to actualize the data release for data analytics is how the private information contained in the data sets should be protected [3].

Without data collecting agencies (curators) that guarantee the protection of private information held in the data sets that they store, only a few individuals would willingly participate in any data collection exercise, and those who do participate may not provide very accurate data [4]. As observed by Cavoukian and Reed in [5], the challenge of safeguarding privacy threatens the willingness to release data and information. However, with a mechanism that provides the guarantee that the privacy of individuals in the data sets is in place, thereby creating anonymous data sets, then a framework that allows data curators to make such data available to third parties or even to the public for analytical purposes can be put in place. This is what is known as private data release. Analysts who interact with such private data can only learn about the population from which the data was obtained, but not about an individual whose data is in the dataset.

The process of making data private involves suppression, aggregation, noise addition, swapping, among other mechanisms [6]–[8]. This, in effect, affects the analytical value of the data, hence reducing the data's analytical utility. If ensuring privacy of the data sets was the only goal, this would be achieved trivially [9]. However, in pursuing high levels of privacy through various mechanisms, data sets may end up losing the analytical utility that is very core for data analytics. On the other hand, to have high levels of analytical data utility, data should not be changed much from its full disclosure form, which in effect makes privacy breach very likely. Data privacy and analytical utility are therefore inversely proportional, requiring delicate balancing act for any private data release aimed at supporting data analytics to achieve its goal [6].

The differential privacy model, a private data release approach that promises to deliver the balance between privacy and analytical utility, is expressed in a theoretical mathematical format, which is not utilitarian. This has made the model face an implementation challenge that needs to be resolved to allow its adoption and wide-spread application [10][7]. In particular, the choice of an appropriate privacy parameter, the epsilon value, that guarantees the privacy without eroding analytical utility has not been utilitarian in a manner that enables software developer be able to implement the model. There is a need for a practical and heuristic approach on how to arrive at the right value. This is what this model review paper aims to achieve.

This paper is a narrative model review that seeks to operationalize the differential privacy by making the choice of privacy parameter become practical and heuristic, hence making it utilitarian. The method used was to synthesize the available literature that was obtained from journals and other academic materials sourced through online searching. A gap that needs to be filled was then identified and a solution to it is provided.

2. Data analytics & data release

Data analytics allows the examination of data sets with a view of extracting useful information by identifying and analyzing behavior and patterns using both qualitative and quantitative techniques [11]. This, however, poses a threat of disclosure of private information about individuals whose data is in the data sets.

If the data sets are used by the curator for analysis, privacy concerns would not exist. However, the curators may wish to a release the data to third party analyst or to the public, who may perform secondary data analytics on the data and the analyst may need to link to other data sets from other sources for the process of analytics to be successful. Data analytics may call for the need to interact with various data sets, in order to attain the hidden patterns and relationships among the data sets. It is important for curators to know which data to release to a third party or even to the general public to enable further analysis using the dataset.

To provide a conducive environment for data analytics, there is need to enable private data release – releasing anonymized data, whose individuals who are the subject matter, are protected from disclosure to unauthorized parties. Such data should retain analytical utility to allow analysts to draw some insights from them. The released data should not have associations with the individuals, who are the data subject [12]. In this way, the data released protects the privacy of the individuals but retains analytical utility.

Private data release is necessary for both privacy preserving data publishing (PPDP) and privacy preserving data mining/analysis (PPDM/A). In PPDP, the aim is to provide the public with deidentified or synthetic data for further investigation. The purpose of PPDP influences the kind of data to be published. If the intention is just to inform, without further analysis expected, then contingency tables – a matrix format showing statistics of one variable in a row and those of another in a column, and histograms – a plot showing the frequency distribution of data, may be used [13] - [17]. However, publishing data intended to be used for analytical purposes needs to retain reasonable analytical utility. In PPDM/A, the data is not released to the analyst, instead, they are allowed to interact with the data set through aggregated queries [10], [18].

3. Privacy and analytical utility

The need for a mechanism that would enable data release that supports data analytics in an environment that guarantees privacy of the individuals whose data is held without sacrificing its analytical utility is the drive behind numerous research work in the area.

Some of the disclosure control mechanisms that have been used to limit privacy loss, such as the anonymization-based techniques, have been shown to diminish the analytical utility of the data due to the alterations made in attempt to mask the individuals in the data sets [18].

To achieve private data release that supports data analytics, it is necessary to balance the two competing goals: privacy of individuals and analytical utility of the data sets [17], [18]. The two are antagonistic in that very private data will not be of much analytical use (utility), while high data analytical utility implies high accuracy, which is likely to cause a privacy breach [19]. Therefore, the choice of an anonymization mechanism to be used in enhancing private data release must be done with a good trade-off between the two goals [12]. We highlight the two main categorizations of the privacy mechanisms used, namely anonymization and differentially privacy models.

4. Anonymization privacy models

A typical data set has three main types of attributes that describe the subject (individual). 1) Explicit Identifiers (EID) – attributes whose values uniquely (directly) identify an individual in the data set. Such includes name, national identification card number, etc. 2) Quasi Identifiers (QID) – attributes whose values on their own may not be able to identify an individual, but when combined with values of other QID, have potential to identify and individual. Examples include gender, age, etc. 3) Sensitive Attributes (SA) – attributes whose values are confidential in nature and individuals in the data set would be uncomfortable if revealed or associated with them. Such includes income, ailment diagnosis, etc. [20].

Any anonymization (de-identification) technique used must remove, hide, or suppress all the explicit identifiers to make sure individuals are not revealed. However, the QID and SA should remain.[18], [21]

K-Anonymity is one of the techniques used for de-identification where a group of records of the dataset with same attribute values is referred to an equivalence class. This technique requires that each equivalence class in the data set, has at least k (a constant number) members, meaning each member of the equivalence class has k-1 other elements that cannot be distinguished from it. The value of k is a constant whole number, i.e., number of records in a given equivalence class. The technique is known to protect against identity disclosure – being able to identify a record in the dataset, but not attribute disclosure – where, from the attribute values, one is able to learn about a group of records. K-Anonymity is further demonstrated to be susceptible to homogeneity attack and background knowledge attack [22]–[24].

The ℓ-diversity model, which is an improvement of k-anonymity, requires that the values of the sensitive attributes in each equivalence class have at least ℓ (a constant number) well-represented values. This means that the values for the sensitive attributes of a given equivalence class are such that there are ℓ indistinguishable records, where ℓ is greater or equal to two (i.e., $\ell \geq 2$). The model is reported to be prone to skewness attack and similarity attack [22], [23].

The t-closeness model improved the ℓ-diversity model by requiring that the distance between the distribution of sensitive attributes in a class is not more than a threshold t [22].

The three models and their affiliates are commonly referred to as anonymization models, and are known to lack mechanism to deal with background information the analyst may have, hence unable to provide guarantee of privacy. Another concern of these models is the loss of analytical utility of the data that is caused by alterations of the original data values in order to achieve the masking aspect [6], [21], [25]. This means the two goals, privacy & utility, sought in private data release may not be achieved using these models. This makes the models unsuitable for private data release.

5 Differential privacy model

This is the privacy model that promises to achieve both the goals of privacy and utility of data sets. Indeed, it has become the de facto model in private data release [25]. The model requires that the probability distribution in the released results area essentially remains the same, irrespective of whether an individual's data is included in the dataset or not. This way, the presence or absence of an individual in the data set does not influence the result of the analysis. This ensures that the analyst does not learn about an individual in the dataset, but is able to learn about the population represented by the members in the data set [26]–[28].

In particular, the ε-differential privacy (ε-DP) provides a provable and quantifiable privacy guarantee, as well as a trade-off between the privacy and data utility. An algorithm, in the context of analytics, a query, is said to satisfy ε-DP, if and only if, the difference in probability (Pr) of any query outcome (S) of two data sets (D1 & D2), which differ only by a maximum of one entry, only varies by a factor of exponential (e) to the power of epsilon (ε) [29], [30]. Formally, a randomized algorithm M, is said to provide ε - differential privacy if for all data sets D1 and D2 differing in not more than one record, and all S ⊆ Range(M), then, equation (1) below applies [26].

$$Pr[M(D1) \in S] \leq e^{\varepsilon} * Pr[M(D2) \in S] \quad (1)$$

The ε-DP mechanism ensures that the computational result of dataset does not change significantly due to the inclusion or exclusion of an individual in the dataset. This is achieved by adding carefully calibrated noise to the true results, making the output insensitive to changes in individual record [15], [31].

One important concept for guaranteeing ε-DP is mechanism sensitivity, denoted as Δf, that measures the maximum change in output of a mechanism as a result of change in individual record. The literature has shown that the sensitivity and the epsilon (ε) value determine the noise to be added for a mechanism to satisfy ε-DP. When dealing with real numbers, Dwork et al. [16] proved that noise from the Laplace mechanism with scale of Δf/ε would satisfy ε-DP, while when using integers [32], Ghosh et al. [33] proved that noise from the geometric mechanism with scale of ε/Δf would do the same [15], [34]. Their views are widely supported in the literature [35]. Mechanism sensitivity (Δf) can easily be computed from the data set but there is not much published work on how the value of epsilon (ε) is obtained.

The effectiveness of ε-DP approach is, therefore, very much dependent on the choice of epsilon (ε) value, which is the privacy parameter, also called the privacy budget. The privacy parameter (or budget) controls the trade-off between the privacy guarantee and the data analytical utility. Small epsilon values lead to higher privacy due to more noise added to mask the data, but it also implies less accurate data, hence low utility. Large epsilon values lead to less noise added, meaning high accuracy; hence high utility, but the individuals are at high risk of re-identification [36]. This follows the fundamental law of information recovery, which indicates that very accurate answers to many questions destroy privacy in a big way [27], [37]. Therefore, getting the right value of epsilon (ε) is an important aspect in operationalizing the differential privacy [30].

Despite the promise of ε-DP achieving the two antagonistic goals (privacy & utility) necessary in private data release, its usage/application is reported to be very low [18]. This is attributed to its theoretical mathematical expression that is not easily implemented [10] and, in particular, to its description of the privacy parameter that is not in a utilitarian format that is readily applicable [38]. Derivation of the privacy parameter (the epsilon value), is not heuristic – i.e., not self-explanatory, meaning not everyone can derive it, for a given dataset.

Differential privacy model has been ascertained to be the one that can give the much-sought balance between privacy and data utility in the private data release. Its practical application has, however, been found to be limited, despite its promising potential. The comprehension and interpretation of its theoretical mathematical formulation that would lead to wide application have been stifled by the challenge of establishing the privacy parameter, the epsilon (ε), which is the guarantee of the privacy being provided by the mechanism.

For ε-DP to get widespread application, a practical and heuristic way of determining the privacy parameter needs to be provided and proven empirically, for software developers to know how to apply it. We take a look at attempts that have been made in trying to arrive at the appropriate value of this privacy parameter.

6. Choice of privacy parameter in differential privacy

Choosing the value of epsilon that satisfies ε-DP has been reported not to be a trivial matter; however, there is dearth of research on how to determine it. In some research works [30], the value is picked without explanation of how it was arrived at, or is simply assumed to be a certain value. If an empirical and heuristic method of determining the value is provided, implementation of ε-DP is likely to be embraced and widely used [30], [36].

Two methods for determining the appropriate value of the epsilon that were found in the literature differ in their approach significantly. One by Hsu et al. [30], views the epsilon (ε) as a factor of: 1) Budget of conducting the study (B), 2) Target accuracy or error margin (T), 3) Expected cost of individual participating or not participating in the study (E) and 4) probability error or confidence measure (α). They proposed a formula to obtain the appropriate privacy parameter as the equation (2) below.

$$\varepsilon \leq \ln\left[1 + \frac{BT^2}{12E\,\ln\frac{3}{\alpha}}\right] \qquad (2)$$

The authors reported that the expected benefits of the participants they studied were in monetary form, which made them quantifiable. Lee and Clifton [36] work used a mathematical formulation approach in their coming up with the formula of computing the value of privacy parameter, the epsilon value. In so doing, they were able to assume certain values of probability and used them to prove their formula. Their approach was theoretical in nature. However, a practical approach method that is validated empirically and is heuristic is needed. A heuristic method would enable analysts and system developers to apply the method to get the probability of re-identification, which is then applied to compute the appropriate privacy parameter for a given dataset on their own.

This probability of re-identification depends on factors that are intrinsic to the dataset as well as external factors that vary from one region to another. The intrinsic factors are 1) the uniqueness – characterizes the amount of unique elements in the dataset, and 2) distinguishing power of each attribute [4], [39]. The external factors are 1) the technical skills and resources available to the analyst and 2) the availability of linkage data sets that can be linked to the anonymous data sets [18]. There is need therefore to model how to arrive at the re-identification considering these factors. They made two fundamental assumptions that made it possible to operate: 1) that participants were afraid of some bad events and 2) that they were able to estimate their expected cost of these bad events [30].

The challenge with this model is that the parameters must be established at the point of data collection. That is, an analyst who gets data sets that were collected without those parameters stated may not have a way of determining the epsilon value, hence not able to implement the model. Therefore, we observe that this approach may be applicable in some circumstances but not in all situations.

The second method is by Lee and Clifton [36], which views the epsilon (ε) as a factor of: 1) Global sensitivity (Δf), 2) Maximum distance between possible solutions (Δv), 3) Size of data set (n) and 4) Probability of being identified (p). They proposed a formula of getting the appropriate privacy parameter as the equation (3) below.

$$\varepsilon \leq \frac{\Delta f}{\Delta v} \ln \frac{(n-1)p}{1-p} \qquad (3)$$

The sensitivity, maximum distance, and dataset size are inherent in the dataset and can be computed or read from the dataset for any given dataset. Once the probability of identification is established, the privacy parameter ε will be known. We find this approach applicable by the analysts, on data sets they collect and those collected by others. Lee and Clifton [36] concluded that a mechanism for establishing the probability of identification is important in making determination of appropriate value of the privacy parameter ε.

7. Modelling reidentification probability

Getting an appropriate estimate of the re-identification probability is critical in computing the appropriate value of epsilon, which is the privacy parameter (the budget). The rest of variables for computing epsilon (ε), i.e. sensitivity, maximum distance, and dataset size, are computed directly from the dataset. The privacy parameter regulates the trade-off between the privacy and utility, as well as the amount of noise to be added so that the output results satisfy ε-DP.

Following factors identified as influencing re-identification [4], [18], [39], the re-identification probability can only be estimated for a given region and for a specific time. This is because factors such as the analytical skills of the analysts and the resources available to them, that were identified as influencing re-identification, will vary from region to another. The same applies to the availability of linkage data sets to the analysts to whom they can refer.

The reviewed literature identified each of the factors influencing re-identification as factor on its own. It is our considered view that the factors do not work in isolation, but instead work together in contributing to re-identification. Therefore, it is necessary to examine their combined cause and effect, and that is what informed the formulation of the proposed model in Fig. 1.

The causal relationship model that is postulated to influence the re-identification probability was arrived at in consideration of the factors stated above. The identified factors are latent variables that need measurement indicators. The appropriate indicators for each factor were identified and, therefore, represent the model as shown in Fig. 1.

8. The proposed model

Fig. 1. represents the model that can be used to estimate the re-identification probability of a given region. The model adopted the Structural Equation Modelling (SEM) for its ability to work with latent variables, also known as a construct. A construct is a representation of factor that cannot be measured directly; instead, its indicators are used to measure it.

There are four constructs that form the structural or inner model.

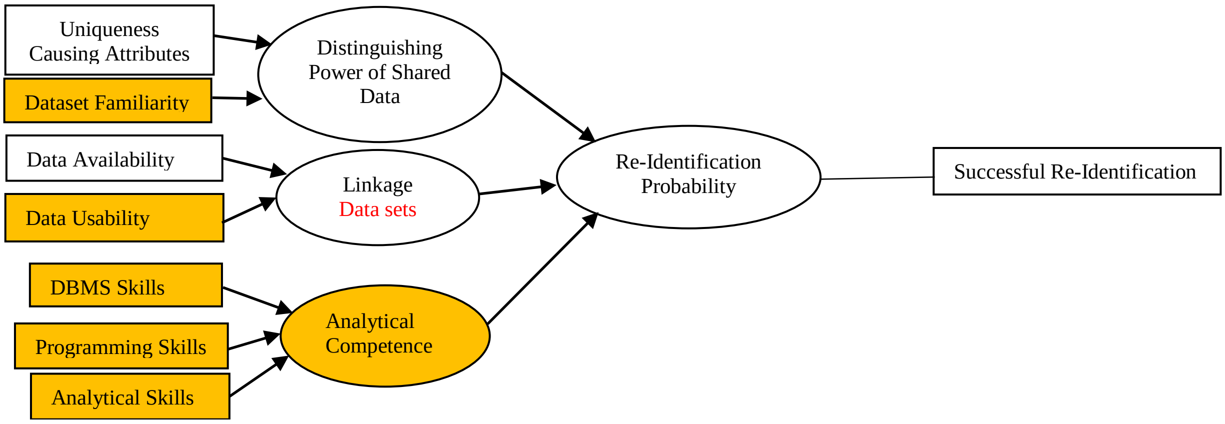


Fig. 1. Proposed Re-Identification Probability Estimation Model

The distinguishing power, linkage data sets, and analytical competence are the independent constructs, also known as exogenous constructs, that predict the dependent construct, re-identification probability, also known as the endogenous construct.

