



A Proactive Chatbot Framework Designed to Assist Students based on the PS2CLH Model

**A thesis submitted to London Metropolitan University for the
degree of Doctor of Philosophy in Computing**

Arlindo Djassi Diogo de Almada

December 2021

Acknowledgments

I am grateful to my God for giving me this opportunity to fulfil my research degree dream. Unfortunately, I come from a country where the level of poverty and the educational system are among the ten worst in the world. I thank my parents, João Almada and Teresinha Diogo, for the unconditional love they have given me throughout my life. I also want to thank my brothers and my sister who have been a constant presence: thanks to Bula, Amilcar and Ginga. I express gratitude to my Grand Uncle Frank Borges for the precious guidance and financial support when the scholarship ceased. My sincere gratitude also goes to Marta, my fiancée, for her emotional support.

I am grateful to my tutors Dr Qicheng and Dr Preeti for their teaching and insights throughout the project. I have benefited from Dr Qicheng, whose patience and wisdom illuminated my journey and who asked questions that forced me to rethink and investigate more deeply. Thank you, Dr Preeti, for your observations and concern, and especially for your touch of magic as we wrote our published paper “*PS2CLH: A Learning Factor Model for Enhancing Students’ Ability to Control Their Achievement*”.

I would also like to express my thanks to the Universidade Católica de Angola for funding my studies and for being the first to believe in this research project. Thank you for authorising the collection of student data for my research and for helping in other ways within the limitations inherent in the project.

Publication

Almada, A., Yu, Q., Patel, P. 2019. *PS2CLH: A Learning Factor Model for Enhancing Students' Ability to Control Their Achievement*. Tokyo, ACE2019. (Publication)

Almada, A., Yu, Q., Patel, P. 2022. *Visual representation of the students' controllable factors, which affects their performance, using the PS2CLH model*, The 2nd IAFOR Conference on Education Research & Innovation, Washington DC, USA. (Conference presentation)

Almada, A., Yu, Q., Patel, P. 2022. *Representation of the Student's Controllable Performance Features Based on PS2CLH Model*, Barcelona Conference on Education, Barcelona, Spain. (Publication)

Almada, A., Yu, Q., Patel, P. 2023. *Proactive Chatbot Framework Based on the PS2CLH Model: An AI-Deep Learning Chatbot Assistant for Students*. In: Arai, K. (ed.s) *Intelligent Systems and Applications*. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 542. Springer, Cham. https://doi.org/10.1007/978-3-031-16072-1_54 (Publication)

Table of Contents

Acknowledgments	
Publication	
List of Equations	
List of Tables	
List of Figures	
Glossary	
Abstract	
Chapter 1 Research Overview	1
1.1 Introduction.....	1
1.2 Research Problem	2
1.3 Research Questions.....	4
1.4 Aims / Objectives of the study.....	5
1.5 Significance.....	6
1.6 Research Contributions.....	7
1.7 Research Scope	8
1.8 Practicalities.....	8
1.8.1 Location	8
1.8.2 General requirements	8
1.8.3 Feasibility.....	8
1.8.4 Achievability.....	8
1.9 Ethical considerations	9
1.9.1 Confidentiality of Student Data	9
1.9.2 Informed Consent.....	9
1.9.3 Lack of regulations	10
1.10 Structure of the Thesis	10
1.11 Summary.....	11
Chapter 2 Literature Review	12
2.1 Introduction.....	12

2.2 Student Assistance – An Overview.....	12
2.2.1 The science behind the ability of the practical Assistant/Coach to improve a candidate’s performance.	13
2.2.2 The Assistant/Coach	15
2.2.3 Different Assistance/Coaching Styles.....	17
2.2.4 Learning style, Multimodality and Personality Types.....	22
2.2.5 The use of chatbots in Universities	27
2.2.6 Why should we use an AI chatbot to assist students?.....	28
2.3 Artificial Intelligence, Natural Language Processing, Deep Learning and Chatbots	30
2.3.1 Artificial Intelligence	30
2.3.2 A brief view of Knowledge Discovery	33
2.3.3 Natural Language Processing	34
2.3.4 Deep Learning.....	37
2.3.5 Similarity Text Measurement	40
2.3.5.1 Text Distance	40
2.3.5.2 Text Representation	42
2.3.6 Chatbots	48
2.3.7 Chatbots frameworks: A short overview	52
2.4 Big Picture of the Proactive Chatbot Framework	54
2.5 Summary	54
Chapter 3 Methodology	57
3.1 Introduction	57
3.1.1 Quantitative Research	58
3.1.2 Qualitative Research	58
3.1.3 CRISP-DM methodology.....	59
3.1.4 Scrum Methodology.....	60
3.1.5 Research Design or Thesis Methodology	61
3.1.6 Research Approach	63
3.2 Methodology to find the correlation among the PS2CLH model factors	63
3.3 The method for building the framework of the proactive chatbot for students.....	67
3.4 Summary	70
Chapter 4 The PS2CLH model: Results and Discussion.....	71
4.1 Introduction	71
4.2 Proposed PS2CLH’s model.....	73

4.3	Psychology and Self-responsibility	75
4.4	Sociology and Communication	77
4.5	Learning and Health & wellbeing	78
4.6	The Experiments: Setup, Results and Discussion	79
4.6.1	First Experiment, without Intervention	80
4.6.2	Second Experiment, with Intervention	89
4.7	Summary	106
Chapter 5 The proposed Proactive Chatbot Framework Designed to Assist Students based on the PS2CLH model, Test and Results		107
5.1	Introduction	107
5.2	The Fundamental connection between the PS2CLH's model and the proactive chatbot framework	108
5.3	The proposed Framework design for the Proactive Chatbot for Students based on the PS2CLH model	108
5.4	Wide-ranging Extended Chatbot	110
5.4.1	Inputs/Questions, Data Preparation	111
5.4.2	Embedding, Vector/Matrix	113
5.4.3	Text Similarity Measurement	114
5.4.4	Q&A model, Transformer model.....	116
5.4.5	Assembly parts.....	117
5.5	The Educational Chatbot Ecosystem.....	119
5.5.1	Interaction Facilitator	120
5.5.2	The Profile Customiser	122
5.5.3	Multimodality	123
5.5.4	Rating.....	124
5.5.5	Suggest Factor.....	125
5.5.6	Knowledge Database	128
5.5.7	Profiles	130
5.6	Testing the proactive chatbot framework and extending the BERT chatbot: Results 136	
5.7	Summary	142
Chapter 6 Evaluation, Conclusion and Future works		143
6.1	Evaluation.....	143

6.1.1	Evaluation of the methodology to find the correlation among the PS2CLH model factors.....	143
6.1.2	Evaluation of the Method to Build the Framework Designed for a Proactive Chatbot for Students	148
6.2	Limitations	152
6.3	Implications.....	152
6.4	Conclusion.....	153
6.5	Future works.....	155
	References.....	156
	Appendix.....	174

List of Equations

Equation 1 - <i>Manhattan Distance</i>	41
Equation 2 - <i>Euclidean Distance</i>	41
Equation 3 - <i>Cosine Distance</i>	41
Equation 4 - <i>Wasserstein Distance</i>	41
Equation 5 - <i>JS Divergence</i>	41
Equation 6 - <i>KL Divergence</i>	41
Equation 7 - <i>Jaro Similarity</i>	42
Equation 8 - <i>LCS Similarity</i>	42
Equation 9 - <i>Dice</i>	43
Equation 10 - <i>Jaccard</i>	43
Equation 11 - <i>TF-IDF</i>	43
Equation 12 – <i>Pearson Correlation coefficient</i>	87
Equation 13 - <i>Cosine similarity calculation</i>	115
Equation 14 - <i>Accuracy</i>	141

List of Tables

Table 1 - <i>Glossary</i>	11
Table 2 - <i>Qualitative vs Quantitative research (British Library, 2015)</i>	58
Table 3 - <i>Experiment 1 Variables identified in the Angolan Context</i>	81
Table 4 - <i>Experiment 2 Variables used in the Angolan Context</i>	90
Table 5 - <i>Data Processing using R language</i>	Error! Bookmark not defined.
Table 6 - <i>Small sample of the questions used in the test</i>	138
Table 7 - <i>Results of a small sample of the questions used in the test</i>	140
Table 8 - <i>Accuracy results</i>	141

Table 9 - Comparison between the proactive chatbot and the extending chatbot.....	151
---	-----

List of Figures

Figure 1 - Finding Flow – State Diagram (Csikszentmihalyi, M., 1997)	15
Figure 2 - Different Coaching approaches diagram.....	21
Figure 3 - Four main learning stages	22
Figure 4 - Impact of multimodal learning (Metiri Group, 2008).....	26
Figure 5 - Deep learning subset of Machine Learning	31
Figure 6 - Steps That Constitute the KDD Process (De Martino et al., 2002)	33
Figure 7 - Diagram of Measurement of text similarity	40
Figure 8 - Word2vec 's demonstrate structures	44
Figure 9 - Bert: Pre-training.....	44
Figure 10 - The model structure of DSSM	46
Figure 11 - Architecture-I for matching two sentences (Hu, B., et al., 2014)	46
Figure 12 - multi-view bidirectional long and short-term memory	47
Figure 13 - MatchPyramid on text matching	47
Figure 14 - Proactive Chatbot Diagram, (Almada's Diagram, 2020)	54
Figure 15 - CRISP-DM reference model (IBM Software Business Analytics, 2010).....	59
Figure 16 - Phases of the Research.....	61
Figure 17 - Methodology to find the correlation among the PS2CLH model factors.....	64
Figure 18 - Method used to build the framework designed for a proactive chatbot for students	67
Figure 19 - Logical Diagram, PS2CLH, 2020	73
Figure 20 - PS2CLH perspectives	74
Figure 21 - P2SCL form wizard	82
Figure 22 - SPSS Modeler CRISP-DM Prediction diagram	83
Figure 23 - Partition table	84
Figure 24 - Models results	85
Figure 25 - Predictor Importance Random trees 1 model	88
Figure 26 - Variables' Correlations	88
Figure 27 – Process for Predicting Students' Performance	91
Figure 28 – Importing required libraries.....	94
Figure 29 – Displaying the dataset	95
Figure 30 – Dataset Information.....	95
Figure 31 – Target variable	96
Figure 32 – Training, test and confusion matrix.....	96
Figure 33 – ROC AUC score, Accuracy, Precision, Recall and F1 score.....	97
Figure 34 - Models best result	98
Figure 35 - Deliberate practice: enjoyment vs effort.....	101
Figure 36 - Activity plan prototype (English)	102
Figure 37 - PS2CLH Visual 3D representation	104
Figure 38 - Students represented in the PS2CLH Visual 3D representation.....	105

Figure 39 - <i>Framework for the Proposed Chatbot</i>	109
Figure 40 - <i>Use Case Diagram Proactive Chatbot Framework</i>	109
Figure 41 - <i>Extending BERT</i>	110
Figure 42 - <i>Framework (first part)</i>	111
Figure 43 - <i>Speech Recognition</i>	112
Figure 44 - <i>Document similarity experiment (Neto, J., 2021)</i>	113
Figure 45 - <i>SQuAD Leaderboard Nov 2021</i>	116
Figure 46 - <i>UML Assembly Parts</i>	118
Figure 47 - <i>The second part framework</i>	119
Figure 48 - <i>Assistant adding Q&A</i>	120
Figure 49 - <i>UML Interaction Facilitator</i>	121
Figure 50 - <i>UML Profile's Customizer</i>	122
Figure 51 - <i>UML Rating</i>	125
Figure 52 - <i>UML Suggest Factor</i>	126
Figure 53 - <i>Chatbot's JSON file structure</i>	128
Figure 54 - <i>Database Diagram</i>	129
Figure 55 - <i>Student's Sign Up</i>	131
Figure 56 - <i>Student's Login</i>	131
Figure 57 - <i>Students' Questionnaire</i>	132
Figure 58 - <i>Personality types selection</i>	133
Figure 59 - <i>Students' PS2CLH coordinates</i>	134
Figure 60 - <i>Student's Profile</i>	135
Figure 61 - <i>Assistant's Profile</i>	135
Figure 62 - <i>Lecturer's Profile</i>	136
Figure 63 - <i>Proactive chatbot vs Extending chatbot</i>	139
Figure 64 - <i>Students' average score in standard deviations</i>	147

Glossary

Term	Definition	Term	Definition
AI	Artificial Intelligence	“Grit”	The power of passion and self-control
API	Application Programming Interface	BERT	Bidirectional Encoder Representation from Transformers
ARC-I	Architecture-I for matching two Sentences	BoW	Bag of Words
ARC-II	Architecture-II of convolutional matching model		
CBOW	Continuous Bag of Words	GNNs	Graph Neural Network
CDSSM	Convolutional Latent Semantic Model	GPA	Grade Point Average
CPU	Central Processing Unit	JSON	JavaScript Object Notation
CRISP-DM	CROSS Industry Standard Process for Data Mining	MV-LSTM	Multi-View bi-LSTM
DL	Deep Learning	OOV	Out-Of-Vocabulary
DM	Data Mining	RNN	Recurrent Natural network
DPA	Data Protection Act	SQuAD	Stanford Question Answering Dataset
DSSM	Deep-Structured Semantic Model	SSE	The sum of the squared errors
GPU	Graphics Processing Unit	SVD	Singular Value Decomposition
ICF	International Coach Federation	TF-IDF	Term Frequency-Inverse Document Frequency
KDD	Knowledge Discovery in Databases	UNICEF	United Nations Children’s Fund
KL Divergence	Kullback–Leibler divergence	VARK	Visual, Auditory, Reading/Writing and Kinaesthetic
LDA	Latent Dirichlet Allocation	Q&A model	Question and Answer model
LSA	Latent Semantic Analysis		
LSTM	Long Short-Term Memory		
ML	Machine Learning		
NLP	Natural Language Processing		
PS2CLH	Psychology, Self-responsibility, Sociology, Communication, Learning and Health & wellbeing		
SSE	Sum of the Squared Errors		

Abstract

Nowadays, universities are using new technologies to improve the efficiency and effectiveness of learning and to assist students to enhance their academic performance. In fact, for decades, new ways to convey the information required to teach and support students have slowly been integrated into education. This development started decades ago with the popularity of e-mails and the Web.