Distinguishing power construct refers to the ability to single out an entity from a dataset, which leads to re-identification of the entity. This is postulated to be determined by the characteristics of the quasi-attributes that are in the data sets and the background information (dataset familiarity) that the analyst may be in possession of. The two become the indicators or measured variables representing the construct. The data set familiarity was not emphasised in the previous model.

The linkage data sets construct refers to the various data sets that the analyst may need to compare with the anonymized data set released by the curator. Such data sets would be containing both explicit and quasi-identifiers. The analyst then matches the quasi-identifiers from the released data to linkage data sets and then uses the explicit identifier to disclose who the entity that had been de-identified is, causing the privacy bleach. This is postulated to be measured using the linkage dataset availability, accessibility, and its usability. Previously, the emphasis was only on the availability, but usability is equally very key.

Analytical competence construct refers to the ability of the analyst interacting with the released data sets to work with data sets and be able to extract relevant data/information aiding in re-identification. The construct was postulated to be measured through analyst's skills in databases, programming statistical mathematics, data mining, and data analytics. This is a new inclusion in the model to emphasise the role of data user in the re-identification process.

The re-identification probability construct refers to the likelihood of an analyst re-identifying an entity that was previously de-identified at the time of data release. The construct is measured by successful re-identification that does happen.

The measured variables (indicators) and the constructs they represent form the measurement or outer model. Our proposed structural equation model hence has the outer and the inner models. Both of them need to be validated empirically for the proposed model to be said to be validated.

9. Conclusions

We have demonstrated the need to come up with a way of determining the probability of being re-identified as the aspect that will make the choice of an appropriate privacy parameter become practical and heuristic. We further demonstrated that the probability of being re-identified will vary from one region to another. This implies that the epsilon value can only therefore be estimated for a given region.

We have improved the model by introducing new indicators for both distinguishing power and linkage data set. We further introduced a new construct (Analytical Competence) to cover the data user or the analyst.

The proposed model needs to be validated empirically, by collecting data and experimenting it in a given region to get the re-identification probability. Once a region has established this probability of re-identification, the value would be plugged into the formula of determining the privacy parameter, epsilon (ε), as stated in equation (3), as the value of P. In this way, we would succeed in making the choice of privacy parameter practical and heuristic, making the application of the ε-differential privacy utilitarian and hence more applicable.

10. Acknowledgment

We thank Dr. Bonface Ngari Ireri, Multimedia University of Kenya, for moral support and also assisting with the editing work during the writing of the manuscript.

References

[1] R. Gupta, "Journey from Data Mining to Web Mining to Big Data," Int. J. Comput. Trends Technol., vol. 10, no. 1, pp. 18-20, 2014.

[2] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: The real-world use of big data," IBM Glob. Bus. Serv. Saïd Bus. Sch. Univ. Oxford, pp. 1-20, 2012.

[3] X. Yao, X. Zhou, and J. Ma, "Differential Privacy of Big Data: An Overview," 2016 IEEE 2nd Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. (HPSC), IEEE Int. Conf. Intell. Data Secur., vol. 9, no. 2, pp. 7-12, 2016.

[4] A. Ramachandran, L. Singh, E. Porter, and F. Nagle, "Exploring re-identification risks in public domains," in 2012 10th Annual International Conference on Privacy, Security and Trust, PST 2012, 2012, pp. 35-42.

[5] A. Cavoukian and D. Reed, "Big Privacy: Bridging Big Data and the Personal Data Ecosystem Through Privacy by Design," 2013.

[6] S. Reddy and O. Prakash, "UTILITY-PRIVACY TRADEOFF IN DATABASES : AN INFORMATION THEORETIC APPROACH," Int. J. Eng. Sci. Res., vol. 4, no. 10, pp. 608-612, 2014.

[7] K. Nissim et al., "Differential Privacy: A Primer for a Non-technical Audience * (Preliminary version)," no. 1237235, 2017.

[8] M. Alfalayleh and L. Brankovic, "Quantifying privacy: A novel entropy-based measure of disclosure risk," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 8986, pp. 24-36.

[9] C. Dwork, "An ad omnia approach to defining and achieving private data analysis," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4890 LNCS, pp. 1-13, 2008.

[10] T. Zhu, G. Li, S. Member, W. Zhou, and P. S. Yu, "Differentially Private Data Publishing and Analysis: A Survey," IEEE Trans. Knowl. Data Eng., vol. 29, no. 8, pp. 1619-1638, 2017.

[11] S. Sruthika and N. Tajunisha, "A Study on Evolution of Data Analytics To Big Data Analytics and Its Research Scope," 2015 Int. Conf. Innov. Information, Embed. Commun. Syst., 2015.

[12] G. Cormode, "The Confounding Problem of Private Data Release," Proc. 18th Int. Conf. Database Theory, pp. 1-12, 2015.

[13] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, "Differentially private data release for data mining," in Proceedings of the 17th ACM

SIGKDD international conference on Knowledge discovery and data mining - KDD ’11, 2011, p. 493.

[14] K. Nissim and A. Smith, “Smooth Sensitivity and Sampling in Private Data Analysis ∗,” Proc. thirty-ninth Annu. ACM Symp. Theory Comput., pp. 75-84, 2007.

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” Proc. 3rd Theory Cryptogr. Conf., pp. 265-284, 2006.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” Theory Cryptogr. SE - 14, vol. 3876, pp. 265-284, 2006.

[17] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, “Differentially private data release for data mining,” Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’11, p. 493, 2011.

[18] S. L. Garfinkel, “NISTIR 8053 De - Identification of Personal Information NISTIR 8053 De - Identification of Personal Information,” 2015.

[19] L. Yin et al., “Re-identification risk versus data utility for aggregated mobility research using mobile phone location data,” PLoS One, vol. 10, no. 10, pp. 1-23, 2015.

[20] H. Vaghashia and A. Ganatra, “A Survey: Privacy Preservation Techniques in Data Mining,” Int. J. Comput. Appl., vol. 119, no. 4, pp. 20-26, 2015.

[21] A.-E.-E. Abdou Hussien, N. Hamza, and H. A. Hefny, “Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing,” J. Inf. Secur., vol. 4, no. April, pp. 101-112, 2013.

[22] N. Li, T. Li, and S. Venkatasubramania, “t -Closeness : Privacy Beyond k -Anonymity and -Diversity,” in IEEE 23rd International Conference, 2007, no. 3, pp. 106-115.

[23] A. Machanavajjhala, D. Kifer, J. Gehrhe, and M. VENKITASUBRAMANIAM, “L-Diversity : Privacy Beyond k - Anonymity,” in Proceedings of the 22nd International Conference on Data Engineering, 2006, vol. V, pp. 1-36.

[24] L. Sweeny, “k- ANONYMITY: A MODEL FOR PROTECTING PRIVACY 1,” Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 10, no. 5, pp. 557-570, 2002.

[25] H. H. Nguyen, J. Kim, and Y. Kim, “Differential Privacy in Practice,” J. Comput. Sci. Eng., vol. 7, no. 3, pp. 177-186, 2013.

[26] C. Dwork, “Differential privacy in new settings,” Proc. Twenty-First Annu. ACM-SIAM Symp. Discret. Algorithms, pp. 174-183, 2010.

[27] C. Dwork and A. Roth, “The Algorithmic Foundations of Differential Privacy,” Found. Trends® Theor. Comput. Sci., vol. 9, no. 3-4, pp. 211-407, 2013.

[28] C. Dwork, “A firm foundation for private data analysis,” Commun. ACM, vol. 54, no. 1, p. 86, 2011.

[29] H. Ebadi, D. Sands, and G. Schneider, “Differential Privacy: Now it’s Getting Personal,” Proc. 42nd Annu. ACM SIGPLAN-SIGACT Symp. Princ. Program. Lang. - POPL ’15, pp. 69-81, 2015.

[30] J. Hsu et al., “Differential Privacy: An Estimation Theory-Based Method for Choosing Epsilon,” Proc. 2014 IEEE 27th Comput. Secur. Found. Symp., pp. 398-410, 2014.

[31] X. Cheng, S. Su, S. Xu, P. Tang, and Z. Li, “Differentially private frequent sequence mining,” IEEE Trans. Knowl. Data Eng., vol. 28, no. 11, pp. 2910-2926, 2016.

[32] L. Fan and H. Jin, “A Practical Framework for Privacy-Preserving Data Analytics,” Www’15, pp. 311-321, 2015.

[33] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally Utility-Maximizing Privacy Mechanisms," SIAM J. Comput., vol. 41, no. 6, pp. 1673-1693, 2012.

[34] A. Ghosh, T. Roughgarden, and M. Sunararajan, "Universally utility-maximizing privacy mechanisms," Proc. 41th STOC, vol. 41, no. 6, pp. 1673-1693, 2009.

[35] R. Sarathy and K. Muralidhar, "Evaluating laplace noise addition to satisfy differential privacy for numeric data," Trans. Data Priv., vol. 4, no. 1, pp. 1-17, 2011.

[36] J. Lee and C. Clifton, "How much is enough? Choosing Epsilon for differential privacy," in Information Security, 14th International Conference, ISC 2011, 2011, vol. 7001 LNCS, pp. 325-340.

[37] C. MIT, "Big Data Privacy Workshop: Advancing the State of the Art in Technology and Practice," Cambridge, Massachusetts, 2014.

[38] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. E. Culler, "GUPT: Privacy Preserving Data Analysis Made Easy," Sigmod, p. 12, 2012.

[39] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets.," BMC Med Inf. Decis Mak, vol. 12, no. September 2009, p. 66, 2012.

Peter N. MUTURI graduated from School of Computing and Informatics, University of Nairobi, Kenya, with MSc. in Information Systems in 2010. He is now a Computer Science PhD candidate, at the School of Computing and Informatics, University of Nairobi. He is currently a Lecturer, Computer Science Department, Faculty of Computing and Information Technology at Multimedia University of Kenya. His domains of work are: Big Data, Data Analytics, and Data Privacy.

Andrew M. KAHONGE graduated from University of Birmingham with MSc. Advanced Computer Science, Specializing in Distributed Systems and Networks, Neural Computation and Virtual Reality, in 2003. He Attained his Doctoral degree from School of Computing and Informatics, University of Nairobi, in Computer Science in 2013, Specializing in Web Security and User Behavior Modeling. At present he is a Senior Lecturer in the School of Computing and Informatics, University of Nairobi. He has published several papers in reputable and international conferences and journals.

Christopher K. CHEPKEN is a Senior Lecturer at the School of Computing an Informatics, University of Nairobi where he has served since 2004. He holds a PhD in Computer Science from the University of Cape Town, South Africa (2013) and a Master of Applied Computer Science from the University of Nairobi (2006), where he also obtained his Bachelor's degree in Computer Science (2004). Christopher has published widely in the area of Applied Computing and other related topics. He has also mentored and supervised several masters and PhD students to completion. His specialization areas include ICT for Development, Systems security and Software Engineering.

String Aggregation Techniques in Oracle
ListAgg – Details, Limitations, and Alternatives

Cristian DUMITRESCU
IBM Romania
mail.cristian.dumitrescu@gmail.com

This paper proposes several ways to overcome the ORA-01489 result of string concatenation is too long error in early versions of Oracle Database (before 12c). Each proposed alternative is presented with pros and cons and may be applied only in specific cases, depending on the requirement or cause.

Keywords: *Relational Databases, Oracle, ListAgg, ORA-01489, User-Defined Functions*

1 Introduction

When working with databases, the aggregation of character fields is a common need in the development of views and reports.

To meet these requirements, Oracle provides a convenient solution through the *ListAgg* function.

As the lifetime support for Oracle 11g has ended in 2020, most companies are looking to migrate to a newer version [1]. However, the 11g version is still widely used.

The paper aims to summarize the options that the database developer has at its disposal to overcome the 4000-character limit imposed on the *ListAgg* function in the Oracle 11g version.

2. The evolution of the built-in *ListAgg* function

ListAgg was first introduced in the Oracle world in version 11gR2.

ListAgg aggregates the values of multiple rows into a single grouping. The rows are, thereby, “denormalized” in a single concatenation of values, optionally delimited by a comma, thus generating a csv (comma-separated-value) result. Similar to other built-in Oracle functions, *ListAgg* can be used as an aggregate function (along with the “group by” clause), or as an analytical function (along with the “*over*” and “*partition by*” clauses).

ListAgg is most often used in conjunction with character attributes, although the function can receive other data types as input.

Another favorable feature of the function is the ability to sort concatenated elements within a group.

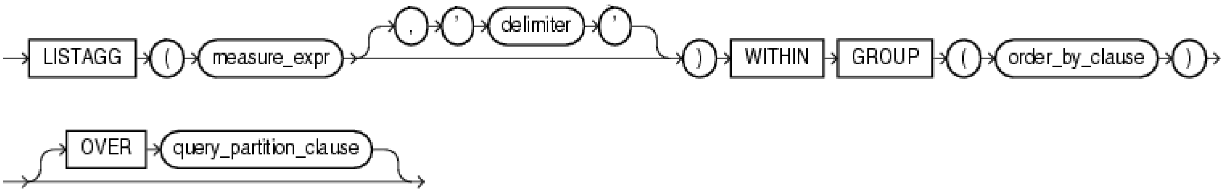
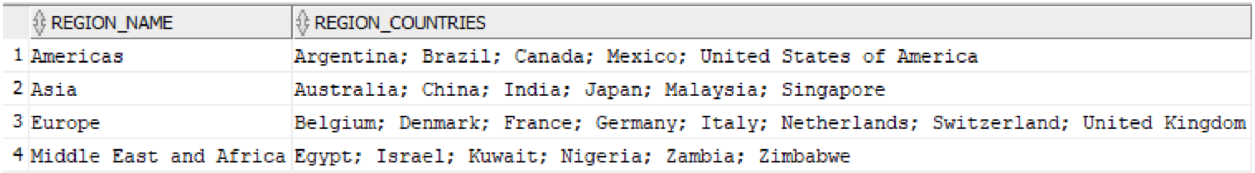


Fig. 1. *ListAgg* syntax [2]

Note: All SQL queries used in this paper will run on Oracle’s well-known HR schema.

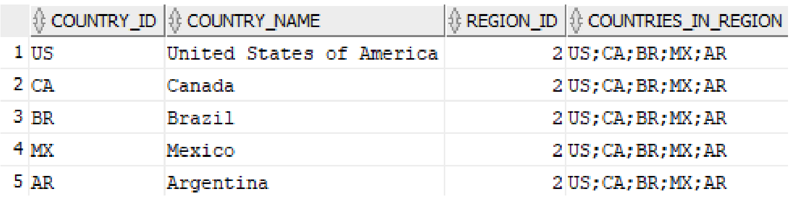
```
select r.region_name,
       LISTAGG(c.country_name,'; ')  WITHIN GROUP (ORDER BY c.country_name)
region_countries
  from regions r
  join countries c on r.region_id = c.region_id
 group by r.region_name;
```



	REGION_NAME	REGION_COUNTRIES
1	Americas	Argentina; Brazil; Canada; Mexico; United States of America
2	Asia	Australia; China; India; Japan; Malaysia; Singapore
3	Europe	Belgium; Denmark; France; Germany; Italy; Netherlands; Switzerland; United Kingdom
4	Middle East and Africa	Egypt; Israel; Kuwait; Nigeria; Zambia; Zimbabwe

Fig. 2. *ListAgg* example in aggregate mode

```
select c.*,
      LISTAGG(c.country_id,';')  WITHIN GROUP (order by null) over
(partition by region_id ) COUNTRIES_IN_REGION
 from countries c
where c.region_id = 2;
```



	COUNTRY_ID	COUNTRY_NAME	REGION_ID	COUNTRIES_IN_REGION
1	US	United States of America	2	US;CA;BR;MX;AR
2	CA	Canada	2	US;CA;BR;MX;AR
3	BR	Brazil	2	US;CA;BR;MX;AR
4	MX	Mexico	2	US;CA;BR;MX;AR
5	AR	Argentina	2	US;CA;BR;MX;AR

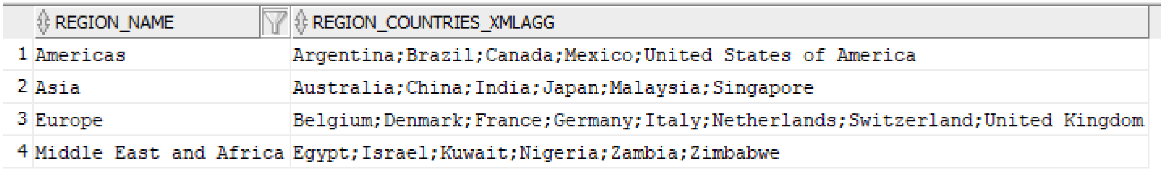
Fig. 3. *ListAgg* example in analytical mode

The optional delimiter should be carefully chosen because in the output it can be confused with a part of the field being concatenated. Best practice dictates choosing a special character that is not present in the aggregated fields.