A review of relevant literature revealed that learning requires more innovative and efficient technologies to cope with natural learning challenges, highlighting a need for more effective tools to establish the interaction between humans and machines, lecturers and students. In addition, the covid pandemic presented additional new challenges for the collaboration/interaction of lecturers and students at universities. This situation led to a great demand for such tools. Researchers have been trying to develop such tools for decades, and have made good progress, but they are still in their infancy. There has been a significant evolution in computer hardware in the last decade, leading to advances in AI machine learning and Deep Learning which have made tools such as chatbots more usable. However, the efficiency and effectiveness of the chatbot are still insufficient to meet many educational needs. According to our investigation, current chatbots are mainly based on subject knowledge and therefore assist users with answers which take no consideration of their personal circumstances, which is essential in education.

This research aims to design a proactive chatbot framework to assist students. The new chatbot framework integrates students' learning profiles and subject knowledge, making the chatbot more intelligent to improve student learning and interaction more effectively. The research consists of two main parts. The first part seeks to determine the most effective students' learning profiles on the basis of the controllable academic factors which affect their performance. The second part develops a chatbot framework to which students' learning profiles will be applied. Due to the different nature of these two endeavours, a hybrid methodology was used in this research.

The literature on learners' characteristics and the academic factors that affect their performance was reviewed in depth, and this formed the basis for developing a new PS2CLH (psychology,

self-responsibility, sociology, communication, learning and health & wellbeing) model on which an individual's web profile can be built. The PS2CLH model combines the perspectives of psychology, self-responsibility, sociology, communication, learning and health & wellbeing to build a student-controllable learning factor model. This study identifies the impact of students' controllable factors on their achievement. It was found that the model was 94% accurate. In addition, this research raised participant students' awareness of PS2CLH perspectives, which helped learners and educators to manage the factors affecting academic performance more effectively.

A comprehensive investigation, including a survey, showed that the chatbot supported by AI technology performed better and more efficiently in various assistant situations, including education. However, there is still room for improvement in the effectiveness of the education chatbot. Therefore, the research proposes a new chatbot framework assistant which will integrate students' learning profiles and develop components to improve student interaction. The new framework uses knowledge from the PS2CLH model AI - Deep Learning to build a proactive chatbot for assisting students' learning of their academic subjects and their controllable factors that affects students' performance. One of the principal novelties of the chatbot framework lies in the communication facilitator between student-lecturer/assistant. The proactive chatbot applies multimodality to the students' learning process to retain their attention and explain the content in different ways using text, image, video and audio to assist the students and improve their learning experience effectively. Furthermore, the chatbot proactively suggests new controllable factors for students to work on, including related factors that influence their academic performance. Tests of the framework showed that the proactive chatbot demonstrated better question response accuracy than the current BERT (Bidirectional Encoder Representations from Transformers) chatbot and presented a more effective learning method for students.

Chapter 1 Research Overview

1.1 Introduction

It is a continuing challenge for resource-constrained universities to assist all students in their academic work for the duration of their programmes. The learning process can be fraught with difficulties that many students find hard to overcome without assistance. Universities are not able to readily provide one-to-one support. This situation often means that students under-achieve or even drop out, resulting in low-performance metrics for institutions. Education researchers are actively looking for effective interventions to assist students in dealing with their learning process. For instance, Professor John Hattie presents a range of interventions and the impact they have on students (Hattie, J., 2009). He points out that assisting students positively impacts on their performance. Indeed, the past decade has seen a proliferation of innovations in the design of educational interventions which aim to support a diverse range of requirements such as online learning, personalisation and multimodality.

Universities have a significant role to play in ensuring a positive student experience in which engagement, academic success, professional development and self-evaluation are central. This research studies the relationship between students' behaviours and lifestyles and controllable factors which affect their performance, using Artificial Intelligence to create a forecasting model and identify the correlation between these factors and their corresponding essential attributes. The wisdom pyramid model (Rowley, J., 2007), will be used to collect students' data, which will be analysed by statistical methods, data mining and deep learning algorithms to turn it into knowledge. This knowledge can then be used to assist students. The applied knowledge is thus transformed into wisdom which can proactively influence students' learning experience.

The aim of this research is to assist students by using a specific model to develop a personalised student' profile, lecturer profile and assistant profile, and then a framework to facilitate their interaction and assist students in their subjects. Furthermore, applying multimodality to students' learning process is intended to retain their attention and explain the subject content in different ways, i.e. using text, image, video and audio to assist students and effectively enhance their learning experience. Finally, this study proposes a framework for implementing

a readily available and student-usable tool using Natural Language Process - Deep Learning (NLP-DL) based on a specific model to enhance academic performance.

This research recognises the importance of providing students with an Artificially Intelligent chatbot's assistant facility. A conversational agent or chatbot is an automated facility designed to reduce the need for live human interaction. In commercial applications, chatbots aim to enhance the consumer experience by providing efficient and accurate responses to basic queries, thereby allowing a 24/7 response platform while at the same time reducing operational costs. In the past decade, the use of chatbots has increased in a range of industries. Virtual assistants or chatbot applications (such as Siri, Alexa, Google Now and Cortana) have been used in the private sector and in broad areas of the public sector such as telecommunications, healthcare and banking, for transactions, tourism and retail (Hoy, B., 2018). Chatbots are also used in the education sector to assist and support learners in learning a foreign language by conversing with a chatbot, and to enhance critical thinking (Sandu, N. & Gide, E., 2020). Another example is the use of a chatbot for medical students for educational purposes, to improve their learning (Sandu, N. & Gide, E., 2020).

Despite the growing sophistication of chatbots, the state-of-the-art question - answer Transformer model for chatbots has some limitations. For instance, when we try to apply it to more than one page or around ten paragraphs, it decreases in accuracy and takes longer to answer questions and answers with long sequence of text (Bird, J., et al., 2020). It seems that the chatbot's usability has not reached the level required for such tasks. In addition, there appear to be gaps in the assistance offered to individual students at university in terms of providing coherent unifying support for multiple behaviours and lifestyles. To achieve these goals, we start by defining the research problem.

1.2 Research Problem

In summary, the problem identified by this research is the lack of efficient, scalable and inexpensive ways to provide individual assistance to university students in relevant areas and with factors which affect their academic performance and which they can control.

Let us go back to the root of the problem. From the moment a woman conceives, she needs assistance to have a healthy child. When the child is born, she needs assistance to survive and

thrive. As children grow, they need assistance and education from their parents, to learn how to develop a personality and a vision of the world. Upon entering school, they have the assistance and care of their teachers and their parents to learn to read and write. When the student arrives at university, it is assumed that he/she is prepared to continue his/her learning journey independently.

This assumption that students have been prepared to continue their journey does not correspond to the evidence. For instance, the number of students dropping out increases dramatically when they arrive at college and university (Tinto, V., 1975). The constant stress they face leads them to spend hours in non-productive activities, which generates a great deal of frustration. In an attempt to cope with this, they use escape mechanisms such as the excessive use of entertainment, games, sex, TV, social networks and more, which may harm them mentally and physically, and which can all have a negative impact on their learning process (Fenta, A. & Kelkay, B., 2018). An example is developed countries such as Japan, South Korea and India. Even though these countries are among those with the highest quality of education, the high rate of stress and suicide among university students is extremely concerning (Huang, F., 2021). On the other hand, the quality of teaching in countries such as Congo, Nigeria and Angola needs to improve greatly. In both cases, students are clamouring for more significant assistance from universities (Acholonu, I. & Njie, S., 2020). In contrast, in countries such as Finland, Norway and Denmark, students have additional assistance at university and at home, and university dropout rates are low. These are among the countries with the highest quality of education (Elken, M., et al., 2015). These contrasts show that students need support to continue their academic journey.

Most students will need assistance and support in their academic life. Nevertheless, given that individuals' circumstances are often different and that one individual's perception and self-awareness also differ from another individual's, any proposed support mechanism must be able to deal with this variability.

A broad range of research (Akama, E., 2017); (Hattie, J., 2009) (Akama, E., 2017; Hattie, J. 2009) highlights numerous learning factors affecting students' achievement. Among them, there are many factors outside of students' control. Even though students are aware of those factors, they cannot address them. For instance, students cannot choose where they are born, and they may not change other people's decisions. However, students can control learning

factors such as their attitude, psychology, behaviour, self-responsibility skills, and in most cases, their physical health. Furthermore, students are responsible for their communication and how they want to study and learn. With so many controllable factors that affect student performance, it is difficult to individually know students' real obstacles and difficulties. Despite the growth of student support at universities, the individual university student's daily guidance has been an issue. It is costly, and a significant human resource effort is required to provide individual assistance for each student at universities. However, there is a gap in the literature (Akama, E., 2017) exploring how these controllable factors affect learning. Consequently, there is a lack of efficient and inexpensive ways to assist students on those controllable factors individually.

1.3 Research Questions

The research problem identified was divided into six research questions which helped us address it systemically, and this enabled us to develop the proposed framework and a specific model of the controllable factors that affect students' performance.

Overall research question - How can a specific model be developed to enhance academic performance which deals with controllable academic factors? Furthermore, how can this model be used within a framework for implementing a readily available student learning assistant tool?

Below we expand the overall research question by presenting four questions identified in the problems described in the previous section that we intend to answer in the thesis.

1. How can the impact of the assistant on the student's academic life be investigated?
2. How can a structure be built to assist students at universities that combines factors perspectives such as Psychology, Self-responsibility, Sociology and more, and which incorporates the university focus on those student-controllable learning factors which most affect students' academic performance to support students in dealing with them?
3. Is there a scalable and cost-efficient way to deal with student-controllable learning factors in students' academic subjects to facilitate student - lecturer interaction at universities using

new technologies such as Artificial Intelligence, Natural Language Processing and Deep Learning?

4. How can an application be built which incorporates an AI student learning assistant tool that could assist students in dealing with students-controllable factors?

These research questions will be addressed through the aim and objectives defined below.

1.4 Aims / Objectives of the study

The aim of this study is to develop a framework for assisting students using a specific model to enhance academic performance. To achieve this aim, we address three main objectives:

Objective 1: to build a model that combines perspectives such as Psychology, Self-responsibility and Sociology and focuses on student-controllable learning factors to assist students in dealing with those factors.

Objective 2: to research a scalable and cost-effective technology to assist university students using the model described above and to investigate how to assist students with their academic subjects and facilitate student - lecturer interaction.

Objective 3: drawing on the first two objectives, to build an application which incorporates an assistant chatbot which will facilitate student-lecturer interaction and assist many individual students in using a model and engaging in knowledge discovery.

We will start by investigating the literature as a basis for building a specific model to facilitate a student-controllable learning factor model. We will then build a methodology to find the correlation among the specific model factors. Next, we will develop a method to build the framework for designing a readily available and student-usable tool that could assist students. The researcher's expectation in achieving these objectives and the aim will have the following significance.

1.5 Significance

The world is complex and constantly changing. We face daily adversity and the unpredictability it brings. Humans are constantly learning. Every day we learn from our interactions. As emphasised at the beginning of this research, we need assistance and guidance to facilitate our growth and learning as human beings.

Studies (Cummings, A. M., 2014) (Prebble, T., et al., 2004) (Shoppe, R., 2019) show that having individual assistance in the areas that influence their academic results has a positive impact on students. This goes beyond simply impacting their academic performance: guiding students in areas they can control will help them manage those areas, and this will be a skill they will use even after leaving university. We believe that in order to achieve academic success it is essential that students should have distinct attributes such as self-control, the ability to communicate their ideas, effective learning techniques and self discipline, and that they have a good work environment and are mentally and physically healthy and more.

Many universities have an application that help students develop these characteristics. Having even more assistance in their academic subjects will give students a new perspective, making them more capable of dealing with the adversities of academic life. The proposed framework will enable the university to provide assistance to students individually from a distance at any time of the day. This has an even greater relevance in the current pandemic, where requiring social distance it is one of the numerous measures that governments have adopted.