The result of the function is a *varchar2*, limited to 4000 characters.

Before the 11g version, Oracle had no direct mechanism that allowed multiple values of the same column to be displayed on a single row in the output. In a post on his blog, Donald Burlescon presents some solutions to this problem: using the *XMLAgg* function (Oracle 9i) and using the *SYS_CONNECT_BY_PATH* operator [3], [4].

```
select r.region_name
      ,rtrim(xmlagg(xmlelement(e, c.country_name || ';') ORDER BY
c.country_name).extract ('//text()'), ';') region_countries_XMLAGG
  from regions r
  join countries c on r.region_id = c.region_id
 group by r.region_name;
```

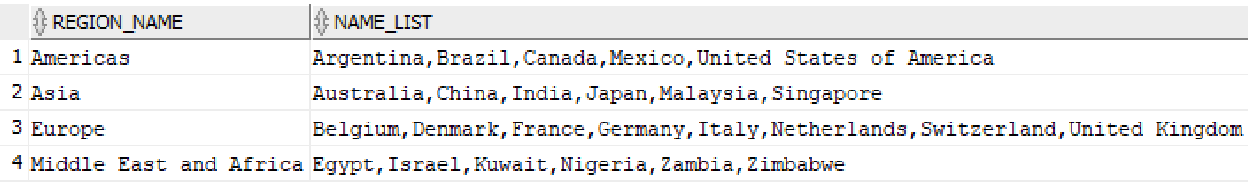


	REGION_NAME	REGION_COUNTRIES_XMLAGG
1	Americas	Argentina;Brazil;Canada;Mexico;United States of America
2	Asia	Australia;China;India;Japan;Malaysia;Singapore
3	Europe	Belgium;Denmark;France;Germany;Italy;Netherlands;Switzerland;United Kingdom
4	Middle East and Africa	Egypt;Israel;Kuwait;Nigeria;Zambia;Zimbabwe

Fig. 4. *ListAgg* alternative – String concatenation with *XMLAgg*

```
select region_name,
        substr(SYS_CONNECT_BY_PATH(country_name, ','),2) name_list
```

```
   from (select r.region_name,
                c.country_name,
                count(*) OVER ( partition by r.region_name ) cnt,
                ROW_NUMBER () OVER ( partition by r.region_name order by
c.country_name) seq
           from regions r
           join countries c on r.region_id = c.region_id          where
r.region_name is not null)
  where seq = cnt
  start with seq = 1
connect by  prior seq + 1 = seq
        and prior region_name = region_name;
```



	REGION_NAME	NAME_LIST
1	Americas	Argentina,Brazil,Canada,Mexico,United States of America
2	Asia	Australia,China,India,Japan,Malaysia,Singapore
3	Europe	Belgium,Denmark,France,Germany,Italy,Netherlands,Switzerland,United Kingdom
4	Middle East and Africa	Egypt,Israel,Kuwait,Nigeria,Zambia,Zimbabwe

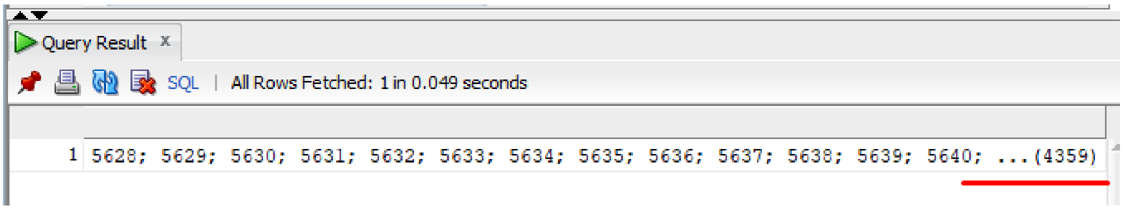
Fig. 5. *ListAgg* alternative – String concatenation with *SYS_CONNECT_BY_PATH*

In the event that the function exceeds the 4000 characters limit, the following runtime error occurs:

```
ORA-01489: result of string
concatenation is too long
```

In Oracle 12c, *ListAgg* has received an upgrade through which this limitation can be elegantly overcome - the *ON OVERFLOW* clause.

```
  with my_query as (select level no
                      from dual connect by level <10000)
select LISTAGG(no,'; ' on overflow truncate) WITHIN GROUP (order by null)
as no_list
  from my_query;
```



```
1 5628; 5629; 5630; 5631; 5632; 5633; 5634; 5635; 5636; 5637; 5638; 5639; 5640; ...(4359)
```

Fig. 6. *ListAgg* with *ON OVERFLOW* clause

The query above would generate an error if we deleted or commented on the "*on overflow truncate*" clause.

If the string aggregation output exceeds 4000 characters, the function truncates the result before this threshold, notifying the user with a customizable message (the example above uses the standard option of "...") [5].

Starting with the Oracle 19c version, the function receives new improvements: The "within group" clause becomes optional, and the concatenation can be performed for distinct values, by using *ListAgg* and *DISTINCT* together [6].

3. Avoiding *ListAgg's* Overflow Error

Runtime Errors such as ORA-01489 are difficult to detect during development because the syntax itself is correct, the error is actually caused by the volume of the data. For example, a view that uses this function can be developed in a test environment with a small volume of data and will not generate this error until it is introduced in the production database. Deleting or restricting the data so that the function falls within the range dictated by the threshold is certainly not a solution. Several options will be explored, along with pros and cons depending on how *ListAgg* is used.

Newer versions of Oracle provide multiple solutions to this common problem:

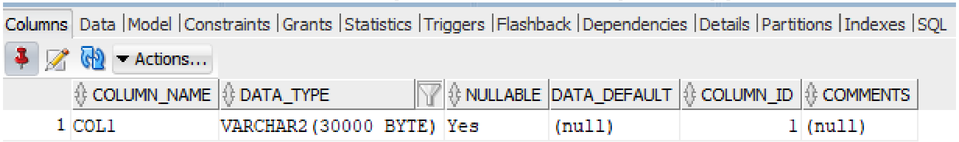
3.1 Replacing ListAgg with XMLAgg

By performing this operation, the view in question will no longer produce a runtime error. This solution should be applied with caution, as the data type of the result changes from *varchar2* to *clob*. If the view is subsequently used by various reporting or ETL tools, they may reject such a field.

In situations where returning the entire result is mandatory, you can choose this option.

3.2 Extending the varchar2 limit from 4000 up to 32767 characters, by setting *MAX_STRING_SIZE* initialization parameter to *EXTENDED*.

```
create table test_large_varchar (col1 varchar2(30000));
```



	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	COL1	VARCHAR2(30000 BYTE)	Yes	(null)	1	(null)

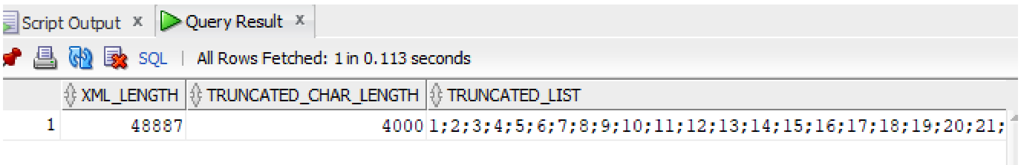
Fig. 7. Creating a table with a *varchar2* field of more than 4000 characters

Admin rights are required to apply this method. The database must also be in *Upgrade* mode.

This option does not directly solve the current issue, but it can prove to be a decent solution if the developer needs to store such aggregation in a table within a non-clob field.

3.3 Replacing ListAgg with Substring, To_Char, XML_Agg

```
  with my_query as (select level no from dual connect by level <10000)
select length(t.no_list) xml_length,
       length(to_char(substr(t.no_list,1,4000))) truncated_char_length,
       to_char(substr(t.no_list,1,4000)) truncated_list
from (
select rtrim(xmlagg(xmlelement(e, no || ';') ORDER BY null).extract
('//text()').getclobval(), ';') as no_list
  from my_query) t;
```



All Rows Fetched: 1 in 0.113 seconds

	XML_LENGTH	TRUNCATED_CHAR_LENGTH	TRUNCATED_LIST
1	48887	4000	1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;

Fig. 8. Replicating *on overflow truncate* clause of *ListAgg* by using a combination of *Substring*, *To_Char*, and *XML_Agg*

With this method, the output is first truncated if the string aggregation exceeds 4000 characters. The result is then transformed from *clob* to character data type.

Note: *XMLAgg* can return the result in both *clob* or character data types, using *getclobval()* or *getstringval()*. In the above query, using *getstringval()* is not an option because it would generate an error similar to the one we were trying to avoid.
This option can be chosen in situations where keeping the character data type prevails over a complete string aggregation output.

3.4 Creating new User-Defined Aggregate Functions

Starting with Oracle 9i, developers can create custom Aggregate Functions. Keith Laker, Oracle's Senior Principal Product Manager, details such an example of a user-defined aggregate function in his blog post [7]. The same topic is also addressed on the famous IT blogs https://www.stackoverflow.com and AskTom [8].

According to Oracle [9], User-defined aggregate functions are used in SQL DML statements, just like Oracle's own built-in aggregates. Once such functions are registered with the server, Oracle simply invokes the aggregation routines that you supplied instead of the native ones. User-defined aggregates can be implemented using `ODCIAggregate` interface routines.

You can create a user-defined aggregate function by implementing a set of routines as methods within an object type, so the implementation can be in any Oracle-supported language for type methods, such as PL/SQL, C/C++, or Java. When the object type is defined and the routines are implemented in the type body, you use the CREATE FUNCTION statement to create the aggregate function.

Each of the four `ODCIAggregate` routines required to define a user-defined aggregate function codifies one of the internal operations that any aggregate function performs, namely:

- Initialize - Initializes the computation;
- Iterate – processes each successive input value;
- Merge – combines two aggregation contexts and returns a single aggregation context;
- Terminate – computes the result.

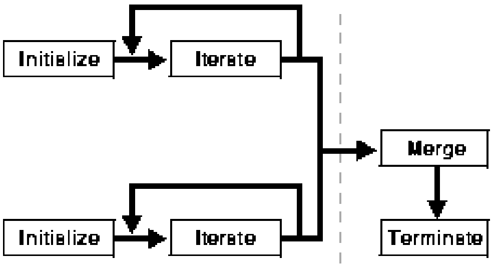


Fig. 9. ODCIAggregate routines [9]

Keith Laker's proposed user-defined function is developed in the following manner:

An initial object is created with the purpose of storing the results from the iterate stage. In this example, the object is created as a table of *varchar2(25)*, but the size can be up to 4000 bytes, even *clob*, depending on the way the final aggregate function is being used. For example, if we plan on concatenating first names and last names, this object should have the size of the two columns combined.

```
CREATE OR REPLACE TYPE string_varray AS TABLE OF VARCHAR2(25);

CREATE OR REPLACE TYPE t_string_agg AS OBJECT
(
 a_string_data string_varray,
 STATIC FUNCTION ODCIAggregateInitialize(sctx IN OUT t_string_agg) RETURN
NUMBER,
 MEMBER FUNCTION ODCIAggregateIterate(self IN OUT t_string_agg, value IN
VARCHAR2 ) RETURN NUMBER,
 MEMBER FUNCTION ODCIAggregateTerminate(self IN t_string_agg, returnValue
OUT VARCHAR2, flags IN NUMBER) RETURN NUMBER,
 MEMBER FUNCTION ODCIAggregateMerge(self IN OUT t_string_agg, ctx2 IN
t_string_agg) RETURN NUMBER
);

CREATE OR REPLACE TYPE BODY t_string_agg IS
 STATIC FUNCTION ODCIAggregateInitialize(sctx IN OUT t_string_agg) RETURN
NUMBER IS
 BEGIN
    sctx := t_string_agg(string_varray() );
    RETURN ODCIConst.Success;
 END;
MEMBER FUNCTION ODCIAggregateIterate(self IN OUT t_string_agg, value IN
VARCHAR2) RETURN NUMBER IS
 BEGIN
    a_string_data.extend;
    a_string_data(a_string_data.count) := value;
    RETURN ODCIConst.Success;
 END;
MEMBER FUNCTION ODCIAggregateTerminate(self IN t_string_agg, returnValue
OUT VARCHAR2, flags IN NUMBER) RETURN NUMBER IS
   l_data varchar2(32000);
   ctx_len NUMBER;
   string_max NUMBER;
BEGIN
   ctx_len := 0;
   string_max := 4000;
   FOR x IN (SELECT DISTINCT column_value FROM TABLE(a_string_data) order
by 1)
    LOOP
       IF LENGTH(l_data || ',' || x.column_value) <= string_max THEN
          l_data := l_data || ',' || x.column_value;
       ELSE
          ctx_len := ctx_len + 1;
       END IF;
   END LOOP;
   IF ctx_len > 1 THEN
        l_data := l_data || '...(' || ctx_len||')';
    END IF;
    returnValue := LTRIM(l_data, ',');
    RETURN ODCIConst.Success;
 END;
MEMBER FUNCTION ODCIAggregateMerge(self IN OUT t_string_agg, ctx2 IN
t_string_agg) RETURN NUMBER IS
 BEGIN
    FOR i IN 1 .. ctx2.a_string_data.count
     LOOP
        a_string_data.EXTEND;
        a_string_data(a_string_data.COUNT) := ctx2.a_string_data(i);
     END LOOP;
    RETURN ODCIConst.Success;
 END;
END;
```

The last step is creating a function – the actual *ListAgg* alternative, which receives a string as input and calls the string-processing object described above.

```
CREATE OR REPLACE FUNCTION string_agg (p_input VARCHAR2)
RETURN VARCHAR2
PARALLEL_ENABLE AGGREGATE USING t_string_agg;
```

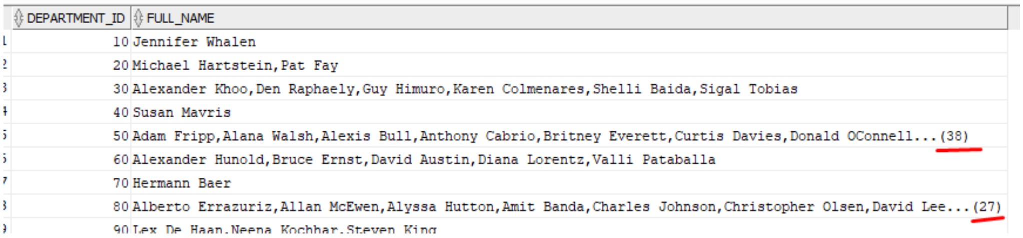
We should observe that in `ODCIAggregateTerminate` routine, `string_max` is set to 4000. This basically reproduces the *ON OVERFLOW TRUNCATE* functionality of *ListAgg*'s 12c version.

The advantage of such an approach is that the function can directly answer the problem we face with a custom solution. The initial object can be as large or as small as we need it to be. The overflow truncate can occur at any given threshold. This versatile approach can be further used in any given SQL Statement, no matter its complexity.

One disadvantage could be that the separator is embedded within the code. The developer cannot change it as easily as with *ListAgg*.

Let us observe the user-defined aggregate in action (in the below example, `string_max` is set to 100):

```
select e.department_id,
       string_agg(e.first_name|| ' ' || e.last_name) as full_name
from employees e group by e.department_id;
```



	DEPARTMENT_ID	FULL_NAME
1	10	Jennifer Whalen
2	20	Michael Hartstein,Pat Fay
3	30	Alexander Khoo,Den Raphaely,Guy Himuro,Karen Colmenares,Shelli Baida,Sigal Tobias
4	40	Susan Mavris
5	50	Adam Fripp,Alana Walsh,Alexis Bull,Anthony Cabrio,Britney Everett,Curtis Davies,Donald OConnell...(38)
6	60	Alexander Hunold,Bruce Ernst,David Austin,Diana Lorentz,Valli Pataballa
7	70	Hermann Baer
8	80	Alberto Errazuriz,Allan McEwen,Alyssa Hutton,Amit Banda,Charles Johnson,Christopher Olsen,David Lee...(27)
9	90	Lex De Haan,Neena Kochhar,Steven King

Fig. 10. *ListAgg* functionality replicated with a user-defined function

4. Conclusions

ListAgg currently remains a powerful string aggregation alternative. Throughout their career, developers usually work with multiple versions of databases, which is why a review on *ListAgg* evolution, limitations, and string processing alternatives may prove useful.