Building the proposed model has the following implications: university students will have a better understanding of the factors that affect their performance. The model will predict their assignment grade results, thus illuminating which variables have the most significant impact on their studies. Later, we will be able to see which variables should be given more attention. Furthermore, we can also make correlations between the variables, thus generating knowledge.

In addition, universities will have an affordable application, and even universities in developing countries will be able to use the tool. The controllable factors for students will be monitored from the beginning of classes. Each student will have a profile that will show them the factors that they should prioritise. The chatbot will also serve as a facilitator for communication between lecturers and students, thus communicating questions relevant to the classes from students to the lecturers. Multimodality will enable the chatbot to shape the presentation of the

answer according to the nature of the question. If the students' subject content is predominantly visual, the proactive chatbot (by 'proactive' we mean that it will anticipate and take a proactive action) will use multimodality to present more visual content, such as a graphic or a video.

The Significance of the Research:

- **Dissemination of knowledge;**
- **This idea proposed in this research was judged to be a Winner at Big Idea Challenge 2020;**
- **Appearance of the researcher on Angolan National television to present the research.**

We now turn to the contributions of the research.

1.6 Research Contributions

In this section, we highlight this research's contribution to knowledge. There are three main contributions to knowledge which have been reworked into academic papers.

Firstly: The creation of the innovative PS2CLH model, which combines the principal student-controllable perspectives that affect students' academic performance the most.

Second: This research adds to knowledge an innovative approach by representing in 3D the factors which are controllable by students, which offers decision-makers a different way to manage and take decisions on factors such as stress, time management, sleep problems and more.

Third: The new AI chatbot framework with innovative components creates an educational chatbot ecosystem, which uses knowledge from the PS2CLH model and AI - DL to build a proactive chatbot to assist students' in learning their academic subjects and in controlling important learning factors.

1.7 Research Scope

Looking at the proposed framework and the Deep Learning spectrum, we find a vast range of academic subjects which could be drawn on for relevant research. However, we will focus on automating the student's assistant and customising it using a specific model and the AI chatbot. For this reason, our research will not address other features required for mass production of the proactive chatbot framework, such as the applications security with standards-required precautions, portability and scalability among others.

The proposed framework is still in the initial stages, and it is natural at such a point to encounter some conceptual limitations; however, we will try to design a multimodal form of interaction with the student. Despite advances in AI, we still have a long way to go to before we dare think that AI could reach human levels of rationality level or become a technology that replaces human interaction. However, it is essential to pay attention to practical aspects of this project.

1.8 Practicalities

1.8.1 Location

England, London at London Metropolitan University and Angola, Luanda at Universidade Católica de Angola, ICT laboratory.

1.8.2 General requirements

The research application will require a host server for the application, possibly an account for training the data and creating models for Natural Language Processing implementations such as predicting students' results, creating the question-answer model, and more.

1.8.3 Feasibility

Outcomes could be affected by participants' willingness to complete a survey and be involved in the research, which is important because these are our intended data collection tools.

1.8.4 Achievability

What makes this research achievable is the successful development of proactive chatbots in recent years and Deep Learning researchers' advances in deep learning algorithms and

architecture and hardware capabilities. These practicalities take into consideration ethical issues and responsibilities.

1.9 Ethical considerations

1.9.1 Confidentiality of Student Data

Ethical considerations are of great importance in our research. The nature of the project requires dealing with aspects of students' characters which may impinge on their studies, and we must respect all student participants. Consequently, students will be able to withdraw from the project or refuse to participate at any time (Research & Enterprise Development Centre, 2014).

Therefore, to reduce the risk of physical and mental distress arising from participating in the research, students will have the freedom to remain anonymous, so (in accordance with the Data Protection Act (DPA) 1998) their data on the server will be encrypted to ensure its confidentiality and privacy (Research & Enterprise Development Centre, 2014).

1.9.2 Informed Consent

Information and informed consent sheets will contain the following information:

- *Name and contact details of the researcher*
- *The role of the participant in the research project*
- *The treatment of the material/information collected*
- *The name and contact details of the participant (optional)*
- *The aims and objectives of the research project*
- *The potential risks to the participant*
- *Sources of advice/help/support/treatment*
- *Confirmation that the project has been granted ethical approval*
- *The contact details of the RERP Chair and Clerk*
- *The option to receive a summary of the research findings*
- *Confirmation that the research is voluntary and that all participants have the right to withdraw at any point*

➤ *The Signature and date of the researcher and the participant*
(Ellison, G., 2013)

1.9.3 Lack of regulations

Angola does not have any clear research ethics approval framework.

We therefore applied for a similar level of research ethics approval as that required for UK-based research (Ellison, G., 2013). In this way, we covered the ethical facets, and the next section is the structure of the thesis.

1.10 Structure of the Thesis

This thesis includes six chapters beginning with an overview of the research that starts with the introduction, statement of the problem, and research questions to be addressed. This is followed by the aims and objectives of the study. This chapter continues by outlining the significance or implications of the research, followed by its limitations / constraints, as well as practical details, including the location in which the data was collected, costs, feasibility and achievability. It then continues with ethical considerations such as student confidentiality and informed consent sheets.

The second chapter is a comprehensive literature review of the science behind practical assistants, different coaching styles and connection and individualisation – in short, all the fields drawn upon in developing the proposed proactive chatbot. It continues by exploring Artificial Intelligence, Natural Language Processing, Measurement of Text Similarity and Deep Learning, and then asks why are we using a chatbot to assist students. This chapter continues with the use of chatbots in universities, then chatbots and chatbots frameworks, and closes with a discussion of relevant graphics.

The third chapter focuses on the methodologies used in this study. First, coverage of the generic methodologies used in this research is presented, i.e. Quantitative and Qualitative Research, followed by the CRISP-DM and Scrum methodologies. There is also a discussion of the methodology for finding the correlation among the PS2CLH model factors by predicting students' academic results and the method used to build a framework designed for the proactive chatbot for students.

Chapter Four starts by presenting the proposed PS2CLH model, and explaining why it was developed. This chapter continues by discussing the literature supporting each area of the PS2CLH model, beginning with Psychology and Self-responsibility. It shows why those two areas were built and how important they are for students' academic performance. In the next section, another critical area of the PS2CLH model is the fields of Sociology and Communication, which show the social impact on students and how important language is in the learning process. Lastly, it presents two essential areas: Learning and Health & wellbeing. The first area shows how vital learning techniques are in students' academic life. The model closes with Health & well-being, where some research shows that if students practise physical activities, their results are affected positively. This chapter ends with two experiments developed by a researcher in Angola and then describes the students' visual 3D representation.

The fifth chapter begins by introducing the fundamental connection between the PS2CLH model and the proactive chatbot. The framework designed for the proactive chatbot for students is based on PS2CLH, and falls into two parts. The first part of the proposed framework presents Questions/Inputs, Data Preparation and Embedding and Vector/Matrix, Text Similarity Measurement, the Question and Answer model and Assembly Parts. The second part of the proposed Framework comprises the components: Interaction Facilitator, Profile Customiser, Multimodality, Rating, Suggest Factor, Knowledge Database and Profiles, and finally, the Test and Results. Chapter six presents an evaluation and conclusion and suggests projects for the future.

1.11 Summary

This chapter presented an overview of the research and then the research problem, i.e. the lack of efficient, scalable and inexpensive ways to assist university students individually. Then, pertinent questions were posed related to creating ways of tackling the student assistance problem. Next, we defined the proposed aim and objectives of the research. Then the scope and limitations of the research were considered, before a discussion of practical research issues such as location, cost, feasibility and achievability. An essential area of the research was ethical issues such as confidentiality, informed consent and the lack of relevant regulations guaranteeing data privacy in Angola. Finally, the structure of the thesis was outlined. The next chapter is the Literature review, in which different areas are explored for building the proposed model and chatbot framework.

Chapter 2 Literature Review

2.1 Introduction

Seeking to answer the research question, we start by laying the psychological foundations for the role of Assistant/Coach with one of the founders of psychoanalysis, Sigmund Freud. Influential Assistant/Coach philosophies from the last decades are then presented, which illuminate our data and strategies for assisting students. To meet the researcher's aim and develop a framework for student assistance using a specific model to enhance academic performance, the research continues by investigating different Assistance/Coaching styles. Then we explore learning styles, multimodality and personality types. Subsequently, we review the use of chatbots in universities, and briefly research the field of Artificial Intelligence; A brief review of Knowledge Discovery is presented at this point; The other topics covered here are Natural Language Processing, deep learning models and the measurement of text similarity. In keeping with the chatbot's present landscapes, we turn to the classification of chatbot models, namely generative vs retrieval-based models, long vs short conversations, closed vs open domains and then the common challenges of building chatbots. The final part of the literature review presents a short overview of the Chatbot framework.

2.2 Student Assistance – An Overview

The first part of the research focuses on the educational area, the impact of the assistant on the student's academic life. We investigated the science behind why a practical Assistant/Coach can improve a candidate's performance, covering the topics of Assistant/Coach, learning style, multimodality, personality types, the visual, auditory, reading/writing and kinesthetic (VARK) system and debugging learning styles. We now turn to the use of chatbots in universities and why we should use a chatbot to assist students, first of all presenting the science behind practical assistants.

2.2.1 The science behind the ability of the practical Assistant/Coach to improve a candidate's performance.

Why should we need an Assistant/Coach to deal with change and improve our performance? Because any improvement is a challenge, and there is a neurological reason why changing bad habits and raising performance is challenging. According to the founder of psychoanalysis Sigmund Freud, a defence mechanism in our brain is activated when it faces pain, and change requires doing painful things, or anything which is uncertain, scary or new. Freudian psychoanalytic theory states that defence mechanisms exist to protect the brain itself from anxiety and psychological pain, to distort reality, manipulating or denying the brain prompts unconscious psychological strategies (Isaksen, V. J., 2013). The moment we want to change, break a habit or do something hard or scary and we hesitate, the protection mechanisms are activated and stop us (Yousry, M., 2011). Researchers refer to these cognitive biases variously as 'the paradox of choice', 'the psychological immune system' or the 'spotlight effect'. (Robbins, M., 2017).

From a programmer's perspective, this is a sort of anti-virus program ready to be triggered at any hint of a suspicious situation. The assistant process is a program that teaches us to deal with that suspicious situation, creating a safe environment, allowing us to be the best version of ourselves. There are many scientifically proven principles in modern psychology; here, however, we will focus on five of them.

Principle 1 - Locus of control: Research has revealed two types of people: those Individuals who believe that they control their lives (internal locus of control), and those who believe life is just happening to them. Hence they are simply the victims of whatever might happen (external locus of control). It seems that people who are in control of their lives are happier, more prosperous and perform better. Dr Julian Rotter initially conceived the locus of control concept in the 1950s (Rotter, J., 1966). How do we create an internal locus of control in ourselves? The article "Internal Locus of Control (and Why It's Important to Success)" by Brandon Hill gives some tips for developing an internal locus of control:

1. Change your mindset, self-talk, and the story you tell yourself.
2. Focus on what you can control in life.
3. Be aware of your habits and build good systems. And more. (Hill. B., 2020)

When we act consistently in daily life, our behaviours help us to build an internal locus of control (Mamlin, N., et al., 2001).

Principle 2 - Behavioural flexibility: Modern psychology has proved that our brain never stops creating new connections and learning new things. This is called behavioural flexibility (Flexibility, Behavioural, 2012). With a practical Assistant/Coach, learners not only create immediate change at that moment, but also over time, as they practice it, they create new habits. This is what the neural pathways in our brain are designed to do, and it is how new neural pathways can develop. In the long term, we are creating lasting behavioural change (Rubin, G., 2015).

Principle 3 - Do good, be good: This is from a psychological researcher at the University of Virginia, Professor Timothy Wilson (Wilson, D. T., 2015). Summing up his book, he argues that we cannot think or plan to be happier and wait for something to happen. We should do something. Going back to Aristotle's thinking, we find the same analogy. To change, we should do things, act and then our mind will follow. With a good Assistant/Coach's, learners could create good habits (Wilson, D. T., 2015).

Principle 4 - The golden rule of habits: Many scientists have researched habits, including Dr Charles Duhigg and Dr Gretchen Ruben, both of whom have written best-selling books. Most researchers believe that habits come from one golden rule. We can never change the things that trigger us, we cannot control how we might feel, but we can always choose how we react or behave (how we behave is a choice). The best way to stop a habit is to replace it with new behaviour (Duhigg, C. & Ruben, G, 2014).

Principle 5 - Activation Energy: Professor Mihaly Csikszentmihalyi at the University of Chicago was a pioneer researcher, studying the state of FLOW (Csikszentmihalyi, M., 1997).