References

[1] https://support.oracle.com/knowledge/Oracle%20Cloud/2068368_1.html

[2] https://docs.oracle.com/

[3] http://www.dba-oracle.com/t_oracle_listagg_function.htm

[4] https://oracle-base.com/articles/misc/string-aggregation-techniques#wm_concat

[5] https://blog.dbi-services.com/oracle-12cr2-sql-new-feature-listagg-overflow/

[6] https://rogertroller.com/2020/01/07/oracle-19c-listagg-enhancement/

[7] https://blogs.oracle.com/datawarehousing/exploring-the-interfaces-for-user-defined-aggregates

[8] https://asktom.oracle.com/pls/apex/f?p=100:11:0::::p11_question_id:15637744429336

[9] https://docs.oracle.com/cd/B10501_01/appdev.920/a96595/dci11agg.htm#1004615

Cristian DUMITRESCU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2010. Cristian works as a Data Engineer for IBM Romania and has 7 years of experience in working with Oracle Databases. His area of expertise includes Relational Databases, SQL, PL/SQL, OLTP, OLAP, and Reporting.

Appointment Scheduling System for a Primary Hospital

Norman GWANGWAVA[1], Kgalalelo D. NTESANG[2]
[1,2]Botswana International University of Science and Technology
gwangwavan@biust.ac.bw, kgalalelo.ntesang@studentmail.biust.ac.bw

Many primary hospitals in developing countries face serious shortages of equipment and skilled personnel to handle cases reporting to them. This article focuses on a primary hospital constructed in the 1970s, with 29 facilities reporting to it. Patients referred to the hospital are usually ferried in ambulances if they are exhibiting critical conditions. Patients deemed uncritical are given referral letters. However, patients are exposed to long waiting times that put their lives at risk or worsen their conditions. The research aims to establish improved ways in which the patient waiting times can be reduced. An appointment scheduling framework for the primary hospital is conceived as a better approach. This is an SMS based queue management system. The system reduces the waiting time of patients in the hospital's outpatient department. A patient registration device that contains a GSM module and a microcontroller which sends messages to and from the patient when booking an appointment for consultation are developed. This queue management system has the potential to reduce patient waiting times by more than 95%.

Keywords: *Queuing System, Queue Management, GSM, Outpatient, Healthcare, Patient Appointment, Scheduling*

1 Introduction

The waiting times for receiving treatment in hospitals are high in many public hospitals, particularly in developing countries. Patients waiting too long before being attended to may acquire some infections, or their current situation may become worse. There is a possibility of some patients going undiagnosed. A triage system is used, patients who seem to be in a critical condition may have some underlying problems, and the extension of their waiting time may lead to their problem growing. Prolonged waiting times are a result of overcrowding in hospitals. Overcrowding is not good, as it leads to doctors having many patients to attend to. Medical errors may also arise because of the pressure on doctors to try to help as many people as possible.

Scheduling too many patients on the same day is one of the causes of long waiting times in queues for health treatment. Referral scheduling is necessary to fill in the gap of long waiting periods for medical treatment. [1] Observes current trends where modern life is becoming too busy to make medical appointments in person and maintain proper health care. This prompted the researchers to provide ease and comfort to patients through an online appointment system. Queuing is a general problem in many service industries. The service provider industries must meet the needs of both the customer and the service provider, hence scheduling appointments can be difficult [2]. Customers resent long waits, whereas service providers are pressed to minimize the idle time of their resources and the use of overtime [3]. The research conducted by [4] implements four appointment scheduling policies, i.e., constant arrival, mixed patient arrival, three-section pattern arrival, and irregular arrival, in an ultrasound department of a hospital in Taiwan. The study helps hospital managers seeking to improve key performance indicators (KPIs) on the patient appointment scheduling system to focus more on searching for optimal solutions or developing better appointment scheduling policies. The current process flow of the outpatient department in the case study hospital is shown in **Fig. 1**.

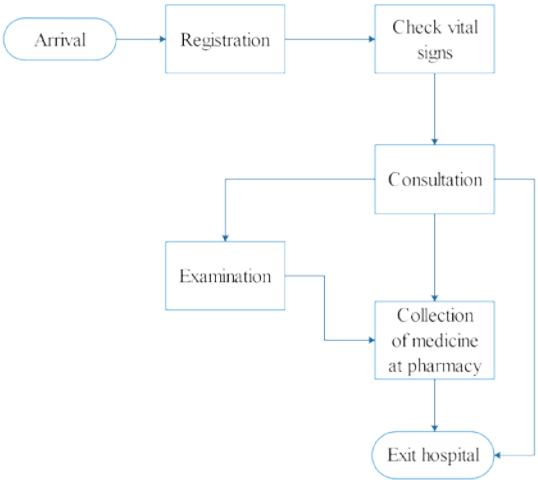


Fig. 1. Process flow for the hospital Out-Patient Department

2. Objectives

Research aims to reduce patients' waiting time in a primary hospital by developing a queue management system. The authors undertake the following steps to come up with the most feasible solution;

- Evaluate the current queue management system used by the hospital;
- Benchmark the existing system against that of successful hospitals;
- Determine where most of the patients' time is spent waiting for assistance;
- Design a more efficient and effective queue management system.

The study uses statistics for patients attended to within a three (3) months period. A framework will be produced to help guide the hospital in the most efficient operational methods to minimize waiting times. Minimizing waiting times in a hospital setup is a critical part of the process, as saving lives is the ultimate goal. The envisaged platform is accessible to patients and hospital staff. It aims to solve the challenges faced by patients while taking appointments and keeping medical files. The advanced system features allow doctors to access and update a patient's medical record after every check-up. Some of the features are as follows: Online follow-up with doctors for distant patients, and linking laboratories and the pharmacy in order to allow the medical administrator to view suggested prescriptions, whilst laboratories can view clinical tests recommended by the Doctor. The system is implemented for all the individuals who seek treatment with the primary hospital. Only registered users can participate. Potential users must create an account through the registration form and should provide their medical history. Patients' records would be updated automatically after each doctor's visit.

3. Queuing in healthcare

The scheduling approaches for healthcare providers are hampered by the challenge of determining how to schedule the number of patient appointments using special time slots based on fluctuations in patients and stochastic patient treatment time [5]. A typical appointment scheduling problem can consist of one or more objective functions, such as minimizing patients' average waiting time, machines'/doctors' average idle time, overtime, and cost [4]. Most healthcare facilities use queueing systems where patients arrive, wait for a healthcare service in a queue, obtain a service, and then depart from a healthcare facility [6]. Queuing theory is used to define analytical techniques, which are closed mathematical formulas, which describe a sequence of dealing with situations where there are congestions and blockages. The services of a healthcare outpatient department involve patients who seek service, wait in a queue before service, and depart the system after being served; an out-patient department is regarded as a queueing system [7,8]. Many countries experience long waiting times. Excessive waiting for treatments may cause deterioration in patient's health, reduce treatment effectiveness, resulting in a barrier in the access to health care services [8]. Patient's waiting is defined as the time between the patient's arrival time and the actual service start time of the patient's service or appointment. Patients' waiting time is affected by;

- Tardiness of earlier patients, causing a delay in the scheduled treatment time;
- Patients having different treatment times, creating a stochastic environment [4];
- Efficiency of the server or personnel providing the service sought by the patient's socioeconomic status [8].

Factors considered in analyzing queuing problems are line or queue length, number of lines, and the queue discipline. Services are mostly offered on a first-come-first-served basis although other services use reservations first and the triage system. The basic queuing model is a single-server, implying that service is provided at a single point from a single line or queue of patients. Queues are classified as finite or infinite by looking at the maximum possible number of patients that a queue can contain. Queue discipline refers to the way in which members of the queue are selected for service. Most healthcare facilities use the First-In-First–Out (FIFO) queue discipline or they categorise patients into sets according to their priorities. However, the priority discipline reduces waiting time for more critical patients, while the average waiting time for lower priority or stable patients increases [9]. Patient flow represents the movement of patients in a healthcare facility. It also shows the ability of a healthcare facility to provide services to patients from arrival to departure by making use of the available resources and ensuring that the quality of the services provided is not compromised [10]. The flow of patients should be quicker to avoid a blockage in the flow which may lead to longer waiting times and throughput time, hence a negative effect on the delivery of services. Short waiting times in all phases of a healthcare system are signs that the patient flow is well managed and that generally improves healthcare services [11]. Factors which have an influence on the operational efficiency based on patient flow include daily patients volume, health care facility policies such as how often patients visit, the type of the healthcare provider attending to patients, the size and the combination of the service providers, and the type of staffing used [12]. Ineffectiveness and inefficiency of appointment systems is one of the factors that lead to prolonged waiting times for patients in the outpatient department [13]. Some of the ways which may be used to reduce waiting times in healthcare facilities are as outlined below:

- Demand Management: Developing a partnership between primary and secondary care with the aim to manage and deliver reduced waiting times is important, since waiting starts in the primary care hospitals. Patients may also be scheduled by matching the demand for medical care with the resources available at the healthcare facility [13].
- Queue Management: Managing queues may ease patient flow in a hospital, as well as reduce the patients' waiting times before they can receive a healthcare service. [14] Suggests the use of an SMS system which notifies patients of their predicted service time as one of the tools that can reduce overcrowding in hospitals, hence reduce waiting times.
- Queue Index System: Index numbers are used to show the position of the patient in a queue. The patient may leave to attend to other issues once they get their index number. Some index cards show the number of people waiting to get the service ahead of the current index number. The system does not indicate the time a patient may have to wait before being attended to.

Patients depart through numerous routes once they are served. Some of the exit fates are listed below:

- Patient is admitted to specialized hospital units;
- Patient may receive the service as expected;
- Patient may be delayed and choose to get the service elsewhere;

- Patient was advised by a health worker to seek services elsewhere if they are unable to provide it.

4. Online outpatient scheduling systems

Online outpatient scheduling systems have been designed to curb problems of high patient no-show rates and long waiting times experienced when using traditional approaches [15]. A study on the effects of an Online Appointment Scheduling System on Evaluation Metrics revealed a significant positive effect on the improvement of the three metrics means, including Patient waiting time, No-show rate, and Physician punctuality [15]. Online systems offer benefits such as scheduling an appointment at the right time and date with the intended physician and 24-hour access to the system, which increase patient satisfaction [16–18]. They also reduce the patients waiting time and permit adjusting healthcare facility capacity through reducing the number of phone calls. Online systems improve the quality of care and the accessibility of patients to outpatient services [19–22]. Other metrics used to evaluate online outpatient scheduling systems are listed below:

- Patient punctuality: The difference between a patient’s appointment time and the actual arrival time; [23]
- Clinic size: The number of patients scheduled per clinic session [23]
- Walk-in rate: The number of patients who walk in without appointments as a percentage of all appointments; [23]
- Service times: The amount of time the physician spends with the patient; [24]
- Panel size: The number of patients covered by the physician. [25]

Real-time appointment scheduling systems have been explored, whereby a patient only fixes a time and date, and the system allocates a doctor available at that particular time and date and also handles the rescheduling of patients with doctors [26]. The object-oriented analysis and design approach were used for development, whilst the android studio was chosen for mobile implementation. [27] Designed a patient appointment and scheduling system, using Angular JS for the frontend, Ajax framework for handling client-server request and Sqlite3 and MYSQL for the backend.

[28] Propose an online patient appointment scheduling system based on the Web Services architecture. The results show that the Web Services architecture provides an ideal design paradigm for the development of an integrated health care information system in the primary care setting.

[29] Developed an Online Doctor's Appointment and Medical Database Management System. The purpose of the application was to enable patients to easily compare, choose, and make an online appointment for a doctor just by sitting at home. The development tools used are HTML, CSS, and JavaScript for the client side, while PHP and MySQL for the server side.

Introducing an appointment schedule reduced the maximum waiting time for all patients by 42% [30].

5. Research method

A stratified sampling method was used to select subjects for the research. The staff in the Out-Patient Department (OPD) were divided according to the nature of their work and then a few were chosen from these small groups to participate in the research. The OPD has patients coming in and out of the hospital, so it was easy to keep track of the time that they waited to get assistance. The data collection methods used for this research are a questionnaire, an oral interview, and observation. A set of questions was prepared prior to visiting the hospital and a total of twelve (12) questionnaires were distributed among the OPD staff. The first part of the questionnaire addressed the first objective of evaluating the current scheduling system used by the hospital. The next point then made a comparison to what other hospitals are doing with regard to addressing the schedule and managing patients' times. The third set of questions looked into statistics to get average times taken before patients get assistance.

An oral interview was used to personally understand the hospital operations from the management of the facility as a follow-up to the questionnaire. A set of questions prepared before the meeting was directed towards the hospital manager to get a clear view of how they deal with long queues using their scheduling program. These questions were to clarify how they deal with patients' flow in terms of staffing, as well as how they cater for shift changes to ensure there are not many delays in helping patients caused by change of personnel on duty.

Observation of patient movement within the hospital was made from the time a patient arrives until they get assistance and leave the hospital. The waiting times at each service station were recorded. From registration, temperature check, consultation, to pharmacy, patients' movements were observed. This method is effective since the information is first-hand and has no distortions.

The data collected from the healthcare facility is presented in graphs and charts for easier interpretation. Trends in patient movement can be easily identified through these representations, thus easing the work towards a solution. **Table 1**. shows the waiting times for 10 patients at each service point. The waiting times and averages are plotted on a graph in **Fig. 2.** On average, patients spend 217 mins waiting for consultation. The proportion of waiting time is 2% for registration, 21% checking vital signs, and 72% consultation as shown in **Fig. 3.**

Table 1. Waiting times at each service point

Patient No.	**Waiting Time Before Registration (Mins)**	**Waiting Time Before Checking of Vital Signs (Mins)**	**Waiting Time Before Consultation (Mins)**
1	8	58	222
2	10	60	213
3	4	63	224
4	5	61	211
5	5	59	229
6	6	60	214
7	5	61	209
8	5	57	218
9	6	59	216
10	4	63	214
Average	**5.8**	**60.1**	**217**

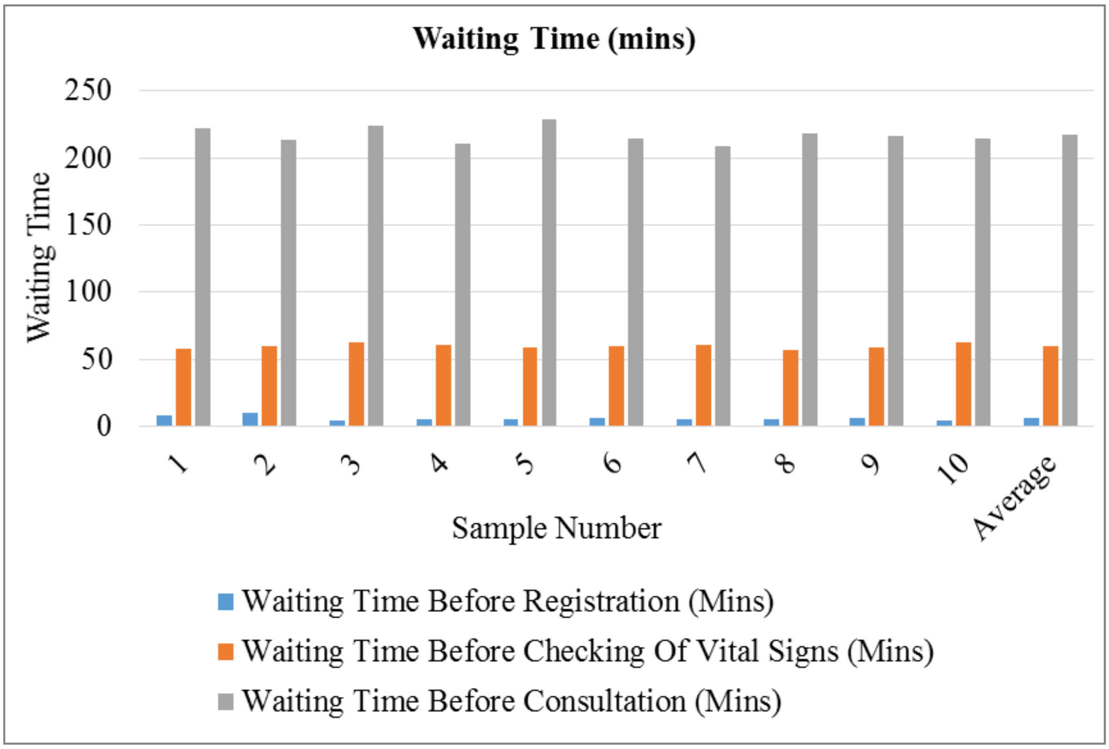


Fig. 2. Waiting times at each service point

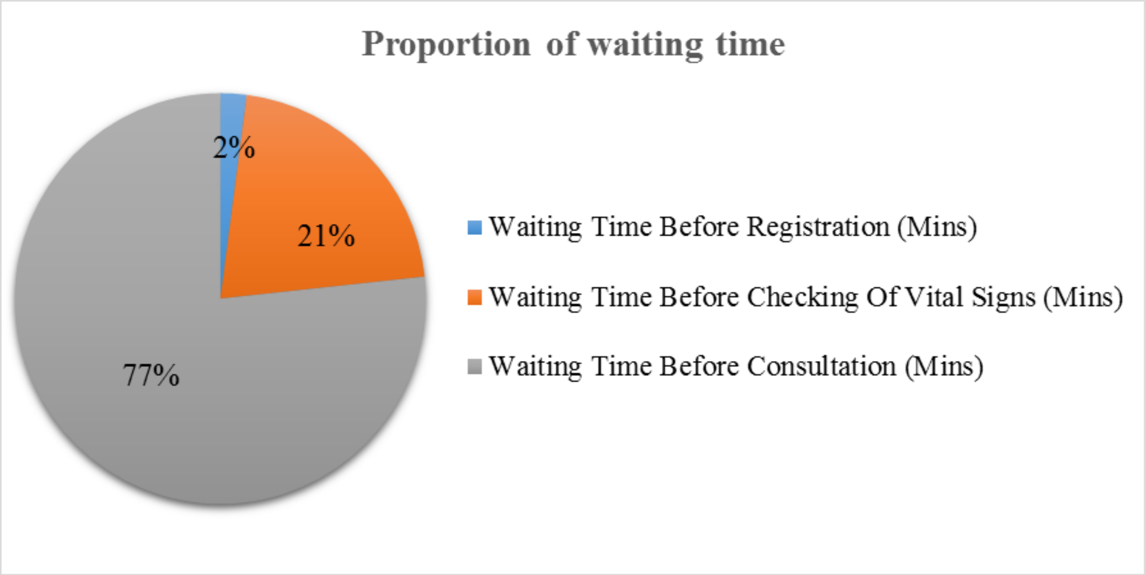


Fig. 3. Proportion of waiting time

6. System modeling and design

Three categories of queue management systems were identified from the literature as the traditional queueing method, online system with tokens, and the SMS-based appointment. These were compared using the weighted matrix method. After the selection of the best solution, the subsequent modeling of the proposed solution was followed through the database design methodologies, conceptual modelling, and logical and physical design [31].

The following attributes were considered in the ranking of the systems:

- User Friendliness: The proposed solution should be easier to understand both for the hospital personnel and the patients.

- Response Time: the solution should give a quick response to the user.
- Flexibility: the solution should give patient freedom to run personal errands while waiting for their time.
- Schedule Visibility: the scheduled times should be visible to the patient and hospital personnel to keep them both in sync.
- Delay Reduction: the solution should reduce the patient's waiting time.

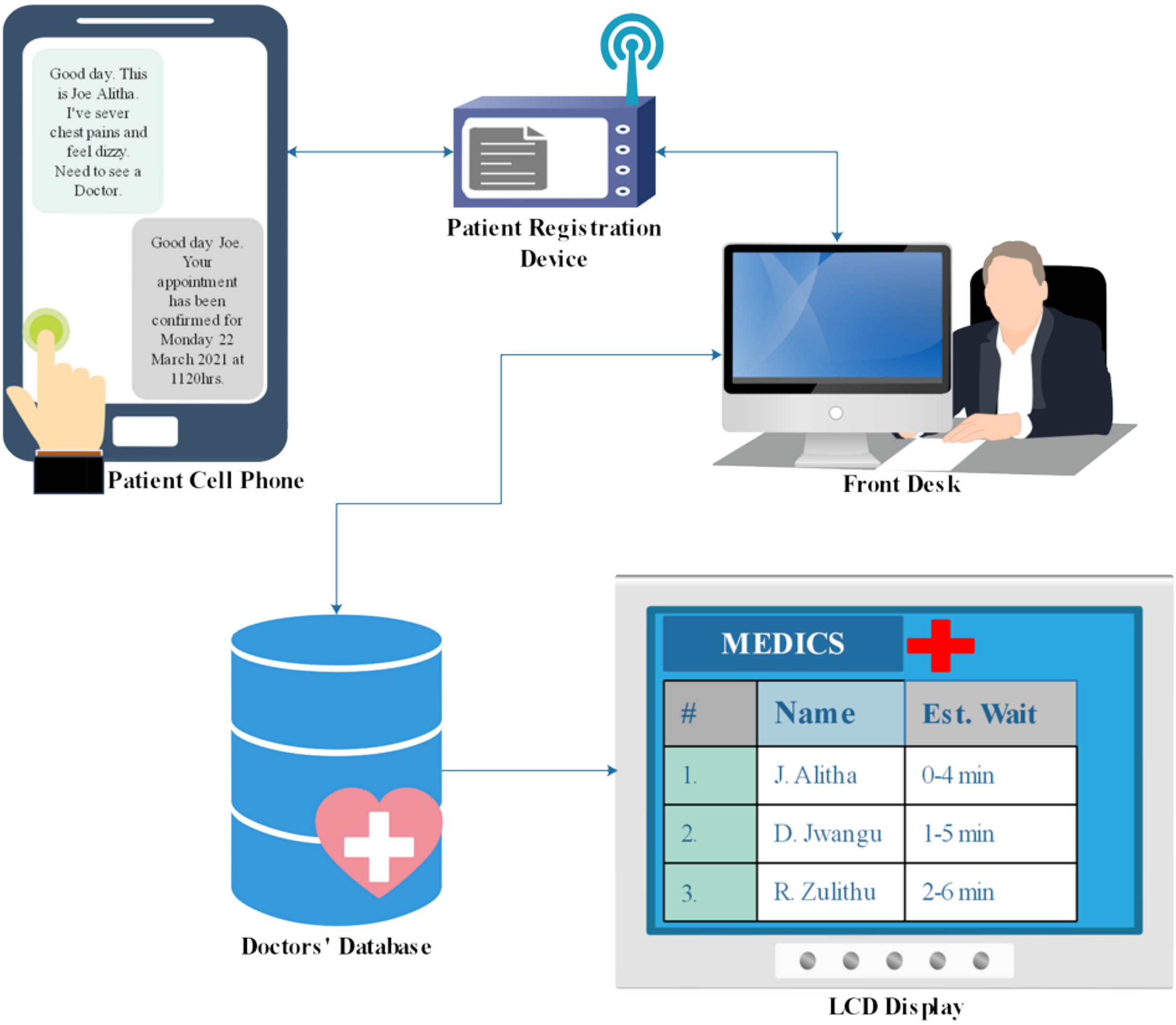
Using a ranking scale of 1 to 5 as outlined below, an SMS based system was selected for further development, as shown from results in **Table 2.**

Rating scale:
5- Very Satisfactory
4-Slightly Satisfactory
3-Fair
2-Dissatisfactory
1-Very Dissatisfactory

Table 2. Waited scoring matrix

Feature	**Weight (%)**	**EQMS with tokens**		**SMS**		**Traditional**	
		Score	Total	Score	Total	Score	Total
User friendliness	20	4	80	4	80	2	40
Response time	10	5	50	4	40	3	30
Flexibility	10	3	30	4	40	2	20
Schedule visibility	25	2	50	5	125	1	25
Delay Reduction	35	1	35	5	175	2	70
TOTAL	**100**		**245**		**460**		**185**

The SMS-based appointment system is used to help patients book their doctor's appointment from the comfort of their homes. The main objective of this system is to reduce congestion in hospitals, consequently reducing patients' waiting times. The system architecture is made up of two parts, the client side (patient) and the server side (hospital). The client-side interaction is through the SMS application of the patient's mobile phone, whereas the server side is the combination of the desktop application and the hospital database. The desktop application consists of written SQL queries that make access to the database possible. A patient sends their details, purpose of their visit, and brief description of their condition through hospital hash tags designed to access the central patient registration platform. The system then schedules the patient to come through at a time when they would not have to wait in queues. The patient receives confirmation of their appointment via a text message. This arrangement allows for the patient to carry out their daily chores and only make their way to the hospital at a time of their appointment. **Fig. 4.** shows the system overview.



Good day. This is Joe Alitha. I've sever chest pains and feel dizzy. Need to see a Doctor.

Good day Joe. Your appointment has been confirmed for Monday 22 March 2021 at 1120hrs.

MEDICS

#	Name	Est. Wait
1.	J. Alitha	0-4 min
2.	D. Jwangu	1-5 min
3.	R. Zulithu	2-6 min

Fig. 4. Proportion of waiting time

Patients book appointments by sending an SMS to the hospital and receive immediate feedback confirming receipt of their appointment details. An internal system scans through the relevant doctor's schedule and the front desk operator confirms the appointment. The system then sends a confirmation SMS to the patient containing the date and time of their appointment. After confirmation of the appointment, the patient details are stored in a database. These details are accessed on the patient's appointment date and displayed on the LCD screen when it is their turn to get consultation. A use case diagram for making an appointment is shown in **Fig. 5.** whilst **Fig. 6.** shows the activity diagram. The primary actor is the patient. The other actor is the front-desk administrator. First-user patients do not necessarily need prior registration in the system. Any patient can dial the hash tags to submit their first intention. Full registration details will be submitted upon visiting the hospital, and records can be retrieved on recurring visits or interaction with the system. The front-desk administrators are registered in the system, search for the doctor, and book an appointment. Other hospital staff – doctor/ physician, pharmacists, can log in to the system with a username and password, accept patient's appointment requests, and update medical record after each visit. Checking the doctor's schedule is done automatically through keyword search based on the patient's inquiry, as illustrated in the activity diagram in **Fig. 7.** The process is executed through retrieving the patient messages from the patient registration device by the hospital application system.

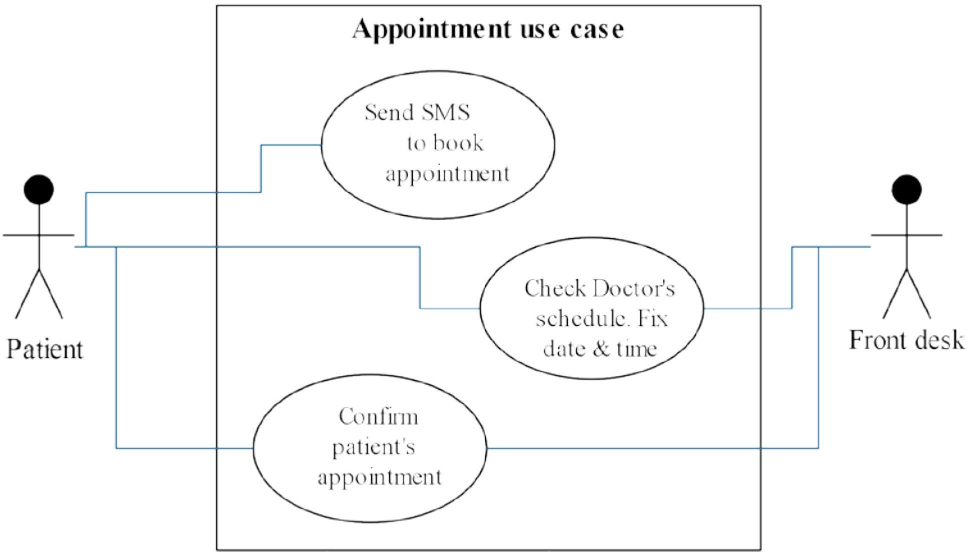

Appointment use case
Send SMS to book appointment
Check Doctor's schedule. Fix date & time
Confirm patient's appointment
Patient
Front desk

Fig. 5. Use-case diagram for making an appointment

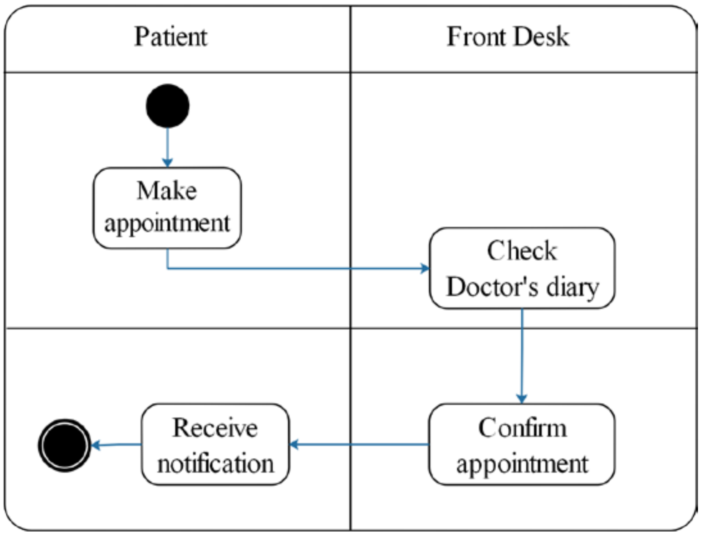


Fig. 6. Activity diagram for making an appointment

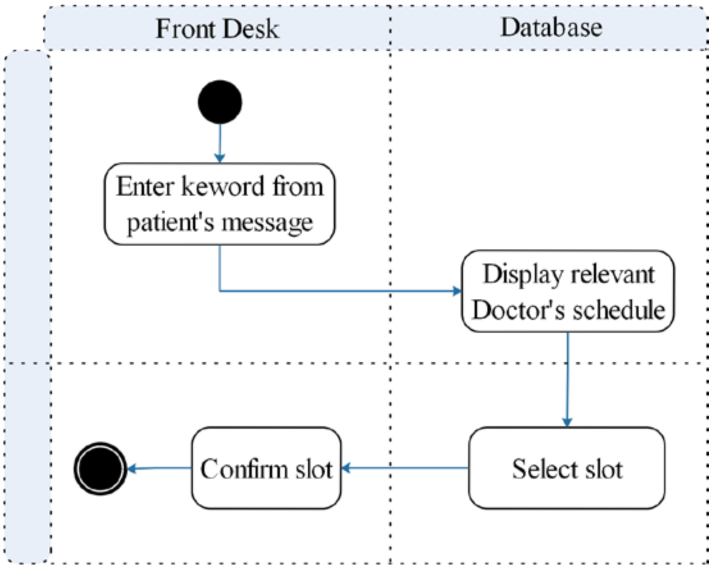

Front Desk
Database
Enter keword from patient's message
Display relevant Doctor's schedule
Select slot
Confirm slot

Fig. 7. Activity diagram for checking the doctor’s schedule

Appointment confirmation is made prior to the patient’s visit to the hospital. The front desk administrator confirms the booking through liaison with the doctor. A confirmation SMS is automatically sent to the patient through the patient registry device. The patient may reject an appointment, and the search process for a suitable booking continues until a favourable slot is found. The activity diagram is illustrated in **Fig. 8.** whilst **Fig. 9.** illustrates the sequence diagram for booking an appointment.

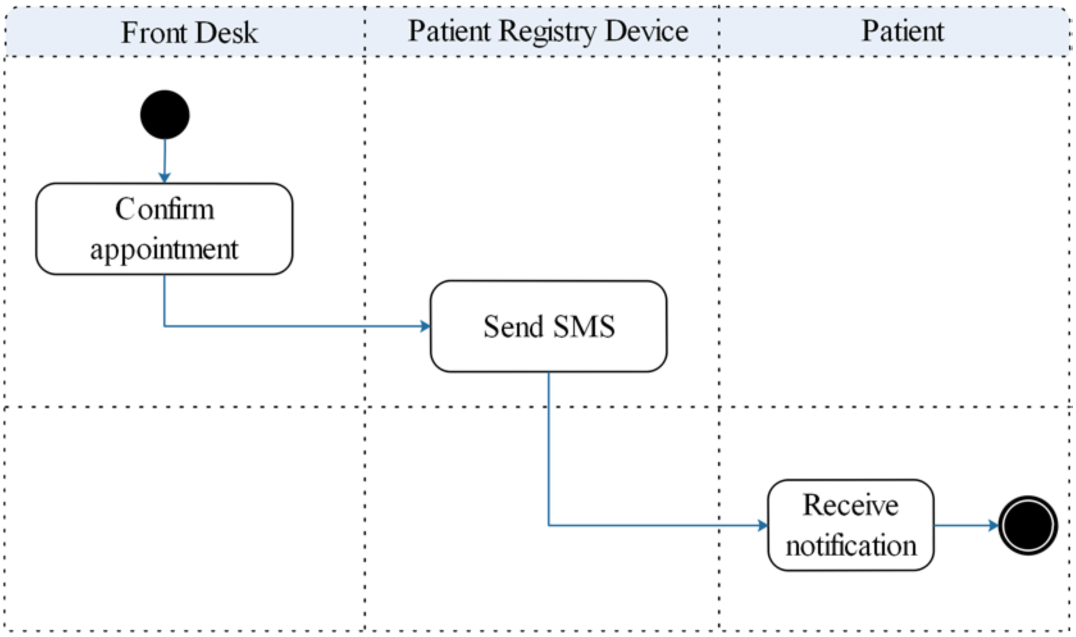


Fig. 8. Activity diagram for appointment confirmation

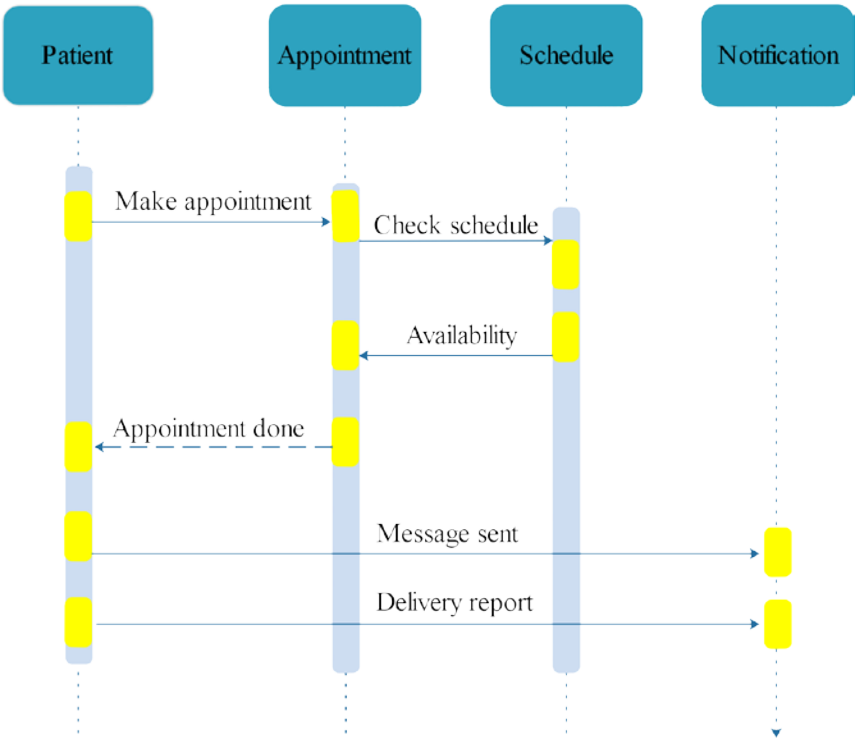


Fig. 9. Sequence diagram for making an appointment

Fig. 10. shows the physical database model for the system. The main entities are the patient, the front desk administrator, the doctor/physician, appointment/booking, the specialty, and the notification. The hardware configuration for the SMS based system is shown in **Fig. 11.**