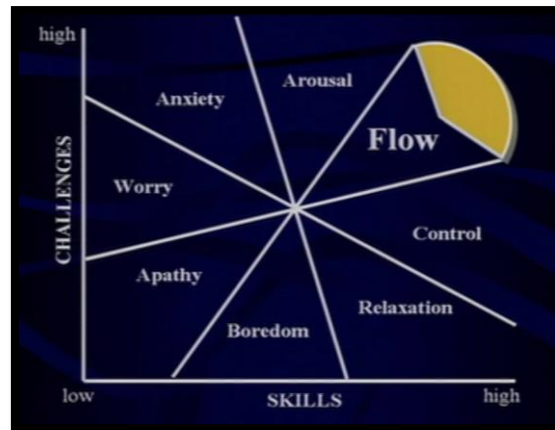


Figure 1 - Finding Flow – State Diagram (Csikszentmihalyi, M., 1997)

The state of Flow is achieved when a person’s very high skills meet a significant challenge. It is the happiest moment of their lives, as they are doing what they love to do in dealing with the most complex challenges. However, they always need to take a first step in facing such a challenge. Dr Mihaly studied human behaviour and human performance. He called this first step “activation energy”, a term from chemistry.

The initial chemical reaction requires a tremendous amount of energy. That is why it is hard for people to start doing what they should do (start studying hard, go to the gym and work as hard as they can, go to work and do their best) and it is that feeling of how hard it is to get started that Mihaly is talking about. The first step is the key to creating any change and reaching our full potential (Csikszentmihalyi, M., 2008). Assistants/Coaches tend to create this activation energy and make learners do what they are expected to do.

These five proven principles have been a foundation for psychologists, coaches and personal assistants for decades. However, there are different methods and principles for assisting and guiding people to help them improve personally, and some of the claims present better results than the five proven principles. After presenting those assistant/coach styles, we will contrast and discuss the five principles and the Assistant/Coach approaches. We now turn our attention to the Assistant/Coach.

2.2.2 The Assistant/Coach

To review the different assistant/coach strategies, we will present some non-scientific but successful Assistant/Coach approaches with proven results. Robbins argues that we are born

with great “potential”. Moreover, when we seize this potential, we will have everything we need to create a life filled with more love, passion, enthusiasm, confidence, and joy than we have ever dreamed. (Robbins Seminar, 2016). Conversely, Pablo Picasso thought that ‘*Our goals can only be reached through a vehicle of a plan, in which we must fervently believe, and upon which we must vigorously act. There is no other route to success.*’ (Richardson, J., 2008). Which processor vehicle a plan (roadmap) can seize this potential that was born with all of us? A writer and director of the OCM Group Ltd said that it is possible to harness it through a coaching/assistant process.

A process of personal development is one that enables learning to happen and performance to improve. Dr Parsloe says; *To be a successful coach requires a knowledge and understanding of the process and the variety of styles, skills, and techniques that are appropriate to the context in which the coaching occurs* (Parsloe, E., 1999). Conversely, mentoring is the continuous following by one person (the mentor) of another (the student) which makes substantial transitions in knowledge, work or thinking for an extended period (Clutterbuck , D. & Megginson, D., 1999). ‘Both coaching & mentoring allow people to take ownership of their development and release their potential’ (Connor, M. & Pokora, J., 2007).

The principal aims of coaching/mentoring are to support people: develop their skills; improve their performance, maximise their potential; and help them become the person they want to be (Jennifer, G., 2013). With the use of an assistant/coaching growing nowadays, many institutions give coach training programs certified by the International Coach Federation (ICF) (Tugend, A., 2015). Critics perceive life coaching as the same kind of thing as psychotherapy. However, there are no restrictions, oversight, regulation, or established ethical policies (Morgan, S., 2012) There are no regulations or licensing requirements for coaches/mentors. Furthermore, most life assistants have no formal training or certification (O'Brien, E., 2014). Therefore, it is expected that future orientation and guidance from the future research application should be regulated and approved by experts in the field before it becomes available to students. However, there are different styles and types of Assistant/Coach, which we investigate below.

2.2.3 Different Assistance/Coaching Styles

Different coaching-styles are connected. The Bible was the initial source for Jim Rohn, who was one of the fathers of “Personal Development” and Tony Robbins’ and Les Brown’s mentor. Tony and Les were in the list of Best Coaches in 2015 and many other years (according to Motivation Grid, wealthygorilla.com and many other references) (to mention that in the scientific world those motivation coaches do not have much credibility). Rohn developed the principle and foundation of seminars related to personal development. Nowadays these are called mentorship/coaching seminars. Rohn states that what we have at the moment we have been attracted by the person we have become. From that viewpoint, it asks a critical question: how can we change our lives? He states that if we change, everything will change for us. We do not have to change what is outside; all we must change is inside (Rohn, J., 1999).

Robbins continued with Rohn’s legacy. Robbins is a practising psychologist. He believes that the best way to improve a person’s performance is by “modelling” the best people in the desired field. To him this is the fastest way to become experts in that field. Look outside and model the inside. He believes that we all act consistently with whom we believe we are. Robbins states that the most potent force in all human personality is the “need to stay consistent with how we define ourselves”. If we define ourselves as a conservative person, we will act conservatively (Robbins, A., 2000).

Psychology experts argue that there is a moment when we define ourselves and create our identity. Consequently, most people make decisions based on beliefs created long ago. We behave according to the definition of ourselves (the identity that we create to ourselves) and what we think we are (Robbins, A., 2000). Exploring Robbins’ approach to coaching, we find that it is based on “Questions”, such as ‘Who are you at this moment?’ ‘Who have we decided to become?’ And more. Then he reframes the person’s identity and proposes that we should make a personal decision, and make it carefully and powerfully to change the old identity and then act upon it. He believes that three essential areas to change are the person’s strategy for creating the identity he is currently living, the story that we tell ourselves to keep that identity, and our emotional and physical state (Robbins, A., 2000). Robbins also uses Neuro-Linguistic Programming (NLP). The Co-Creator of NLP is Dr Richard Bandler, an author and trainer who has more than 40 years experience in the coaching field.

He claims that all thoughts are images or pictures in our mind. Rather than look at the past, he focuses on the present and creates a new reality where he uses hypnotic techniques to amplify the consciousness and change the reality. Another way of describing his view is that he mentally changes the size, colour, and previous reality. He has a similar idea on modelling, the systematisation of what works well, or the way people do things to improve performance. For Bandler, humour is a tool to make people aware of their real problems and see that the strategy they have taken has led to undesired results. People are willing to change their status and the picture they create in their minds, and he replaces them with a better view of themselves (Bandler, A., et al., 2014).

In the same realm, and with more than 40 years in the coaching field, Bob Proctor has based his coaching style on Napoleon Hill's book "Think and Grow Rich", which basically sets out 13 principles for personal growth and how to achieve any goal. He believes that peoples' results are related to their paradigm (what they think, the dominant idea of their capabilities); his approach focuses on changing the paradigm, replacing the old idea, changing to a new level of vibration (Proctor Gallagher Institute, 2015).

On the other hand, some coaches or mentors believe that the best way to improve people's performance is by looking at themselves and finding the best in themselves. For instance, martial arts legend Bruce Lee said, "*Always be yourself, express yourself, have faith in yourself, do not go out and look for a successful personality and duplicate it*". It seems that the fastest man in the world, Usain Bolt, has developed his style of running according to a former champion Michael Johnson, a retired American sprinter. Conversely, Vishen Lakhiani's meditation expert focuses on making an outside change by changing the inside by Bend Reality. The idea of bending reality uses our consciousness to affect and shape our Universe (Lakhiani, V., 2016). The Kybalion, seems to be a mind of all Hermetic classic, it is an ancestral philosophy used by Mayas tribe. Use resistance to generate power. The general idea is that we do not have to change ourselves; instead, we use the situation and what we are to grow ourselves (Three Initiates, 2010).

In the same vein, in the words of the educator Nietzsche, "*No price is too high to pay for the privilege of owning yourself.*" (Nietzsche, F. W., 2014). Our instinctive and intuitive reaction is to avoid all pain and suffering. The technology and the ease of achieving everything have made us ungrateful, and we have forgotten that suffering is an integral part of life. However,

according to Nietzsche, it is only when we are willing to face the challenges of life that it is possible to grow spiritually. In 1873, when he was 29 years old, Nietzsche addressed this fundamental question of how we find ourselves in a beautiful essay titled "Schopenhauer as Educator". In this essay, he argued that *"if someone wishes to be somebody in this life, to maximise their potential, they need to take the difficult path, which often leads to isolation."* (Nietzsche, F. W., 2014)

Being a loner is not easy, but this is one of the prices someone must pay for the privilege of owning themselves. To keep ourselves from being overwhelmed by the tribe, we must distance ourselves from others. We need to strive to be free, and this might lead to severe difficulties in our life. We should refuse to take an easy path, and we should decide to embark on a quest to gain the freedom to be ourselves, no matter how frightening it might be.

To be free means also to be free from all physiological and psychological needs - in other words, not to let them drive us, but for us to drive them instead. For example, whenever we feel an emotional urge to do something, complaining to somebody in a very passionate way, we must first try to become conscious of this impulse and then decide if we should act upon it or not. Nietzsche's philosophy on this point is in

some ways similar to the ideas of many modern-day motivational gurus and 'thought-leaders', but the similarity is only on the surface. Nietzsche is more profound than motivational gurus, who focus on self-development to achieve worldly success and a fulfilled life on a material and relational level. The fight is an inner fight, the struggle is to find oneself, and this quest is a much more difficult mission, requiring a very different kind of sacrifice.

For instance, a motivational guru may teach us how to be more confident to become famous and attract investors to our business. Yet Nietzsche teaches us first to analyse the root cause of our desire to be confident. Usually, we will find it is the desire to impress other people such as our loved ones and our friends or to prove a point about ourselves to society. A simple analysis might make us give up this desire and focus more on what matters in our life, i.e. much deeper issues such as self-discovery. Moreover, this endeavour might make us a loner.

Refusing to compromise ourselves can very well put us in conflict with many people. It means changing our lifestyle. It means giving up friendships or other relationships, looking deep into our fears, analysing our deepest emotions, including our darkness, and rising above them. We

have to break down the chains of opinion and fear. Nietzsche encourages us to “challenge our own demons”; nevertheless, we should not simply cast them out, as beyond them there is a deep meaning which we should try to understand. We need to get out there in the world, do things, experience different temptations, and be always present with our entire consciousness. In the end, we emerge as individuals with a distinct strength of character and a much richer inner nature.

If we do not go out and experience life first-hand in a fully aware state, we cannot claim we have lived our life. How far we can go depends on how much we are willing to pay for that. To reach the state of self-ownership and avoid going through life in a meaningless way, we must learn how to find our inner genius ourselves. To get in touch with our inner genius, we must walk a path no one has walked before, as we are unique, and no one can walk that path on our behalf. Finding ourselves is finding our uniqueness, that unique set of values and things we truly love which represent us (Nietzsche, F. W., 2014).

Discussion: We start with the fifth psychological principle for improving performance, which has to do with activation energy, the enormous initial energy required to begin the transformation. The principle is that we have to take the first step towards change and thus reach our potential. The truth is that if we do not take any steps, we remain in the same place. However, if we take steps in the wrong direction, we will certainly not reach our goal. However, we often learn from the experience of going in the wrong direction, even though it sometimes takes us a long time to realise that we have taken the wrong step. In this case, the external guidance of an assistant who knows the best way to achieve the desired result will make all the difference, thus helping to minimise the learning time. However, a person has to take the first step if he wants to improve his performance.

Along the way to reaching our goals, there will always be external triggers and events, according to the fourth principle. We cannot change the trigger, but we can always choose how we react to the trigger. The best way to change a habit is to replace it with a new habit, but this is a complex and slow process. In some cases, if we do not have the ability to change to a new habit which brings us closer to our goals, we will end up creating habits that we find most comfortable for us, in which case orientation to a better habit and the experience of outside assistance can be a significant factor. The third principle is similar to the fifth principle, which says that to change, we must act. After making the plan, we must act to make it happen.

The second principle is about behavioural flexibility. Our brain does not stop learning new things, but on the contrary always creates new connections. Indeed, the capacity for learning and the speed with which we learn decrease with the ageing of cells, especially as there are certain things that we only learn when we are children- for example, to distinguish acoustic tones from musical notes. Another example is that dyslexia can be treated in children, but adult brains no longer relearn to the point of curing this learning disability.

Finally, according to the first principle, research has revealed two types of people: those with an internal locus of control and those with an external locus of control. Given these two types, the principle emphasises that we should focus on having control of our lives rather than being guided by events beyond our control. The next section will show that different assistant/coaching-styles are reflected in the two categories mentioned above. Some have as their methodology a focus on internal control and others on external control, such as modelling an expert in the field to reduce learning time, avoid predictable errors, and apply proven and effective techniques to obtain the expected result. In this case, people with an external control focus tend to have better results than those with an internal control focus. Nevertheless, the real change comes from the inside. According to many experts in personal development, many of the techniques with external references and short paths do not have very long-lasting effects because when we stop being externally influenced, we return to what we are internally.