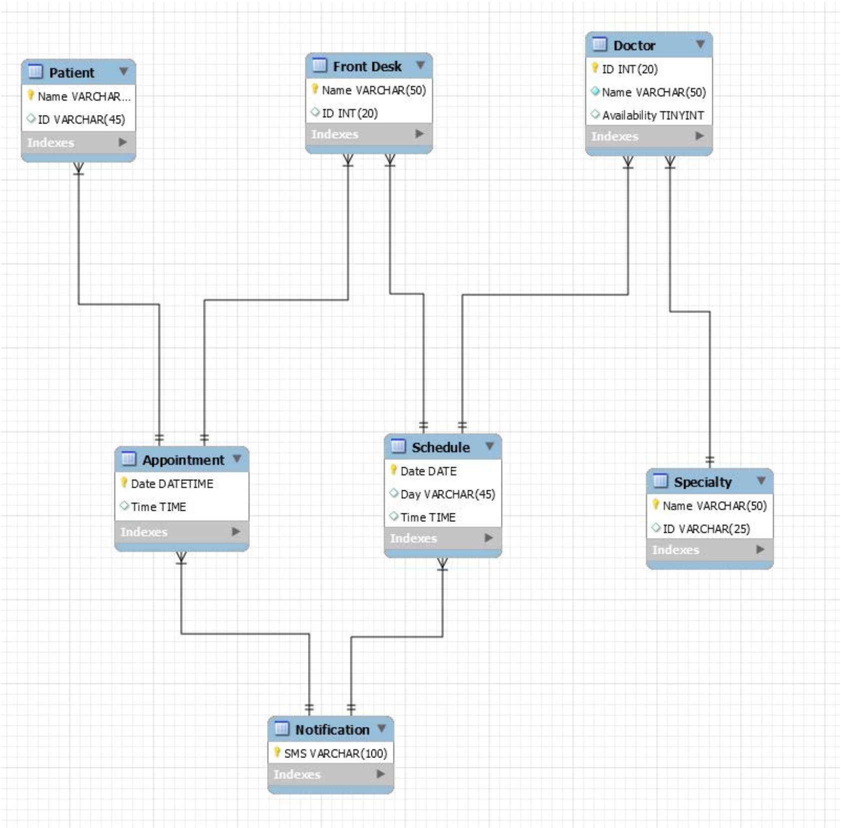


Fig. 10. Physical model for SMS based appointment system

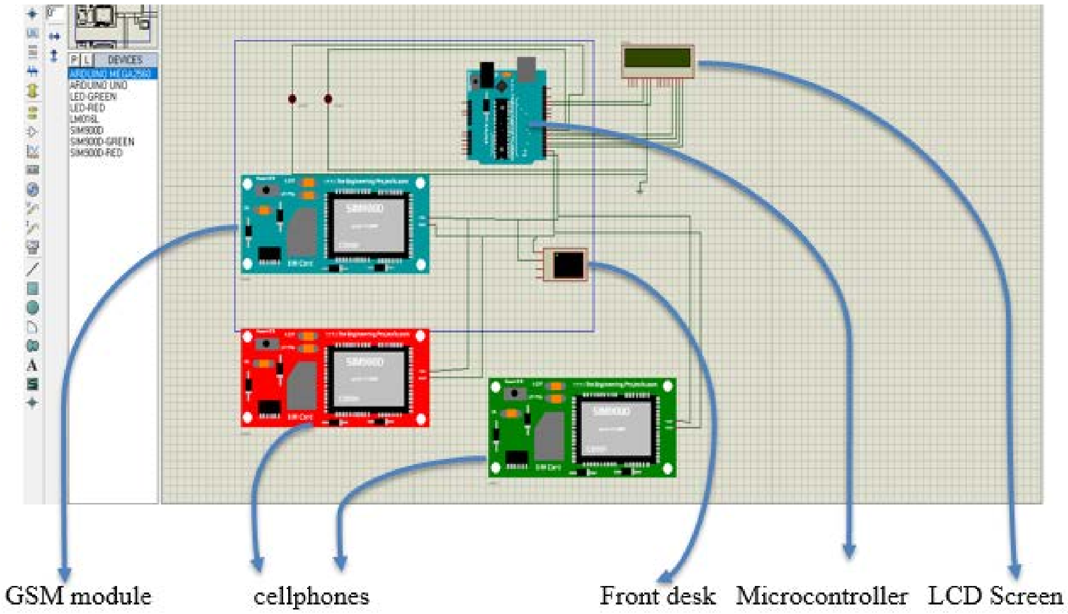


Fig. 11. System hardware configuration

7. Conclusions

The proposed system has the capability to drastically reduce waiting times. This is possible through the slotting of patients at their specific times. Patients only go to the hospital at their allocated time slot without having to wait for a longer time before being attended to. The SMS appointment system is the most optimal and effective system that can be used in remote areas and patients of all classes because it does not need an Internet connection or the use of smartphones. The benefits of implementing this technology cut across a wide spectrum of patients and staff involved in the scheduling process, such as administrators and doctors who will be able to conduct their tasks more efficiently and accurately. Patients have the ability to book their appointments and reservations quickly and more conveniently.

There is a wide range of improvements to the proposed system, as outlined below:

- Accommodate patients cancelling appointments;
- Integrate and customize it to allow patients to choose their own time slots. This will increase the flexibility of the system,
- Allow for rescheduling of appointments. Rescheduling will enable the hospital to give the available slots to the next patient in line if a patient is unable to come on their scheduled time and wishes to reschedule to another date or time;
- Adding sound to the announcement system. The use of a sound to announce the next patient in line will ensure that the patient is aware that it is their time to get a service. Patients may not pay much attention to the LCD screen, which might lead to a minor delay and increase idle time.

8. Acknowledgment

Acknowledgments go to the case study hospital authorities and the responsible ministry for paving the way for the research.

References

[1] Malik S, Bibi N, Khan S, Sultana R, Rauf SA. Mr. Doc: A doctor appointment application system. arXiv preprint arXiv:1701.08786. 2017 Jan 17.

[2] Rinder MM, Weckman G, Schwerha D, Snow A, Dreher PA, Park N, Paschold H, Young W. Healthcare scheduling by data mining: Literature review and future

directions. Journal of Healthcare Engineering. 2012 Sep 1;3(3):477-502.

[3] Rau CL, Tsai PF, Liang SF, Tan JC, Syu HC, Jheng YL, Ciou TS, Jaw FS. Using discrete-event simulation in strategic capacity planning for an outpatient physical therapy service. Health Care Management Science. 2013 Dec 1;16(4):352-65.

[4] Chen PS, Robielos RA, Palaña PK, Valencia PL, Chen GY. Scheduling patients' appointments: Allocation of healthcare service using simulation optimization. Journal of healthcare engineering. 2015 Jan 1;6(2):259-80.

[5] Gupta D, Denton B. Appointment scheduling in health care: Challenges and opportunities. IIE transactions. 2008 Jul 21;40(9):800-19.

[6] Conrad M. Patient waiting time and associated factors at the assessment center, general out-patient department Mulago Hospital Uganda. Kampala, Uganda: Makerere University. 2013 Aug.

[7] Landi S, Ivaldi E, Testi A. Socioeconomic status and waiting times for health services: An international literature review and evidence from the Italian National Health System. Health Policy. 2018 Apr 1;122(4):334-51.

[8] Landi S, Ivaldi E, Testi A. Socioeconomic status and waiting times for health services: current evidences and next area of research. Health services insights. 2019 Aug; 12: 117863 29198 71295.

[9] Siddharthan K, Jones WJ, Johnson JA. A priority queuing model to reduce waiting times in emergency care. International Journal of Health Care Quality Assurance. 1996 Sep 1.

[10] De Souza LB. Trends and approaches in lean healthcare. Leadership in health services. 2009 May 1.

[11] Olorunsola SA, Adeleke RA, Ogunlade TO. Queueing analysis of patient flow in Hospital. Department of mathematical sciences, Ekiti State University of Ado Ekiti, Ekiti State, Nigeria. 2014.

[12] Wanyenze RK, Wagner G, Alamo S, Amanyire G, Ouma J, Kwarisima D, Sunday P, Wabwire-Mangen F, Kamya M. Evaluation of the efficiency of patient flow at three HIV clinics in Uganda. AIDS patient care and STDs. 2010 Jul 1;24(7):441-6.

[13] Safdar K. *Better queue management in a busy public hospital of a developing country without appointment system: an application using data envelopment analysis* (Doctoral dissertation, Aston University).

[14] Burungale S, Kurane K, Mhatre S, Vora D. Patient Queue Management System. International Journal of Engineering Science Invention (IJESI). 2018;7(2):39-41.

[15] Habibi MR, Mohammadabadi F, Tabesh H, Vakili-Arki H, Abu-Hanna A, Eslami S. Effect of an Online Appointment Scheduling System on Evaluation Metrics of Outpatient Scheduling System: a before-after MulticenterStudy. Journal of medical systems. 2019 Aug;43(8):1-9.

[16] Walters, B. A., Danis K, editors. Patient Online at DartmouthHitchcock– Interactive Patient Care Web Site. AMIA Annu Symp Proc. 2003:1044, 2003.

[17] Grain H, editor. Patients' adoption of the e-appointment scheduling service: A case study in primary healthcare. Investing in E-Health: People, Knowledge and Technology for a Healthy Future: Selected Papers from the 22nd Australian National Health Informatics Conference (HIC 2014); 2014: IOS Press.

[18] Wang, W., and Gupta, D., Adaptive appointment systems with patient preferences. Manuf Serv Oper Manag. 13(3):373–389, 2011.

[19] Cao, W., Wan, Y., Tu, H., Shang, F., Liu, D., Tan, Z., Sun, C. et al., A web-based appointment system to reduce

waiting for outpatients: A retrospective study. BMC Health Serv Res. 11:318, 2011.

[20] Maeder A, Martin-San chez F, editors. Patients' perceptions of web self-service applications in primary healthcare. Health Informatics: Building a Healthcare Future through Trusted Information: Selected Papers from the 20th Australian National Health Informatics Conference (HIC 2012); 2012: IOS Press.

[21] Gupta, D., and Denton, B., Appointment scheduling in health care: Challenges and opportunities. IIE Trans. 40(9):800–819, 2008.

[22] Paré, G., Trudel, M.-C., and Forget, P., Adoption, use, and impact of e-booking in private medical practices: Mixed-methods evaluation of a two-year showcase project in Canada. JMIR Med Inform. 2(2), 2014.

[23] Cayirli, T., Veral, E., and Rosen, H., Designing appointment scheduling systems for ambulatory care services. Health Care Manag Sci. 9(1):47–58, 2006.

[24] White, D. L., Froehle, C. M., and Klassen, K. J., The effect of integrated scheduling and capacity policies on clinical efficiency. Prod Oper Manag. 20(3):442–455, 2011.

[25] Robinson, L. W., and Chen, R. R., A comparison of traditional and open-access policies for appointment scheduling. Manuf Serv Oper Manag. 12(2):330–346, 2010.

[26] AJAYI OO, AKINRUJOMU OS, DASO OS, Paulina AO. A Mobile Based Medical Appointment And Consultation (Mmac) System. *International Journal of Computer Science and Mobile Computing*, Vol. 8, Issue 5, May 2019, ISSN 2320-088X, pp. 219-233

[27] Akinode JL, Oloruntoba SA. Design and Implementation of a Patient Appointment and Scheduling System. International Advanced Research Journal in Science, Engineering and Technology, Vol. 4, Issue 12, December 2017. ISSN (Online) 2393-8021, Department of Computer Science, Federal Polytechnic Ilaro Nigeria. 2017 Dec.

[28] Zhang X, Yu P, Yan J, Hu H, Goureia N. Developing an online patient appointment scheduling system based on web services architecture. InChinese Academy of Sciences EET ALAPAMI 2012 Conference Proceedings 2012.

[29] Tufail, Maryam. "Online polyclinic and database management system." (2018).

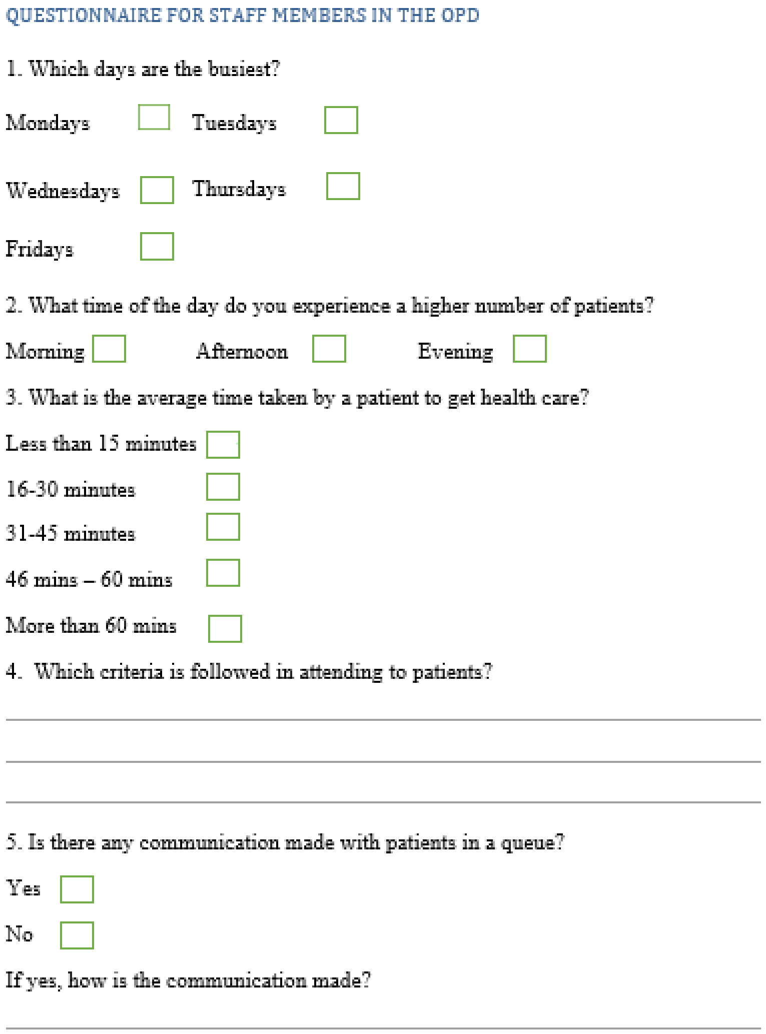
[30] van Brenk D. Reducing Waiting Times In The Pre-Anaesthetic Clinic Of Vu University Medical Center. Master Thesis university of Twente. 2016.

[31] ROSSEL G, MANNA A. A Big Data Modeling Methodology for NoSQL Document Databases. DATABASE SYSTEMS.:37.

Norman GWANGWAVA is a professional Engineer with experience from industry and academia. He is currently a Senior Lecturer at the Botswana International University of Science and Technology (BIUST), department of Mechanical, Energy and Industrial Engineering. Research interests are in; Reconfigurable Manufacturing Systems (RMS), Cyber-Physical Production Systems (CPS), Collaborative Product Design and Closed Loop Life-Cycle Systems, Manufacturing Information and Database Systems. He holds a DTech in Industrial Engineering from Tshwane University of Technology, South Africa and a Master of Engineering in Manufacturing Systems and Operations Management from the National University of Science and Technology, Zimbabwe. He is a member of the SAIIE-ZA and ZIE-ZW.

Kgalalelo D. NTESANG Industrial and Manufacturing Engineering graduate from the Botswana International University of Science and Technology (BIUST). Has high grasp in Production Planning and Control, Six Sigma, Business Process Mapping Models, Supply Chain Engineering, Systems Modeling and Simulation, and Enterprise Resource Planning.

Appendices



QUESTIONNAIRE FOR STAFF MEMBERS IN THE OPD

1. Which days are the busiest?

Mondays □ Tuesdays □

Wednesdays □ Thursdays □

Fridays □

2. What time of the day do you experience a higher number of patients?

Morning □ Afternoon □ Evening □

3. What is the average time taken by a patient to get health care?

Less than 15 minutes □

16-30 minutes □

31-45 minutes □

46 mins – 60 mins □

More than 60 mins □

4. Which criteria is followed in attending to patients?

____

____

____

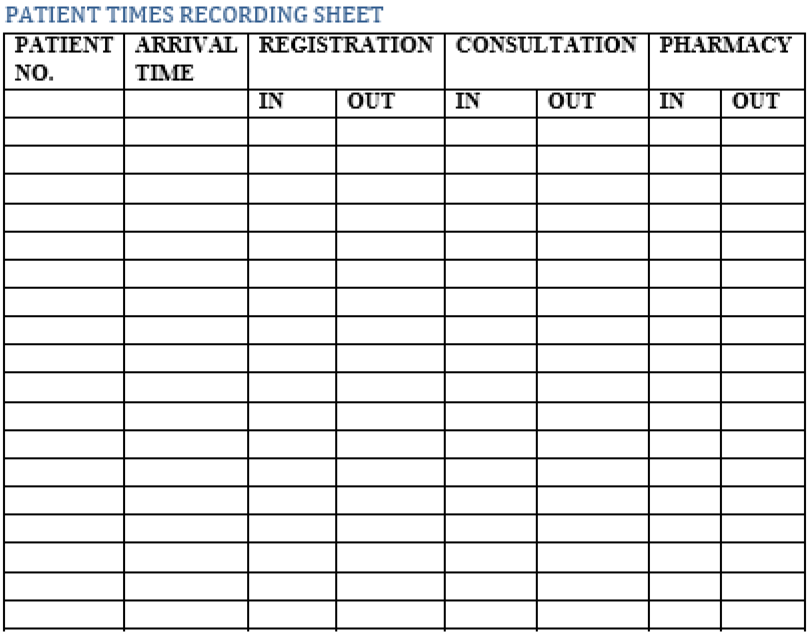
5. Is there any communication made with patients in a queue?

Yes □

No □

If yes, how is the communication made?

____



PATIENT TIMES RECORDING SHEET

PATIENT NO.	ARRIVAL TIME	REGISTRATION		CONSULTATION		PHARMACY	
		IN	OUT	IN	OUT	IN	OUT

Challenges and Ethical Solutions in Using the Chatbot

Cîmpeanu Ionuț-Alexandru
The Bucharest University of Economic Studies
Faculty of Cybernetics, Statistics and Economic Informatics, Romania
ionut.cimpeanu@csie.ase.ro

Artificial intelligence is making its mark on more and more different areas of our lives. No matter what business we are talking about, a smart application offers customers solutions and added value in a more digitalized and automated world. In business, those who do not participate in the development and implementation of innovative solutions exclude the company from the market. However, it uses these necessary IT solutions and ethical challenges related to applicable AI applications, managing and responsibly using the information stored in the applications, the content of the messages, and the way users relate to other users and the chatbot. The paper is structured in four sections. In the Introduction, we talked about the chatbot, about its necessity and usefulness, and about the permanent appearance of some ethical challenges related to the use of the chatbot in different fields of activity. In the next section, we listed a number of ethical challenges that chatbot developers / users face detailing these challenges and setting an example of concrete ethical / unethical approaches. Section III offers solutions to some of the ethical challenges found in the paper. The conclusions provide an overview of the topics addressed in the paper and the directions of perspective in the ethical approach to the issue.

Keywords: *Chatbot, ethics, business, challenges, solutions.*

1 Introduction

Society evolves and with technology and science, the world is revolutionized. These two must keep up with social requirements, with the personalized needs of each field of activity, with the challenges that can be encountered in finding solutions to solve these challenges, with the progress and desire for better people, with making a profit in the activities carried out. 71 years ago, the revolutionary idea of the chatbot promoted by Alan Turing appeared. So, in 1966 the first chatbot was developed, Eliza, who had conversations with people [19]. Since then, new chatbot models have continued to appear, with applicability in different fields of activity and having benefits for people with various health problems, in education - IT solutions that help pupils / students / teachers to collaborate and achieve very good results. , in tourism - offering help and information for buying a flight ticket, booking a stay, canceling a stay, obtaining information about the price of a ticket compared to other flights, information about visiting tourist spots, travel options in the locality, in business - solutions that allow employees / managers to collaborate and exchange professional information in the field of work, to upload applications for leave / employment / retirement / transfer to another salary category, to modify or upload documents based on chatbot data, to have a dialogue with the chatbot to find out the answer to professional tasks [18]. The chatbot is tested and used to provide and gather health information from people and in some cases to provide treatment and counseling services. There are mobile chatbot apps that help people manage depression or anxiety by teaching them self-care and attention techniques. Over time, chatbots have become an innovative technology that promises to change the world by taking over much of human activity or tasks. By using the chatbot in companies / firms, opportunities are opened to improve and streamline the activity in a multitude of fields.

In addition to the many advantages of using the chatbot in different industries, there are also ethical challenges related to the development and implementation of the chatbot, the form and content of communication between the chatbot and users [1]. Solutions to these challenges are found as a lot of research is consulted, read, and studied, studies that address the ethical issues and ongoing challenges that arise and need to be constantly investigated, as society evolves, technology and science move rapidly toward different, innovative, and creative approaches. Ethics helps institutions and individuals by guiding society to what is good and useful to do, to choose thinking, communication, activity using reason and common sense, dignity and well-being of all, to acquire and apply useful skills to distinguish good from evil, when creating a climate of collaboration, trust, solidarity [17]. Respecting a code of norms and rules, a civilized environment of conversation and relationship is created in which each individual feels appreciated and balanced, but also the collective requirements are met. In society, in everything we do, in all areas of activity in which people create and find innovative solutions, ethics is necessary and mandatory because it defines how production relations will evolve in close connection with technical progress and production forces. with which they must adorn themselves. This goal requires constant efforts and all people to understand the need for ethics in everything we do, in everything we think, in our actions, in our perceptions of ourselves and those around us. As ethics is necessary in all fields of activity, ethical norms and rules must be known and applied in business, in the activity of research, development, and implementation of AI.

The topic of research on the ethical challenges of using chatbots in conversation is a current issue that involves many factors in resolving it and holding all those who can make decisions in this regard accountable [23]. The multitude of ethical issues that arise show the need to carefully research this issue and to find solutions related to human-machine conversational processes and their integration with AI principles. When a chatbot develops and appears on the market for consumers, people need to be aware of the features of the IT solution and be warned about the implications of their interaction with the chatbot. Conversations between users and chatbots are not just words used for data transfer and to form a robot conversation pattern. Conversations are strategic processes that have the role of strengthening users' trust in the chatbot, in the information it gives, in the company that transmits this information to users, in reducing uncertainties about people, processes, services, knowledge.

2. Ethical challenges in using the chatbot

A). Responsibility

As with any new technology, the use of the chatbot has ethical challenges and a lot of implications that we need to consider to ensure that the implementation and use of the chatbot in the chosen field of activity are done responsibly. An example of an IT solution that was used irresponsibly and attracted a lot of ethical implications and challenges was the Microsoft Tay chatbot [22]. It aimed to conduct a dialogue with people on Twitter by learning to reply and find answers to their questions based on the conversations they develop. In the first conversations, the way of exchanging remarks was kind, in a friendly way, each user using respectful and polite words when requesting information or when he wanted to have a dialogue with the chatbot. As these conversations evolved and the number of chatbot users increased, the dialogue gained a strong negative character, with insults and harsh words addressed to people of different ethnicities, especially Roma, but also words with strong racist content. This negative change in the way we communicate, the transformation of mood, and the pleasure of speaking in a civilized way was caused by chatbot users who felt the pleasure and need to constantly address such words with offensive and racist content, whether or not they were given the opportunity to talk. Thus, because the chatbot had a program through which it learned to respond mainly using communication models from previous conversations, this chatbot experiment was a failure, the ethical value of the content of the information transmitted from the user to the chatbot, and vice versa, getting an increasingly a negative nuance.

B). Transparency

Another ethical challenge in using the chatbot is transparency. This entails knowledge of how to communicate. It is very important to inform customers when communicating with a real person or a chatbot. Thus, the level of trust in the information obtained in the dialogue is higher when the user communicates with a person. If users find out that a company is using a chatbot to interact with them, they may feel betrayed or even turn against the company. As robots become more human, emotions are implemented, abilities to react in conversation to the user's mood, people's emotional response to these robots increased first, and then suddenly decreased. People have high expectations of the chatbot when it communicates very similarly to human language. However, when a chatbot does not behave exactly like a real person, expectations are replaced by distrust. The question is who owns the information and what is the confidentiality of the chatbot in relation to customers, how confident are users that the information communicated between them and the chatbot will not be passed on and will not reach different parts [20]. If a chatbot makes a shopping list based on orders in previous requests and based on user preferences, does this information belong to the chatbot or user? Can this information be retrieved from the chatbot dialog, or can it be sold or passed on to others? If the answer is YES, then the user must be informed about this, know, and agree. Companies wishing to implement a chatbot must address these issues in the development, implementation, use of the chatbot and be transparent in communicating the conditions of service provision, in the privacy policy, and inform users, from the beginning, if they will converse. with a chatbot or a person.

C). The ability to replace man

However, a great challenge related to the chatbot is reflected in its ability to replace the man in the profession he practices. Researchers point out that all human jobs will be automated in 100 years and that there is a 50% chance that this will happen in the next half century [2]. Analysts and researchers argue that the highest efficiency in an area of activity is achieved when people and chatbots work together [1]. Chatbot is used to answer questions and quickly search a large database in a much more efficient way than a person, while people can retrieve from the chatbot when the situation becomes more complicated and a lot more documentation is needed of to give an answer. The user logs into the application and requests information from the chatbot. Through the initial interaction with customers, the chatbot records the information and the detailed content regarding the request, after which the information is transmitted to the most qualified person in the company, using the skills, knowledge, and professional values according to his qualification, makes a decision on the answer to be sent. Research shows that technological advances in chatbots are reducing unemployment and increasing employment by creating jobs in new sectors [3]. In other words, the idea that AI leads to unemployment and a lack of jobs is being combated. In the future, chatbots will not take jobs from people but the man and the machine will work together, and complement each other to achieve maximum efficiency. Moreover, even though chatbots take over some of the human tasks and some of the jobs from people, the use of chatbots leads to a technological revolution that results in an increase in the employment rate of people in various fields. activity.

Developers can design chatbot models that have a pleasant physical appearance and manners, movements that can be modified, language and dialect of speech that match the cultural environment, the way a user thinks and likes to communicate, including the race, ethnicity, social or economic environment in which he lives and works. This helps to establish relationships between chatbot users by contributing to a good understanding between them and an effective dialogue.

D). Design biases

Using the chatbot brings other ethical challenges. The appearance and behavior of the chatbot, the way it dialogues, the responses it offers to users attract design biases, such as: preference for racial or ethnic communication, background and information in racial or ethnic chatbots [23]. The chatbot knowledge base used to train machine learning algorithms and the way the chatbot generates responses to users suffer from even systematic errors or incorrect results in sending a chatbot message to the user. by privileging one group of users over other groups. This favoring is done from the datasets used to develop the chatbot conversation and may contain problems with missing data, misclassification, and measurement errors, small sample sizes used in determining datasets, and keywords resulting in underestimation and inaccurate predictions responses to users. Implementing the chatbot in different countries and cultures may suffer from these design shortcomings if technology companies do not take into account the demographic characteristics and specific needs of the target user groups. Thus, the developers of the chatbot must also take into account the prejudices and the way of thinking of the human groups for which they develop the chatbot when designing and testing the IT solution. It is necessary to include data from the target population and from different communities in a population that would suffer or could be affected by ambiguities, dysfunctions, irregularities in the design, testing, and implementation of these technologies.

The chatbot works with a certain level of autonomy; it cannot be fully controlled in the way it acts and answers. There is a potential negative risk to people if the chatbot does not adequately address the conversation scenarios in which the implemented system detects potential or safe security risks. We can exemplify these statements if a person converses with a chatbot and during the dialogue they reveal that they have suicide plans and that they want to put them into practice [25]. Also, in such situations patients with mental illness, psychotic symptoms, cognitive disorders or other problems such as depression are included that make it difficult for the chatbot to find an answer, but also alter its knowledge base with negative information. In order to find solutions to the conversation and in the case of these problems mentioned above, the developers of the chatbot must set the limit for the disclosure and disclosure of a user's personal data as well as the intended use of such data. The issue of a possible verification of users before they have access to the conversation with the chatbot is also discussed. As a precaution against users, the chatbot technology must be designed to automatically monitor the risks that may arise in the conversation, as well as the measures that the chatbot may take in the circumstances. Some IT solutions have been designed to meet these requirements as well. So when a situation like the one mentioned earlier appears in the conversation, the chatbot quickly displays help resources, such as a crisis line or an alarm announcing that something is wrong or someone urgently needs help. In these cases, competent and well-trained people should add dialogue procedures and additional information to allow even the user to be contacted and after the dialogue or chatbot can even make a recommendation for assistance for people who have used such expression. of negative words.

E). Confidentiality

Another ethical challenge is confidentiality [21]. When users use a chatbot, their personal data must not be disclosed without their consent. User privacy must be viewed with the utmost responsibility by people who develop, implement, and test chatbots because poor security of these solutions could harm users. The chatbot can collect large amounts of personal data when people talk to it. The possibility of storing this personal data and more sensitive information, such as those related to the health of users, their accounts, finances, e-mail addresses or online accounts, trips or stays by announcing them on social accounts thus so that it can be easily found when owners leave home and buildings are unattended, diminishes the confidentiality that a chatbot can provide to the user by posing serious problems of user security and more rigorous and careful design of robots because it can even affect security and the security of the robot, as well as the developer of the chatbot or the manager / owner of the company that uses the chatbot.

F). Manipulating users perceptions of a particular issue

Other ethical issues encountered in using chatbot-based conversations are reflected in the large amount of information it stores and on the basis of which the chatbot creates templates that provide additional intelligence to chatbot owners. Thus, they could program the chatbot to use this data in the process of manipulating users' perceptions of a particular problem. An example of manipulating users using the chatbot would be to evaluate products or services offered by the company by including in the knowledge base of the chatbot biased information that would favor the company by attracting customers but would disadvantage users by a mass misinformation and a misperception about poor quality products and services, but which are sold or perceived as having great value. This potential imbalance reflected in the informative power of the chatbot could increase the risk for users to interact with different companies through a chatbot. Another example of chatbot manipulation is political propaganda. In online policy research, the most widely used strategy in more than 38 countries is to develop political robots or automated accounts designed to mimic human behavior [15]. These robots were specially designed for the use of social manipulation techniques by spreading false news, political propaganda made during elections in different countries, forming and spreading a false popularity of a personality in social life, spreading stories that denigrate certain people in favor of others, giving political support to people who want to reach high positions without regard to ethical values and fairness.

G). The credibility of the chatbot among users

We also consider the fact that the chatbot is not a moral and independent agent, it does not have the capacity to make moral reasoning, to divide the information received into positive and negative, and to direct to the user only those data that it considers beneficial [5]. Depending on the information stored in the database and on the conversations with users, and customer experiences, the chatbot forms communication models memorizing the way you want to talk and the type messages you need to send to users. The chatbot can improve its customer experience and customer interactions based on the data collected by providing personalized messages tailored to users' needs based on sets of information collected during human-chatbot conversations. However, the chatbot cannot have human qualities such as: judgment, empathy, discretion, it will not be able to secure this information and will not be able to keep it secret from other users by disclosing the data collected to all customers. The chatbot makes decisions but does not make judgments. Its decisions are based on algorithms included in ways that benefit the business owner or chatbot developer. There are also cases where the chatbot risks spreading rumors or misinformation or may even verbally attack people who post personal thoughts and opinions in messages [4]. The chatbot must include protection against the types of people who use obscene language. One solution to these cases would be to apply rules when the chatbot identifies these negative forms of expression and exclude people from the conversation, followed by a final sanction - closing the user's account.