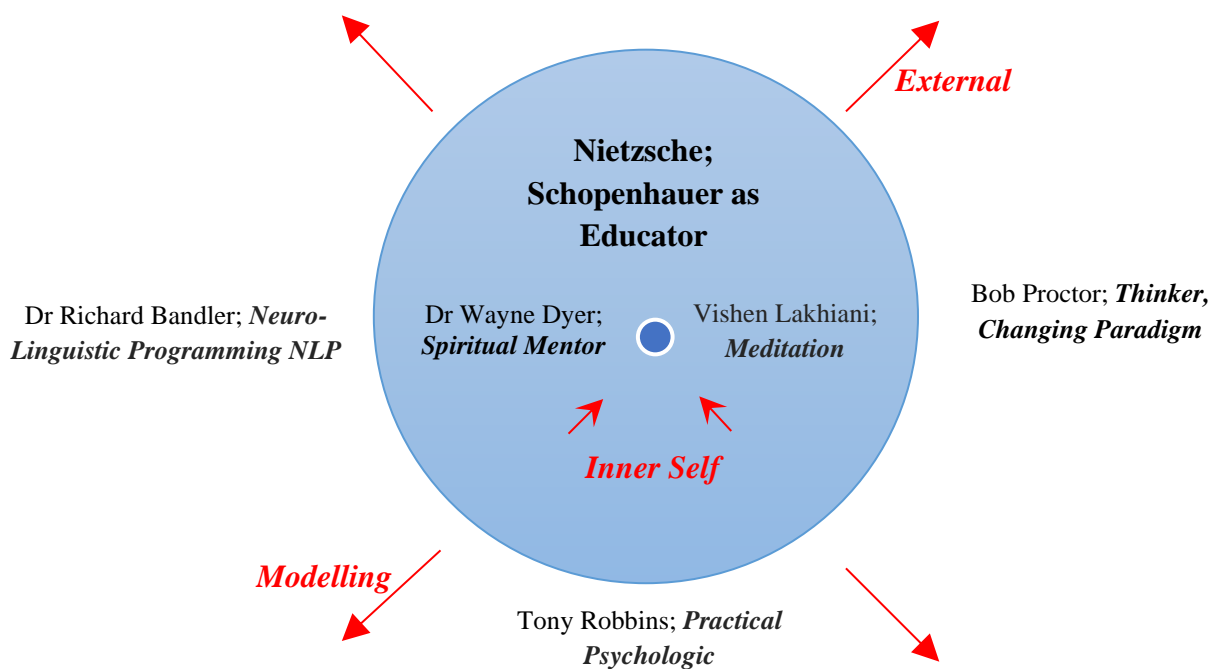


Figure 2 - Different Coaching approaches diagram

Figure 2 summarises the key features of two coaching styles, the external approach and the inner self approach. These two approaches will support our understanding of how we can assist students.

Discussion: Rather than discarding one approach and exclusively using the other, we believe that both forms of assistance or coaching - looking inside or externally - can be helpful depending on the student and their circumstances. For example, the academic year is short, and we do not have enough time for the student to be guided using only internal control focus strategies. In addition, many students have insufficient focuses on internal control. Therefore, we believe a hybrid strategy should be used, in which students are assisted according to their personality. The next section presents learning styles, multimodality and personality types.

2.2.4 Learning style, Multimodality and Personality Types

The classic view of learning is that there are four main learning stages: attention, encoding, storage and retrieval (Hoose, N., 2021).

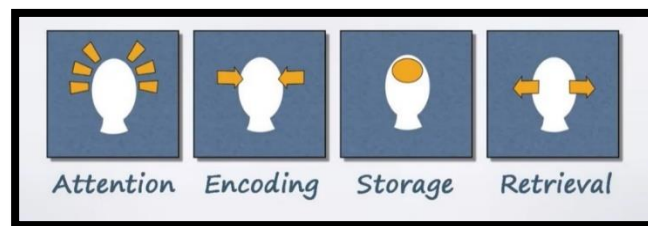


Figure 3 - Four main learning stages

Figure 3 above shows the four main learning stages, starting with capturing the attention or focus of the student, then giving a meaning to the message by encoding it as information, and then storing or memorising the information and finally retrieving the saved information. This is a process, and during the learning process we do not want to tune out the content we are learning, but to remember it. This process is often compared to the way a computer works. When we are type a Word document or create a PowerPoint presentation, we put all the relevant information in it and save it. We can store it because we do not want every single document on our computer open all the time. Therefore, we close things down and file them away.

However, when we need the document again, we can open it up on the computer and refer to it. Our memory works in the same way. We have to encode ideas and information: we want

store them away because we do not always think about everything. However, we want to be able to recall them again when we need them (Hoose, N., 2021).

What Do We Mean by “Learning Styles”? Cognitive aptitudes such as using visual-spatial stimuli and/or language processing can be strengths or weaknesses of specific individuals. For instance, student A may prefer listening to a lecture, whereas student B prefers reading an article related to the subject. This is based on the “learning styles” theory which suggests that teaching students in a style that conforms to their preferences improves their learning capabilities. Different learning styles have been categorised on the basis of student self-reports/questionnaires (The University of Kansas, 2021).

Howard Gardner from Harvard University suggests that there are seven styles of learning. Cognitive research studies the extent to which people with dissimilar minds learn, recall, perform, and comprehend in different ways (Gardner, H., 1983). He claims that *“Everyone is capable of perceiving the world through logical-mathematical analysis, musical thinking, language, spatial representation, and the use of the body to make things or to solve problems, an understanding of other individuals, and an understanding of ourselves.”* (Gardner, H., 1983).

Gardner’s studies help us understand different learning styles, namely visual, aural, verbal, physical, logical, social and solitary.

VARK Learning Styles: It has been found that the most common learning styles used in schools by students, classified as the VARK system, are Visual, Auditory, Reading/Writing and Kinesthetic styles (Cuevas, J., 2015). Firstly, students are asked 25 multiple-choice questions that range from how they enjoy their lecturers. Then their sense of direction is tested by determining how they give directions to a neighbour’s house (e.g. saying them aloud, writing directions, drawing a map or walking with the person) (Cuevas, J., 2015). The system then identifies them as Visual, Auditory, Read-write, and/or Kinesthetic learners and recommends specific learning strategies based on their answers (The University of Kansas, 2021).

A survey conducted in the UK and the Netherlands of almost 400 lecturers found that over 90% of them strongly believed that when receiving information in their preferred learning style individuals tend to learn better (Prof Sims, J., 2019). Although this brief study may offer some

insights for lecturers, it should not be seen as definitive to avoid categorising students to reflect “desired learning styles” and thus running the risk of simplifying the complex issue of how to teach and learn and thus disadvantaging students.

Debugging Learning Styles: Learning styles are a preference - so how strongly do learners stick to their preferences? In a 2018 study, over 400 students at a university in Indiana completed the VARK questionnaire during the first week of the semester and were classified according to their learning style. When the semester finished, the same students completed a study strategy questionnaire to determine how they had studied during the term. It was found that there were significant discrepancies between the study strategies used by students and their learning style in most cases. Only a minority of the students did not perform significantly differently on the assessments in the course (May, C., 2018). Previous research has also shown that learning styles have made no difference to students’ performances.

Discussion: Neil Fleming, a school inspector in New Zealand, created the VARK model. Describing the origins of VARK, he explained that while observing different lectures in class, he had found that an excellent lecture could not reach some learners while poor lectures could indeed reach them. While trying to solve this puzzle he found many reasons, but one topic stands out, - Preferred modes of learning, or ‘modal preferences’ (Khazan, O., 2018) This led to the development of VARK. No study has revealed that students naturally cluster into four distinct groups, but there seems to be some magic that might explain why some lecturers can reach students while others cannot. How can this be? Suppose we accept that some people are more skilled at interpreting and remembering certain stimuli than others, such as visual or auditory. Why do we see no differences in learning or recall with different presentations? It is because what we want people to recall is not the precise nature of the images or the pitch or quality of the sound, but the *meaning* behind the presentations. Some tasks obviously require the use of a particular modality. Learning about music, for example, should have an auditory component.

Furthermore, some people will have a greater aptitude for learning one task than another. Someone with perfect pitch, for example, will be better able to recall certain tones in music. Someone with excellent visual-spatial reasoning will be better at learning the locations of countries on a map. Nevertheless, learning style theories claim that these preferences will be

consistent across learning domains – for example that the person with perfect pitch should learn everything better auditorily. Yet that is not the case. Most people will learn geography better with a map. Articles which review learning styles consistently conclude that there is no credible evidence that learning styles actually exist.

In a 2009 review, the researchers noted that even though the learning styles approach has gained immense popularity within the education environment, any practical ability to classify students' learning styles remains to be demonstrated (Khazan, O., 2018).

Multimodality: It has been demonstrated that multimodality applied to students' learning process does have practical utility. 'Multimodality' refers to the idea that more than one semiotic mode, such as all forms of verbal, nonverbal, and contextual communication, can be effectively used in communication, meaning-making, general representation or specific situations (Bouissac, P., 2007).

Multimodality and synaesthesia are synonymous for Bill Cope and Mary Kalantzis at the Illinois College of Education. They observe that when children are born into the world, they touch, make noises, draw and feel, and become more human as they grow. They use synaesthetic ways of learning and understanding. Once we put them into a learning environment, however, the mission of the traditional school from kindergarten through to university is as far as possible to strip out as much as possible the touching, the noise and the interrupting when they are not being asked to speak. All those things involved in making meaning are stripped out in favour of alphabetical literacy through to the top. What counts is the essay, the quiz. It is what we can write in alpha, in some symbolic form, that is highly valued as evidence, as the artefact of our knowledge.

Educational institutions acknowledge that traditional aspects of classroom education are changing under the influence of multimodality in the 21st century. The rise of digital and internet literacy has brought new modes of communication into the classroom, from visual texts to digital e-books. However, rather than replacing traditional literacy values, multimodality has introduced new ways to communicate and increased literacy for educational communities. Miller and McVee, authors of *Multimodal Composing in Classrooms*, argue that instead of replacing traditional literacies with these new literacies, they should be integrated

into the system since students still need to know the basics, such as how to read and write (Miller, Suzanne M. & McVee, Mary B, 2013).

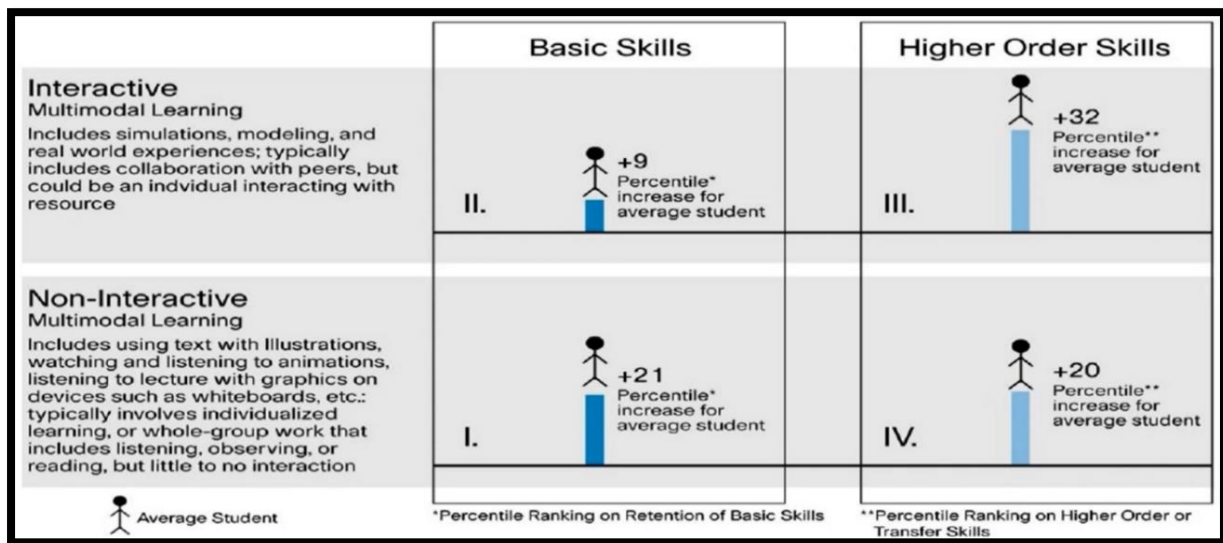


Figure 4 - Impact of multimodal learning (Metiri Group, 2008)

Figure 4 shows that multimodal videos effectively improve student performance for both basic and advanced skills and in interactive and non-interactive environments. However, they are most effective when used to present complex material in an interactive environment.

Discussion: Instead of showing pictures according to student learning styles, we should choose pictures if that is the best way to explain that content. For instance, if the question has to do with geographical location, it is best to show a map. On the other hand, if a student prefers to listen to the answer, he/she should be able to choose this option. Using a proactive chatbot and applying multimodality to students’ learning process can hold students’ attention and explain the content in different ways, using text, image, video and audio to help students and improve their learning experience effectively.

Personalities types: Carl Jung presented the theory of psychological types, which has the same purpose of categorising people according to their different functions and attitudes of consciousness (Jung, C. G., et al., 2014).

16 distinct personalities are the result of permutations of the four dichotomies below. These are categorised by their preference for wide-ranging behaviours and three areas of preferences and dichotomies as follows: Extroverted (E) vs. Introverted (I); Sensing (S) vs. Intuition (N); Thinking (T) vs. Feeling (F). Dr Isabel B. Myers further developed Jung’s work, and added one more field as a fourth pair, namely Judging (J) vs. Perceiving (P) (Briggs, M., 1980).