Many models of chatbots imitate man in both communication and physical appearance. We notice a spectacular evolution of the chatbot so that closer and closer to man is possible. The chatbot imitates man, even though it has nothing human in it. The credibility of the chatbot in front of a large number of users also depends on how the chatbot manages to promote and encourage human-robot dialogue by attracting potential interlocutors through secure messages, which include correct and desired information by users, through the trust gained in interactions achieved by reducing or minimizing people's perceptions of the risks of using the chatbot in conversations, by transmitting personal data to robots. The messages of the chatbot must be correct, without mistakes in written or verbal expression, without words not understood by the user, aspects that could affect the human-robot conversational relationship [14]. Studies show that not taking into account the form of the message, the way the chatbot communicates with users through clear and precise statements said or written correctly and with direct address to the meaning of words in the statements expressed, can lead to ambiguity in expression and loss of credibility. [10]. Approaching a proper, pleasant, and balanced tone diminishes the possibility of having a tiring or unpleasant conversation. Conversation patterns in the chatbot's knowledge base must also include information from real facts by which the chatbot gives assurances in the conversation that the data received is correct, secure, and there is no room for misinterpretation. Chatbot developers and companies that want to implement a chatbot must include in the chatbot approach how to classify the different types of conversations according to different purposes, this refers to the percentage of meeting the company's objectives, to the degree of user satisfaction, and to the kind of conversational interaction relationships. Four types of chatbot scheduling conversations have been identified [14]:

- the first type of conversations includes: statements that show initiative and use forms of expression such as: “I will do…”, “I have to do…”, “I will solve…”; statements that show requests and use forms of expression such as: “You will approve…”, “You will request…”; statements that show promise and use forms of expression such as: “We will bring…”, “We will achieve…”, “We will integrate…”, “We will benefit…” which focus users' attention on what could be achieved in the future or what should be done.
- the second type of conversation includes statements used to understand the message and helps the chatbot find the right meaning of a request or problem. In these conversations, evidence is given, hypotheses are made, and information is examined; additional data are requested to help formulate a correct and clear statement; users' emotions and feelings are researched and used; the beliefs and habits of those are observed and analyzed. interrogated.
- the third type of conversation includes statements used for performance using a multitude of exhortations to action, requests, and promises that interact in order to achieve the desired result.
- the fourth type of conversation includes statements used for closing using forms of communication that direct the user to the end of the task, the chatbot ensuring that all participants in the conversational interaction will give positive feedback and be satisfied with the way the dialogue went.

3. Solutions to current ethical issues

While people have the opportunity to be helped by the chatbot when addressing new areas of activity and new services worldwide, chatbot developers need to think responsibly about the security, dignity, and respect that users must constantly maintain to ensure the ethical use and application of technology. One solution by which we can address and come up with solutions to the current ethical challenges associated with using chatbots is to review or think about and develop new codes and principles of professional ethics and practical guidelines to assist in the process of communication between people. Many codes of ethics and practical guidelines in various fields of activity do not address the use of technologies that replace people, namely chatbots. Another solution that I find appropriate and that would bring immediate solutions to these problems is to set up working groups at the international level to review existing principles and ethical guidelines with concrete reports on situations that arise and that disrupt the smooth running of communication and even making recommendations to ensure the ethical use of AI-based tools, including chatbots.

Discrepancies may arise in the relationships and conversations between users and the chatbot in the sense that the chatbot developer has more information about the services it provides or the information it wants to provide to the chatbot users [7]. Here are two ethical issues: a problem of selection and collaboration [6] that occurs when the user cannot know and assess all the skills, knowledge, and information of the person with whom he is in dialogue and who gives him answers to questions leading to a certain situation resolution; the second ethical issue is that the information transmitted by the chatbot affects users' perceptions of the quality of a product or service [8]. Based on the research and documentation, we have identified possible solutions to solve these problems:

- achieving a clear and precise information of the user before solving a task or reaching a stable conclusion by carrying out with priority a private exchange of information with details about the requested product / service requested / necessary information / customer need. This private exchange of information must involve both parties - the user and the chatbot - in a responsible and useful interaction for both parties.
- achieving and developing a communicative and conversational direction that promotes a common culture based on meanings and meanings of situations / events / circumstances discussed or may occur in conversations to avoid misunderstanding messages and failing to achieve dialogue objectives.
- the development of a communication in which the user is put first, to be satisfied with the way he is informed and the quality of the information received, as well as the promotion of a system through which the user can provide feedback for how the chatbot found quickly and effective response to the requested problem.

Studies based on users' conversation with the chatbot show that IT solutions that have implemented a chatbot for conversation are required to reveal their identity immediately and present themselves from the beginning of the dialogue so that users can clearly understand whether they are talking to a robot or with one person. Finding out the identity of the information agent helps to stimulate the conversation so that one moves to a position of dialogue made on an equal footing, to a respectful and sincere listening on both sides [9]. Given all the above, I think it would be helpful if the chatbot revealed its identity from the beginning of the conversation with the user by making a statement similar to it: “Hello! I am Ioana chatbot and I'll try to find an answer to your problem. How can I help you?"

Chatbot developers must include specific mental models in its knowledge base to understand the messages through which meanings are interpreted and assigned meanings to the reality about which they dialogue [10]. These models help the chatbot process information by sending messages about how the world works, reports on human values and beliefs, psychological dispositions related to emotions a person goes through or feels, and interpretations of original messages based on similar words. used in previous expressions. The social and cultural context of chatbot users influences the conversation through the personal opinions of registered chatbot users through cooperation between users in order to achieve a goal or find a solution [11].

Studies and research that discuss the ethical issues of communicating with chatbots also talk about the need to immediately declare the purpose of the conversation and about communicating the intention of the chatbot in the dialogue with the user [12]. There are two intentions in users' conversations with the chatbot:

- the operational intention is characterized by the chatbot's use of a pleasant, low, understanding, conciliatory tone, by using words of understanding, apology, thanks, balance in conversation. For example: "I'm glad to hear this", "Greetings on this beautiful day!", "I'm glad to meet you", "Thank you for your help". The operational intention also consists in sending the user some informative materials to help him find an answer to the problem for which he used the chatbot, such as: graphics, links, emails, reports, diagrams, tables [14].

- the strategic intention is characterized by the orientation of the questions and the direction of the answers in order to collect new data, ideas that have not been discussed so far, and different perspectives to address issues discussed by users. In this case, the tone used by the chatbot is completely different, being provocative, insistent, insistent, with the use in expression of statements that represent questions based on which new information can be discovered and collected. For example: "What do you want to know?", "What do you mean?", "What do you intend to achieve?", "Why are you interested in this information?", "What is the use of the information found?" [13].

Other studies that discuss the ethical issues of communicating with chatbots talk about three sets of questions that should be included in the development of chatbot conversations:

- the first set of questions must be asked before a user has access to public conversations and contains mental patterns, objectives, and intentions of users. The chatbot asks questions at the beginning of the interaction with a new user, the dialogue taking place only between the chatbot and this user. Thus, the chatbot can discover the mental models of the users and could frame the type of conversation in a certain model that corresponds or does not correspond to the conversation ethics [10]. These questions and answers should not be included in the knowledge base of the chatbot, they have the role of identifying and classifying the conversation and the user's intention, and based on them the user has access to the account or is not allowed access because the intention contravenes ethical norms of conversation. By analyzing the initial and individual chatbot- user conversation, the main task of the chatbot is to find out the information about the user, his role in possible conversations, the goals of the conversation, intentions, and results expected by the user. None of the parties involved in the conversation should be considered only a source of information, but a discussion partner from which each party has something to gain and both partners are responsible and accountable for the information given.
- the second set of questions contains ethical principles and norms extracted from the developer from the ethical conversation literature. These questions help increase the chatbot's credibility and find answers to the ethical challenges the chatbot faces in dialogue with users. The second set of questions is asked by the chatbot users in order to find out if they want to address a broader topic in the conversation or want to focus on an issue. In forming messages, the chatbot will take into account both the unique contribution of each user, as each person thinks differently about solving a problem, and the statements and thinking of a group of people who have a greater share in the use of different patterns. conversation. Identifying these conversation patterns aims to avoid possible conversations with unethical, negative content that would affect users' trust in other users and even users' trust in the chatbot. Studies show that users' trust in other users and in the chatbot can be created through a process of discovery through dialogue with others and the quality of the dialogue [17]. The quality of dialogue can be improved by the developer by including in the way the chatbot expresses empathic statements, with understandable and pleasant forms of addressing, attractive to users, or with statements in which the chatbot expresses its agreement or disagreement for a form of communication or a topic of conversation [24].
- the third set of questions contains conversation rules that must be made known to users by the chatbot from the

beginning of a conversation. Studies show that the way a conversation will take place, as well as its content, has a lot to do with the social and cultural environment from which the users come. The third set of questions helps developers achieve security and privacy for chatbot - users. Developing a conversation based on ethical rules and regulations allows the developer to implement an IT solution in which the chatbot can find ideas and solutions to share with users. At the same time, users feel free to dialogue with each other or with the chatbot, but have the obligation to follow ethical rules and regulations in order to use the application [18].

4. Conclusion

In this paper we talked about chatbot, its need and usefulness in various industries, we addressed issues related to ethics and ethical challenges that developers and users of smart solutions constantly face, we found solutions to some of these challenges. The research activity is based on the multitude of:

- studies;
- articles;
- journals;
- books;
- courses;
- tutorials.

The results of these studies concluded in a number for arguments of the need and usefulness of business ethics, in the use of chatbots in different sectors of activity through individual and collective responsibility, by applying ethical norms and rules and by the ethical approach of each problem in solving work tasks.

The introduction presents the usefulness and necessity of the chatbot and the need for customization in the construction of a chatbot so that it fulfills particular tasks depending on the needs of the company, employees, depending on the needs of the market, and the company's objectives. Also, here we talk about a series of ethical challenges that arise as society evolves and the degree of use of these applications increases. These challenges are different and present for both the developer and the users of the chatbot, which can be overcome, and solutions can be found to improve or solve all problems through personalized, different approaches, depending on the social and cultural environment in which come from users, from the communication models implemented in the knowledge base of the chatbot and on the basis of which it formulates answers to users, from the observance of ethical norms and rules in the development and implementation of the chatbot, in the conversational process between users or between users and chatbot.

In Section II, Ethical Challenges in Chatbot Use, we have listed and developed some of the ethical challenges that arise in using chatbots, giving concrete examples of situations in which these problems have manifested and constantly coming up with examples. of studies from which we learned the concrete situations or information about the problem.

In Section III, Solutions to Current Ethical Problems, we have gathered solutions and ideas that can be successfully applied by always referring to the accessed study or the article / paper studied.

The research topic on ethical challenges related to the use of chatbots in conversation is a topical issue, which meets the needs of solving situations in which unethical, aggressive conversations, verbal violence, ethnic / racial discrimination are manifested in user conversations, and permanent adherence to norms and principles, to ethical values and rules becomes a necessity. The ethical approach also applies to the chatbot, which stores a large amount of data, including personal data. For this data that does not remain in the system and is not used only by the person who transmitted the personal data to the chatbot, there is the issue of confidentiality and security of this data that becomes known by all users of the application.

As society evolves and with it, technology and science, new ethical challenges arise that can be improved or resolved by laws that give direct applicability to ethical norms and rules by sanctioning institutions / people who break the rules. These challenges can also be solved by constantly reporting each individual / community to the norms, rules, and ethical principles, and by the deliberate adoption by each person of an ethical behavior and attitude in any circumstance.

References

[1] Von Malitz, B. (2016). Why chatbots won’t replace humans. Retrieved January 15, 2018, Available: https://memeburn.com/2016/04/chatbots- wont-replace-humans/

[2] Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When will AI exceed human performance. Evidence from AI Experts. CoRR.

[3] Stewart, I., De, D., & Cole, A. (2015). Technology and People: The great job- creating machine. Deloitte, London: UK.

[4] Radziwill, N., & Benton, M.C. (2017). Evaluating quality of chatbots and intelligent conversational agents. Computer and Society. Available: https://arxiv.org/pdf/1704.04579.pdf

[5] Marr, B. (2018). Machine Learning In Practice: How Does Amazon’s Alexa Really Work? Forbes, Available: https://www.forbes.com/sites/bernardmarr/ 2018/10/05 /how-doesamazons-alexa- really-work/#465dde371937

[6] Nayyar, P. R. (1990). Information asymmetries: A source of competitive advantage for diversified service firms. Strategic Management Journal, 11(7), 513–519. Available: https:// www.jstor.org/stable/2486325.

[7] Mishra, D. P., Heide, J. B., & Cort, S. G. (1998). Information asymmetry and levels of agency relationships. Journal of Marketing Research, 35(3), 277-295. Available: https://doi.org/ 10.2307/3152028

[8] Holmstrom, B. (1979). Moral hazard and observability. Bell Journal of Economics,1(1),74-91. Available: https://doi.org/10.2307/3003320

[9] Beech, N., MacIntosh, R., & MacLean, D. (2010). Dialogues between academics and practitioners: The role of generative dialogic encounters. Organization Studies, 31(9-10),1341-1367. Available: https://doi.org/10.1177/0170 84061037439 6

[10] Mengis, J., & Eppler, M. J. (2008). Understanding and managing conversations from a knowledge perspective: An analysis of the roles and rules of face-to-face conversations in organizations. Organization Studies, 29(10), 1287-1313. Available: https://doi.org/10.1177/01708406070 86553

[11] O’Neill, A., & Jabri, M. (2007). Legitimation and group conversational practices: Implications for managing change. Leadership & Organization Development Journal, 28 (6), 571-588. Available: https://doi.org/10.1108/014377307107 809 94

[12] Von Krogh, G., & Roos, J. (1995). Conversation management. European Management Journal, 13(4), 390-394. Available:

https://doi.org/10.1016/0263-2373(95)00032-G

[13] Skordoulis, R., & Dawson, P. (2007). Reflective decisions: The use of Socratic dialogue in managing organizational change. Management Decision, 45(6), 991-1007. Available: https:// doi.org/10.1108/002517407107 62044

[14] J. D., & Ford, L. W. (1995). The role of conversations in producing intentional change in organizations. Academy of Management Review, 20(3), 541-570. Available: https:// doi.org/10.2307/258787

[15] Bradshaw, S., & Howard, P.N. (2018). Challenging truth and trust: A global inventory of organized social media manipulation. Oxford Internet Institute. Available: https://comprop.oii. ox.ac.uk/research/cybertroops2018/

[16] Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. Interactions, 24(4), 38-42.

[17] Bowen, S. A. (2016). Clarifying ethics terms in public relations from A to V, authenticity to virtue: BledCom special issue of PR review sleeping (with the) media: Media relations. Public Relations Review, 42(4), 564-572. Available: https://doi.org/10.1016/j. pubrev.2016.03.012

[18] Hammond, S. C., & Sanders, M. L. (2002). Dialogue as social self-organization: an introduction. Emergence: Complexity and Organization, 4(4), 7-24. Available: 10.emerg/10.17357.82a56a167e96cf 9 c286a13c30ad9f893.

[19] Weizenbaum J 1966 Eliza—A computer program for the study of natural language communication between man and machine. Commun. ACM 9(1): 36-45 Available: https://dl.acm.org/doi/10.1145/365153. 365 168

[20] David D Luxton (2020), Ethical implications of conversational agents in global public health. Available: https://www.ncbi.nlm.nih.gov/pmc/arti cles /PMC7133471/

[21] Grazia Murtarelli, Anne Gregory, Stefania Romenti (2020), A conversation- based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots. Available: https://www.sciencedirect.com/science/ arti cle/abs/pii/S0148296320305944

[22] Hayco de Haan (2018), Chatbot Personality and Customer Satisfaction. Available: https://research.infosupport.com/wp-content/uploads/Chatbot-Personality-and- Customer-Satisfaction-Bachelor-Thesis- Information-Sciences-Hayco-de-Haan.pdf

[23] Fernando Fogliano, Fernando Fabbrini, Andre Souza, Guiherme Fidelio, Juliana Machado, Rachel Sarra (2019), Edgard, the Chatbot: Questioning Ethics in the Usage of Artificial Intelligence Through Interaction Design and Electronic Literature. Available: https://link. springer.com/chapter/10.1007/ 978-3-030-22219-2_25

[24] Stoll, B., Edwards, C., & Edwards, A. (2016). “Why aren’t you a sassy little thing”: The effects of robot-enacted guilt trips on credibility and consensus in a negotiation. Communication Studies, 67(5), 530-547. Available: https://doi.org/10.1080/ 10510974.2016.1215339

[25] Stewart, J. (1994). The welfare implications of moral hazard and adverse selection in competitive insurance markets. Economic Inquiry, 32(2), 193–208. Available: https://doi.org/10.1111/j.1465-7295.1994.tb01324.x

Ionuț-Alexandru Cîmpeanu graduated from the Faculty of Economic Cybernetics, Statistics and Informatics within the Bucharest University of Economic Studies in 2018. He holds a master's degree in economics by graduating from the E-Business master's program within the same faculty in 2020. In the same year, he started the research program as a doctoral student, at the Bucharest University of Economic Studies, specializing in Economic Informatics. He currently works as a programmer at TotalSoft and, in parallel, works in doctoral research at the Bucharest University of Economic Studies, specializing in Economic Informatics, participating in national and international conferences, publishing articles in specialized journals, and participating in scientific sessions on research topics in the field. His work focuses on analyzing the quality of software applications and developing these applications.