Although it is not the focus of this research, we intend to build the proposed chatbot framework on the basis of Jung's personality types. To start with, , it is necessary to research the use of chatbots in universities.

2.2.5 The use of chatbots in Universities

Universities' traditional approach to handling students' frequently asked questions is a physical help desk, welcoming visitors and assisting students, or a mailbox and a website. Dealing with the number questions received each day requires dedicated staff (Memon, A. R., et al., 2015). Moreover, to get updates and real-time information, students need to go to the University. E-mails can tackle the distance issue in many cases, and are a frequently used tool to interact with students. It is efficient to send a single communication to many students. However, E-mails are slow, and ineffective for dealing with single requests, as was shown during the pandemic. The asynchronicity of e-mail interaction can lead to an unsatisfactory experience. Research shows that chatbots take advantage of positive benefits of the academic environment ecosystem for students, lecturers and employees. As a result, using a chatbot can be a valuable and effective solution for students due to its conversational synchronicity. It can also reduce University expenses and the workload of staff.

Usually, chatbots carry out two functions in an academic environment, namely dealing with specific area tasks or questions-answers (ICX Association blog, 2016). For example, a questions-answers chatbot application can be used to involve students in Computer Science topics (Benotti, L., et al., 2014). These chatbots are task-oriented, as the issues they try to tackle are related to specific problems.

University chatbot assistants usually provide generic information, for example about courses, admissions and the University's facilities. The information usually offered by chatbots is about University facilities, upcoming events and academic matters (Ranoliya, B. R., et al., 2017). Over the course of the academic year, the chatbot will present different types of information. During the application process, a bot could help students with enrolment and payment of their fees. The objective is for students to retrieve information without browsing several web pages searching for answers to frequently asked questions. Often, users find it challenging to locate relevant information quickly (Ghose, S. & Barua, J., 2013). The proposed framework chatbot

belongs to the “virtual assistant” category. The goal is to assist students with their academic subject and with the daily academic problems (Psychology, Self-responsibility, Sociology, Communication, Learning and Health & wellbeing - PS2CLH model) that affect their performance, such as time management, setting and achieving goals, stress, and sleep problems. The chatbot is available 24 hours a day to answer students’ questions and proactively suggest areas to improve.

Furthermore, the proposed chatbot also aims to create a “human-like” connection when it interacts with the student. A chatbot allows students to retrieve valuable information in more natural ways. It is not only a recommendation to the first steps to enhance the University communication environment, but also a significant source of help during the entire academic year (Morshed, J., 2016). So why should we use a chatbot to assist students during their entire academic year? Below we try to answer this question.

2.2.6 Why should we use an AI chatbot to assist students?

The word ‘chatbot’ is used throughout this dissertation to signify an Artificial Intelligent (AI) chatbot. A critical question must be asked: why is it needed in this context? It assists a university’s students in relation to their controllable factors and their academic subjects, with the aim of improving their learning experience by suggesting scalable and cost-effective solutions. But is an alternative available? If the answer to this last question is ‘no’, then a chatbot is a natural choice.

Looking at possible ways to assist students at universities, we can see the following methods: creating a student assistance department, providing assistance via e-mail or post, using the university website, using an external company to provide assistance for students, developing an intelligent tutoring system or providing a one-to-one personal assistant – i.e. a chatbot.

In relation to the aims of this research, using e-mail, post, a website or an assistance department with a few assistants would not be a practical way of individually assisting students in relation to the controllable academic factors that affect their results and assisting students in their academic subjects. Similarly, using one-to-one personal assistants would require a considerable staff increase and would be an expensive solution. Therefore, from our list of possible solutions, the chatbot is the option that best meets the requirements.

A systematic review of chatbot applications in education was carried out by Chinedu Okonkwo and Ade-Ibijola in 2021 (Okonkwo, C. W. & Ade-Ibijola, A., 2021). They found that chatbots are used in the education system to improve students' interaction skills and assist teaching personnel through automation (Dsouza, R., et al., 2019). Chatbots can also be used in education to increase efficiency and connectivity and reduce uncertainty when students interact with the system (Ondas, S., et al., 2019). They can thus function as a personalised, focused and result-oriented online learning platform (Cunningham-Nelson, S., et al., 2019), which is precisely what today's educational institutions require.

According to (Rapp, A., et al., 2021), the design of a chatbot should not be evaluated solely on the basis of its ability, utility and effectiveness. (Paschoal, L. N., et al., 2018) come to similar conclusions, suggesting that the process for evaluating a chatbot is often based on a small and insignificant data sample to determine its effectiveness.

One of the first systematic reviews of the effectiveness of chatbots during COVID-19 was conducted by Biplav Srivastava at the AI Institute, University of South Carolina (Srivastava, B., 2021). Srivastava observed minimal chatbot use during COVID-19, and asked if chatbots had missed their "Apollo Moment".

The report observes that most chatbots dealt with simple scenarios and there were questions about usability, effectiveness and handling user privacy. It found that the critical value of developing a chatbot is that its success derives not just from the specific technology developed but rather the ecosystem that makes it safely available to people (Ho, C. C., et al., 2018).

Advisory interactions in education are another crucial area in which chatbot technology has been applied. Analyses indicate that chatbots are being used to help students make vital decisions on their various academic programmes or activities by providing recommendations on academic matters. (Ho, C. C., et al., 2018); (Ismail, M. & Ade-Ibijola, A., 2019); (D'Silva, G., et al., 2020) built a Chatbot with the intention of helping individuals gain a better understanding of themselves as well as employment trends, allowing them to make more informed career and education selections.

At the moment, a chatbot is perceived as a valuable tool in educational settings to deliver an engaging experience to students (Clarizia, F., et al., 2018); (Hobert, S., 2019); (Sandu, N. & Gide, E., 2019).

Artificially intelligent chatbots simplify students' *learning process by making it more engaging, short and sharp and attention-grabbing* and assisting lecturers by supporting their teaching. Additionally, chatbots can also help to reduce the workload of administrative staff (Khan, A., 2020).

When someone develops a chatbot, a key question is why it is needed in preference to any available alternative with the aim of improving students' learning experience, this research creates a scalable and cost-effective solution to assist students in developing and developed countries. That is why we are considering the AI-related approach for chatbots, as it could retain students' attention and effectively assist students and improve their learning experience. In the next section, we will investigate Artificial Intelligence, Natural Language Processing, Deep Learning and Chatbots.

2.3 Artificial Intelligence, Natural Language Processing, Deep Learning and Chatbots

In this section, we will investigate the technology behind the proposed framework. Starting with a short history of Artificial Intelligence, a brief review of Knowledge Discovery, Natural Language Processing, Deep Learning and Measurement of Text Similarity, we will investigate text distance, representation and text matching. This will be followed by a discussion of chatbot research which will focus on the classification of models, namely Generative vs Retrieval-Based Models, Long vs Short Conversations, Closed vs Open Domain, common challenges in building chatbots and a brief overview of chatbot frameworks: The section will close with a broad overview of the proactive chatbot framework.

2.3.1 Artificial Intelligence

The world has awakened to Artificial Intelligence. Scientists define *Artificial Intelligence* as the ability of a machine to replicate human behaviour or functions related to the human mind. A machine which has this ability is an intelligent machine (Russell, J. & Norvig, P., 2009). The differences between Artificial Intelligence, Machine Learning and Deep Learning can be seen as follow: *Machine Learning* is seen as a technique aimed at achieving artificial intelligence. Deep Learning is a sub-area of Machine Learning. Machine Learning involves learning from data and experience. As a subarea, Deep Learning was inspired by the human

brain's functionality which is based on connections between neurons. At present, Deep Learning produces the best results in Natural Language Processing.

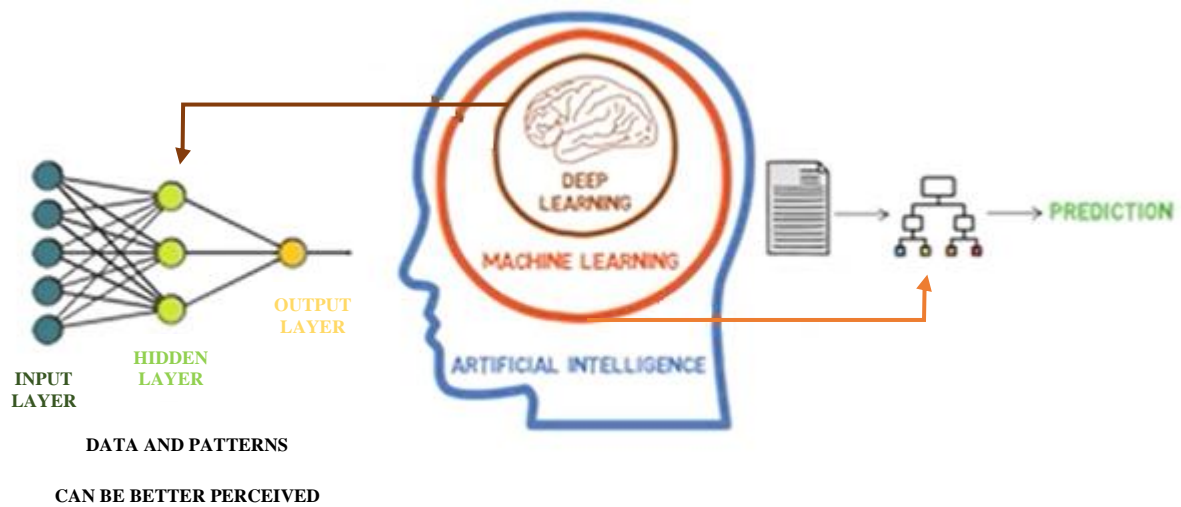


Figure 5 - Deep learning subset of Machine Learning

Vannevar Bush's seminar "As We May Think" began to research the possibility of a machine being able to think like a human (Bush, F., 1945). The first time the term Artificial Intelligence was used was in 1956 by John McCarthy in his first academic lecture on AI (McCarthy, J., 1959). The system proposed by Bush was concerned with the amplification and exploitation of human knowledge. Five years later, Alan Turing wrote a paper that saw the machine as capable of replicate the human behaviour (Turing, A., 1950). In the summer of 1956, a group of researchers in Computer Science, Mathematics and Linguistics got together to develop a research programme. Their vision was to develop self-learning systems, which could solve tasks that require human intelligence. In this research, they used the term "Artificial Intelligence".

The primary AI program demonstrated mathematical theorems. Programming for symbolic computing structures appeared soon after. In the 1960s, AI pioneers were dreaming about the super machine of the future, envisioning the creation of an all-purpose problem solver (McCorduck, P., 2004). The world's first autonomous robot Shakey, which executed simple tasks in the laboratory, was created by researchers investigating machine language. Then Joseph Weizenbaum developed the first known chatbot, Eliza (McCorduck, P., 2004). Eliza could chat in such a realistic way that users at times were led to think they were chatting with a person rather than a Natural Language Processing computer program although ELIZA had no consciousness of what she was talking about. The process was about replacing or repeating

what they said and rephrasing her response using some grammar rules. Eliza mimics a psychotherapist's response to a patient during a consultation. Although those interactions were deemed realistic and the future of the field was promising, the real world was far more complex than the issues they had tried to tackle (McCorduck, P., 2004). Early enthusiasm declined significantly in an 'AI winter', during which research funding was withdrawn (McCorduck, P., 2004).

In the 1970s, scientists wondered whether AI systems would become autonomous or would only simulate behaviour. However, the focus on practical problems changed in the 80s, when the first expert system mixed expert knowledge from specific disciplines to carry out medical diagnoses and other more practical tasks. For example, passengers could request ticket purchase information from chatbots via telephone (Schuchmann, S., 2019). The significant AI breakthrough came only in 1997 when grandmaster Gary Kasparov lost against a robot in a chess match. That was a case of a real expert system demonstrating superiority over a human (Schuchmann, S., 2019). Then, as the amount of data increased, the adaptability and autonomy of service robots increased. That was the birth of self-learning systems (Schuchmann, S., 2019).

2010 marked a turning point in AI applications, and the world embraced it. The growth of computing power and storage capacity played a crucial role. In parallel, there was an improvement in the methodology of Machine Learning; with the development of the Long Short-Term Memory (LSTM) model, for example, which is considered a fundamental element of speech recognition. Other developments contributed to the popularisation of AI, such as the internet, industrial sensors and social media producing vast volumes of data (Research Blog, 2016). As a result, renowned applications demonstrated improved performance, such as Watson, who won the US quiz Jeopardy against a world champion. In addition, applications like the Smart Assistance question-answering chatbot was able to schedule meetings, equipment in intelligent houses responded to voice commands, and the latest Alpha Go beat a master in Go, a board game (Research Blog, 2016).

Discussion: Are we trying to replace human assistance with an AI chatbot? This may be a basic topic of discussion for the scientific community in developed countries, but unfortunately, this question is still taboo in underdeveloped countries like Angola. For example, the first and only scientific paper by Angolan students related to the use of AI in Education was published in 2019 as part of the current research. This example demonstrates

that it is necessary to rectify the frightening images of AI produced for example by Hollywood, which for years shows AI as more intelligent than humans and taking on a life of its own.

It is essential to mention that the intention is to assist thousands of students simultaneously and proactively, that is, anticipating the problem by acting before the problem has escalated and affects the student’s academic performance. Achieving this means that universities have to know the controllable factors that most affect their students at the beginning of the academic year, and this would be costly. We therefore use Artificial Intelligence to create a forecast model and find the correlation between the factors and their corresponding essential attributes, which is what we intend to do in this research. We believe that Artificial Intelligence is one more complementary Intelligence Replica, an extension and interpretation of the knowledge and data obtained by natural Intelligence. Below we present a brief view of Knowledge discovery.

2.3.2 A brief view of Knowledge Discovery

A proactive Chatbot seeks to develop an appropriate environment to gather personal knowledge from our students. Gregory Piatetski defines Knowledge Discovery in Databases (KDD) as the procedure for finding knowledge from data (Piatetski, G. & Frawley, W., 1991). KDD is a combination of technologies for data management such as machine learning, data warehousing and big database. This technology allows the global development of learning valuable knowledge from big data, in which data mining technology is one step in the process (De Martino et al., 2002).

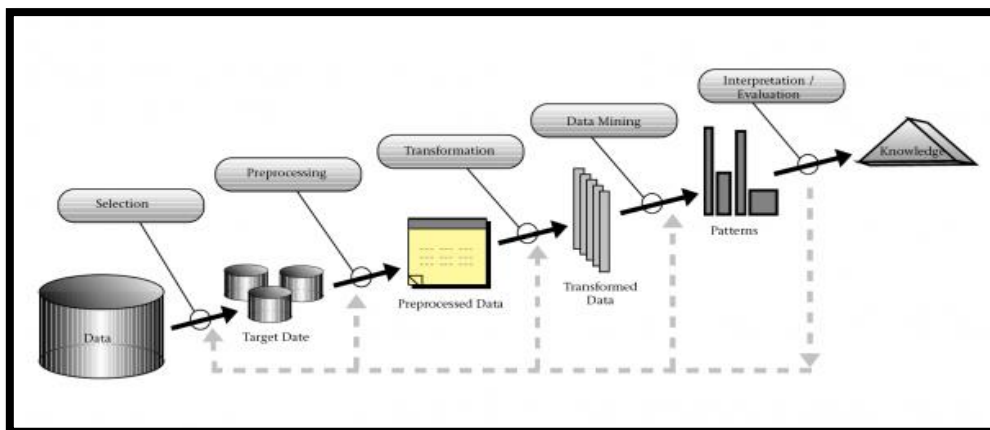


Figure 6 - Steps That Constitute the KDD Process (De Martino et al., 2002)

The principal issue addressed by KDD is related to the initial mapping of data into higher-levels or more abstract and related data. According to Fayyad, KDD focuses on the extensive picture process, including *data collection and access, scalability and efficiency, how results*

can be interpreted and visualised, and how the overall man-machine interaction can usefully be modelled and supported (Fayyad, U., et al., 1996.). In the next section we explore Natural Language Processing in greater depth.

2.3.3 Natural Language Processing

Artificial Intelligence has many fields, including Natural Language Processing (NLP), which focuses on communication between computers and human languages. NLP faces challenges, including empowering computers to discover knowledge from people or natural language input, understanding natural language, and natural language generation (McCorduck, P., 2004). Humans have always been able to communicate with each other through a common language, and despite communications barriers, humans still understand each other.

In contrast, the computer has a machine language known as ‘code’. Since humans have to write all this code manually, it makes the language extremely limited in vocabulary and severely restricted. For instance, if a programmer makes a minor spelling or syntax error, the program will crash, and the problem will not be solved. Old NLP systems cannot understand the context that words are in. But understanding a word all boils down to understanding the words around it, because a word by itself does not have much significance – it all depends on the context. For this reason, older AI systems could not make sense of the data, and ultimately produced lower performances.

However, given how powerful computers are in terms of memory and speed, imagine how more useful they could become if there were some ways to understand and process our language instead of only incomprehensible 1s and 0s. The good news is that there is no need to daydream about this anymore because, with Natural Language Processing, this is the new reality. With NLP, new doors open. People can now analyse the heaps of unstructured data within minutes. Smart assistants such as Siri, Alexa and Meena can understand our words and context clues to improve our lives. Google can predict results as we start typing something into the search bar, and look at the whole picture and recognise more than just the words we type. Our phones and computers can predict what we will type, finish our sentences and suggest relevant words. NLP can solve more significant problems, such as crimes and even diseases, by identifying clues and patterns in emails and written reports (Ameisen, E., 2018).

Parts of speech, such as nouns and verb phrases, articles, and others, offer Natural Language Processing the power to comprehend the context. For instance, in the sentence “The man leaves the bank”, a computer can understand that the main subject of this sentence is the man, and what he is doing is leaving a place, which is a bank. However, an older computer might have been confused as to whether the word LEAVES refers to the act of exiting a place or the plural of the word ‘leaf’, as well as whether the word BANK refers to a building that stores money or the land next to a river. A computer powered by NLP will see that LEAVES in this context has to be a verb and not a noun. As a result of analysing hundreds of sentences and different word patterns, it can work out that it is far more uncommon for a person to claim that they are leaving a bank of a river than it is to say leaving a bank that stores money. Therefore, it must be this bank (Devlin, J., et al., 2019).

Furthermore, it is not only part-of-speech tagging and chunking that allow NLP to uncover the meaning of a person’s words. Other techniques can broadly be classified into two categories: syntax and semantics. Syntax and semantics are the two basic categories that other approaches fall into. These words, their parts of speech, and their arrangement in the sentence provide the computer with information and context about the meaning of the sentence. However, how does the computer know the parts of speech, and how can it figure all this out even if it knows the parts of speech? This is where the power of today’s society comes in, and specifically the power of the big data available in today’s world (Devlin, J., et al., 2019).

A programmer can gather up all sorts of words, phrases, sentences, grammatical rules, and word structures and input this all into a Machine Learning algorithm. Using all this information, this algorithm can learn what words usually end up next to each other, how a sentence should be formed, why certain words fit into a sentence better than others, and more. Furthermore, this is precisely how a computer will eventually figure out why one meaning of the word ‘leaves’ make more sense than another in that sentence. Moreover, data is collected for NLP algorithms every day.

In our research, we will certainly need to process the students’ questions in order to answer them, and as a result we need to explore syntax and semantics, the broad categories under which NLP’s techniques fall. Therefore, we present different techniques used to deal with syntax and semantics in NLP, starting with tokenisation, which falls into two categories, namely sentence tokenisation and word tokenisation. Sentence tokenisation separates a paragraph into distinct

sentences, and word tokenisation separates a sentence into distinct words. This allows the computer to learn the potential meanings and purpose of each word. Following this, we have stemming. Stemming is the process of reducing a word to its root or stem. It does this by chopping off universal prefixes and suffixes such as -es, -s, -ing and -ed. Stemming is a powerful technique, but it involves crude chopping based solely on common prefixes and suffixes, and this sometimes cuts necessary components off a root and changes the original word's meaning (Devlin, J., et al., 2019). Lemmatisation on the other hand reduces a word to its root form by analysing it. Let us take a look at what this means. For example, we have the words 'am', 'are', and 'is', the root form for these words is 'be', which can be seen through lemmatisation.

Stemming would not have been able to work this out, as chopping any letters off of these words would not have outputted 'be'. Now for semantics. Hence, Named Entity Recognition allows the machine to categorise specific words in a sentence (Devlin, J., et al., 2019).

To summarise, Natural Language Processing provides the necessary mechanisms for machines to get human inputs, transform them into machine language, and then process them and respond to us. As we have seen, it uses a variety of techniques to accomplish this: part-of-speech tagging, tokenisation, stemming, lemmatisation, named entity recognition, and natural language generation. As a result, NLP allows computers to comprehend the context and meaning of our words. One of the techniques to allow machines to understand a conversation is deep learning (Devlin, J., et al., 2019).

Discussion: Despite significant advances in Natural Language Processing, where the interaction through language between humans and machines seems more like a conversation between humans, we still have many challenges in NLP, all because of the complexity that the language presents. As humans, we have always been able to breeze past differences or errors in speech, such as mispronunciations, different accents, different contexts, homophones, slang expression, slurs, and more. All differences aside, we can still understand each other perfectly. Thus, we can say that the machine has not yet replaced human interaction in specific domains where a high degree of understanding is needed for communication to be permanent and continuous.

Natural Language Processing has been growing thanks to new technologies such as Deep Learning. Below we present the basics of Deep Learning.

2.3.4 Deep Learning

Deep Learning was born from Machine learning, or more specifically, from deep neural networks, where the architecture is fully connected, comprising an Input layer, one or many Hidden layers and an Output layer. Key elements in building the pro-active chatbot are the application of error backpropagation to minimise the error between the desired outcome and the actual outcome, and drawing on ‘Neural Networks and Deep Learning’ (Goodfellow et al, 2017).

Nowadays, chatbots behave in a more human-like way, and the most relevant cause of this evolution in chatbot behaviour is the growth in Artificial Intelligence, and specifically, the way models have developed in terms of Natural Language Processing. A chatbot can understand the context of a conversation context with the help of a well-trained Deep Learning model. Moreover, with enough data is possible to distinguish different users’ personalities using classification models (Guess, A., 2011).

Below, we present a short overview of Deep Learning models in recent years.

To train a model, what is required is a group of organised words or data, which is known as a document. The computer only understands numbers, so we must convert the document into a sequence of numbers, or vector and matrix. It is a challenge to represent the document as a fixed-size vector. The problem is that the documents are variable in length. Therefore, we must come up in some way, and represent it in a size vector. The traditional way of doing this is by using Bag of words models (Sivic, J., 2009), so that we have one dimension per unique word in the vocabulary.

However, the English language has approximately 100.000 words in its vocabulary. The problem is that this vector will have almost all values as zero, since most words will not be present in one document. This scenario leads to Sparse data (Wang, et al., 2014) which it does not store the zeros.

In a document, sequence matters: for instance “work to live” is very different from “live to work” so these two documents have a different meaning, but a Bag of words model will score them identically because they have the same vector for words present. The solution in this

context is N-grams (Jurafsky, D. & Martin, H., 2020), Dimensionality V^N . The response lies in maximising exponentially the vector representation, which generates other problems. The natural way to resolve this issue is by using Recurrent Neural Network (RNN) models (Pascanu, et al., 2014). These use a for-loop in maths that recursively defines the output at any stage as a function of the input of the previous stages of the previous output. The problem with RNN is the vanishing and exploding gradients. Using matrix multiplication creates a high number of exponents in linear algebra, generating high results and leading to exploding gradients. On the other hand, if the exponent is negative, the results tend to zero, leading to vanishing gradients (Pascanu, et al., 2014).

The evolution of RNN was Long Short-Term memory LSTM (Informatik, F., et al., 1997) models. This is a form of RNN, but has more sophisticated cells, two hidden states, and was invented in 1997 at a time regarded as the dark ages of AI. In order to tackle the vanishing gradient problem in backpropagation, LSTMs apply a gating mechanism which controls the memorising process. This allows information in LSTMs to be stored, written or read via gates that open and close, which solves the vanishing gradients problems (TimesMojo, 2022).

The Transformers model performed better than previous models. It was first described in 2017 in the famous paper “Attention is all you need” (Vaswani, A., et al., 2017). Transformers use Encoder & Decoder. However, for supervised learning problems, we only need Encoder. The main components of a Transformer Encoder model are the input embedding, positional Encoder, multi-head attention, Add & Norm and feed-forward. One of the main components is the Attention Mechanism, which has an all-to-all comparison, in which each layer is $O(N^2)$ for the sequence of length N, and every output is a weighted sum of every input. The weighting is a learned function. Thanks to the evolution of GPUs, All-to-all comparisons can be made fully parallel. This comparison allows computing parallelism.

The main difference between this and the old models is that RNN/LSTM must be computed in serial per token, meaning that we cannot do anything with token eleven until we have completely finished with token ten. That is the crucial advantage of transformers. They are much more computation-efficient. In 2020 Nikita Kitaev, Łukasz Kaiser and Anselm Levskaya published the Reformer: The Efficient Transformer (Kitaev, N., et al., 2020). They introduced *two techniques to improve the efficiency of Transformers .Firstly, they replace dot-product attention with one that uses locality-sensitive hashing, changing its complexity from $O(L^2)$ to $O(L \log L)$. Furthermore, they use reversible residual layers instead of the standard residuals,*

allowing storing activations only once in the training process instead of N times. Consequently, it results in much more memory-efficient operation and works much faster on long sequences (Kitaev, N., et al., 2020). This research will apply deep learning models to our proposed proactive chatbot.

Discussion: The history of deep learning has been written with resilience driven mainly by the desire to overcome limitations, so we have observed a continuous evolution of deep learning models. The significant limitation was the hardware, supercomputers were needed to process the models. With the evolution of the technology, it is now possible to train a small model on a regular computer. However, deep learning models still have limitations in representing complex human languages. Due to its unpredictability, mapping “slang” with its meaning and context is still tricky. Our investigation deals with students in London, many of whom are international students, with different accents, mispronunciations, their own slang and more.

For this reason, it is essential to create a flexible framework for using the model and to learn from the interaction between the chatbot and the students as a basis for making the system more effective and adapted to the students’ reality.

Although Deep Learning models has shown sustained growth during the last few years, Deep Learning models still have some limitations, especially in the context of a large dataset. For instance, the BERT question-answer model only works well in less than one page or ten paragraphs. For this reason, collaborative solutions have been implemented, using text similarity measurements to select only those paragraphs which are most similar in the corpus, and then applying the question-answer model to those few paragraphs. The following section describes the measurement of text similarity.

2.3.5 Similarity Text Measurement

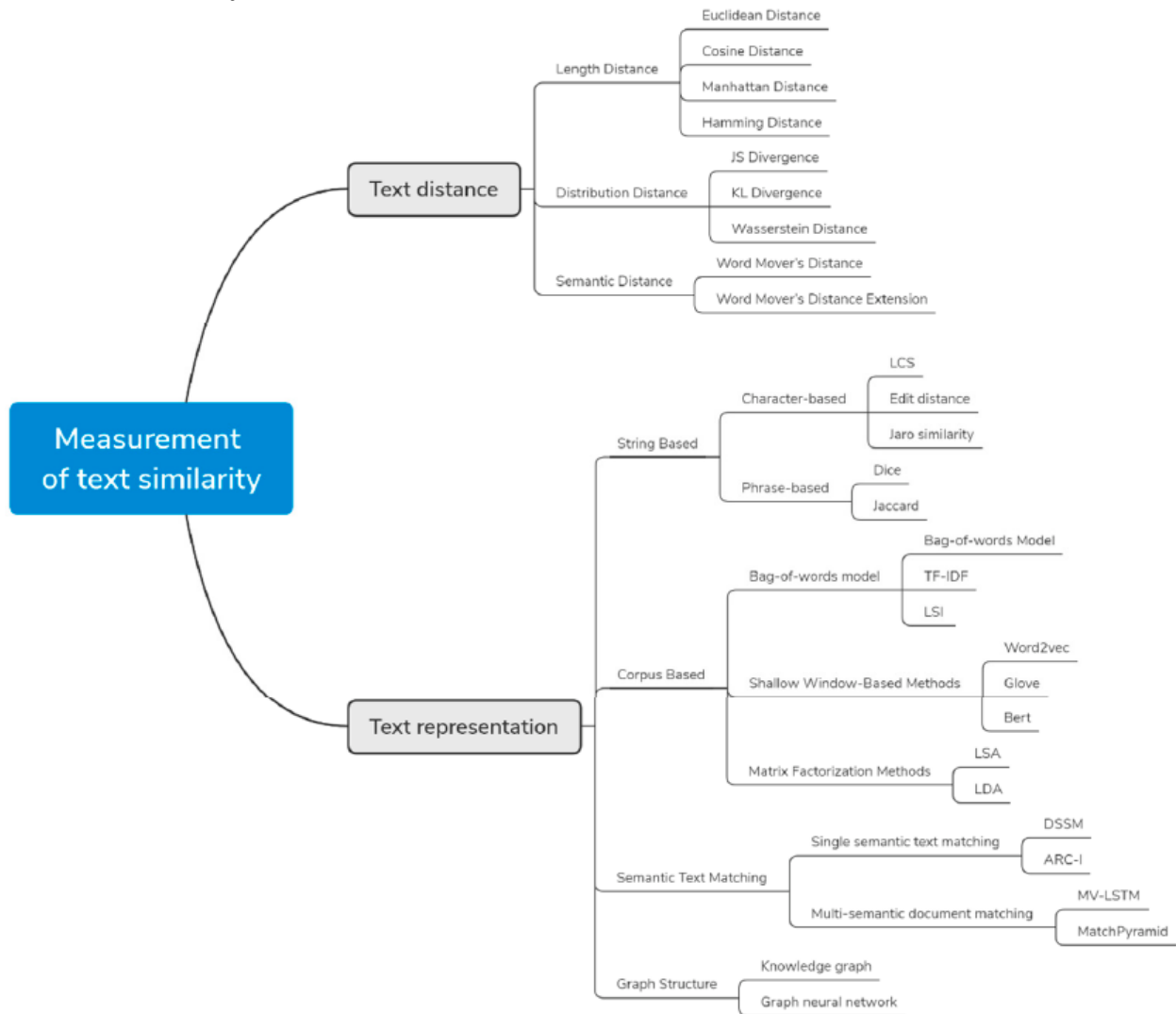


Figure 7 - Diagram of Measurement of text similarity

2.3.5.1 Text Distance

The Length, Distribution and Semantics distance can be seen in the Text distance similarity, which defines the semantic proximity of two documents from the viewpoint of distance.

Length Distance: Using numerical characteristics of the text to calculate the length distance of vector text has been the basis of text-similarity measurement. We briefly present the most popular types. **Euclidean Distance** is defined mathematically by *the straight-line distance between two points in Euclidean space* (Deza & Deza, 2009). **Cosine Distance:** *The cosine of the angle between two vectors* is used to calculate the cosine similarity of the two vectors. The size of the document might mean that using Euclidean distance might be far away from Euclid. Therefore, using the cosine distance to measure similarity gives better results. **Manhattan**

Distance: The sum of the absolute differences between the two vectors is called the Manhattan distance. Manhattan distance obtains the similarity after one-hot encoding (Deza & Deza, 2009). **Hamming Distance:** This is the measurement used to compare two binary data strings (Norouzi, M., et al., 2012). Below are the Euclidean, Cosine and Manhattan distance formulae.

$$d(S_a, S_b) = \sqrt{\sum_{i=1}^n (S_a^{(i)} - S_b^{(i)})^2}$$

Equation 2 - Euclidean Distance

$$\text{Sim}(S_a, S_b) = \cos \Theta = \frac{\vec{S}_a \cdot \vec{S}_b}{\|S_a\| \cdot \|S_b\|}$$

Equation 3 - Cosine Distance

$$\text{Sim}(x, y) = |x_1 - x_2| + |y_1 - y_2|$$

Equation 1 - Manhattan Distance

Distribution Distance: Length distance similarity presents two significant problems: First, it works well in the face of symmetrical problems, for instance $\text{Sim}(A, B) = \text{Sim}(B, A)$. However, it does not perform well in non-symmetrical cases, for example for question Q to retrieve answer A. Second, not knowing the statistical characteristics of the data can result in misjudging similarity when using length distance (Deza & Deza, 2009). The distribution distance is applied to compare the similarity between documents using distribution and to determine if the documents come from the same distribution. There are two popular methods for determining Distribution distance, JS divergence and KL divergence.

JS (Jensen–Shannon) Divergence: This is defined as a technique to size the similarity between two probability distributions (Manning, C.D. & Schütze, H., 1999). JS divergence and LDA (latent dirichlet allocation) are usually combined to establish the relationships among similar documents in distribution (Nielsen, F., 2010). **KL (Kullback–Leibler) Divergence** measures different levels of a probability distribution and another reference of the probability distribution (Kullback, S. & Leibler, R.A., 1951). **Wasserstein Distance** is used as a measure of the distance between two probability distributions (Weng, L., 2019). The formulae are as follows:

$$\text{JS}(P_1 \| P_2) = \frac{1}{2} \text{KL}(P_1 \| \frac{P_1 + P_2}{2}) + \frac{1}{2} \text{KL}(P_2 \| \frac{P_1 + P_2}{2})$$

Equation 5 - JS Divergence

$$d(p \| q) = \sum_{i=1}^n p(x) \log \frac{p(x)}{q(x)}$$

Equation 6 - KL Divergence

$$W(p_r, p_g) = \inf_{\gamma} \int \int (p_r, p_g) E_{(x,y) \sim \gamma} [|x - y|]$$

Equation 4 - Wasserstein Distance

Semantic Distance: Distribution measurements or length similarity measurements may be reasonably small when the text has no ordinary words. In such cases we calculate the distance at the semantic level (Kusner, M., et al., 2015). **Word Mover’s Distance:** The earth mover’s distance method is used to represent the text as a vector space (Andoni, A., et al., 2008). Its other function is to find in the semantic space the *minimum distance required for a word in one*

text, reach a word in another text, and reduce the cost of moving text 1 to text 2 (Wu, L., et al., 2018). **Word Mover's Distance Extension:** This uses Euclidean distance. Euclidean distance normalises every dimension in space as the same weight. Using the enhanced Mahalanobis distance takes into account (De Maesschalck, R., et al., 2000) in Euclidean distance place.

2.3.5.2 Text Representation

Text is represented as numerical features, which can be calculated directly. Texts could be similar in two different ways, semantically and/or lexically.

String-based methods: these are simple to calculate. For example, string similarity measures work on character composition and string sequences, which measures dissimilarity or similarity distance for approximate string comparison or matching between two text strings.

Character-based methods: character-based calculations of the similarity among characters in a document shows the similarity between texts. **LCS:** LCS matching (Irving, R.W. & Fraser, C.B., 1992) is the method used to measure the similarity between two strings (S_a , S_b). Taking each text as a string, LCS characterises their similarity on the basis of the length of the longest substring. **Edit distance:** The edit distance shows the required conversion string from S_a to S_b , by the minimum number of transformations. L-distance (Levenshtein, V.I., 1966) and D-distance (Damerau, F.J., 1964) are two ways of defining the editing distance. D-distance includes delete, replace, insert and adjacent exchange operations. On the other hand, there are three atomic operations of L-distance, namely delete, insert and replace. While L-distance deals with multiple editing errors, D-distance can only handle a single editing error due to the number of adjacent operations it includes. **Jaro similarity:** Jaro similarity represented for strings S_a and S_b is as follows (Winkler, W.E., 1990):

$$LCS(S_a, S_b) = \begin{cases} 0, & \text{if } S_a = 0 \text{ or } S_b = 0 \\ 1 + LCS(S_a - 1, S_b - 1), & \text{if } x[S_a] = y[S_b] \\ \max \begin{cases} LCS(S_a, S_b - 1) \\ LCS(S_a - 1, S_b) \end{cases}, & \text{if } x[S_a] \neq y[S_b] \end{cases}$$

Equation 8 - LCS Similarity

$$Sim = \begin{cases} 0, & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|S_a|} + \frac{m}{|S_b|} + \frac{m-t}{m} \right) \end{cases}$$

Equation 7 - Jaro Similarity

Phrase-based methods: The character-based method and phrase-based methods differ in the basic unit they use. For phrase-based methods, it is a phrase word, and two of the primary methods are the Jaccard and the dice coefficient. **Dice:** the comm (S_a , S_b) in the dice method shows the number of collinear phrases, namely the number of characters common to the strings

S_a and S_b (Dice, L.R., 1945). **Jaccard:** *the size of the intersection divided by the size of the union of two sets defines the Jaccard similarity* (Jaccard, P., 1912). The Jaccard method solves the similarity using a set; the similarity is smaller when the text is long. Consequently, Jaccard usually is normalised to calculate similarity. For instance, words can be minimised to the same root as English words.

$$\text{Dice}(S_a, S_b) = \frac{2 \times \text{comm}(S_a, S_b)}{\text{len}(S_a) + \text{len}(S_b)}$$

Equation 9 - Dice

$$S(S_a, S_b) = \frac{S_a \cap S_b}{S_a \cup S_b}$$

Equation 10 - Jaccard

Corpus-Based: *The crucial difference between string-based and corpus-based methods is that the corpus-based method obtains the basic information it requires from the text corpus to measure similarity.* The information can be a co-occurrence probability or a textual feature. The corpus-based approach has been represented using matrix factorisation methods, that is distributed representation and the bag-of-words model.

The Bag-of-Words Model: The text representation is like a mixture words in series, with no regard to how words appear in the document, and this is considered the basic idea behind the bag-of-words model (Wang, S. & Manning, C.D., 2012). TF-IDF, BOW (bag of words) and LSA (latent semantic analysis) are the main methods of using the bag-of-words model.

BOW: Count vectorisation is the core of the BOW method, and describes a document by enumerating the appearance of the number of words it contains by using these word counts to measure similarity (Salton, G. & Buckley, C., 1988).

TF-IDF (term frequency - inverse document frequency): This works on the basis of words which appear frequently in a document where many documents contain that word. Nevertheless, the words have no relevant meaning to the document (Robertson, S.E. & Walker, S., 1994).

$$tf-idf(w, d, D) = tf(w, d) \times idf(w, D)$$

where: $tf(w, d) = \text{Freq}(w, d)$

$$idf(w, D) = \log \frac{|D|}{N(w)}$$

Equation 11 - TF-IDF

Shallow Window-Based Methods: While it recognises the semantic distance between words, the bag of words model does not capture the semantic distance between words. Shallow window-based methods generate word vectors. In unstructured text with no mark, low-dimensional real vectors can be trained.

Several strategies have been developed to represent word vectors, and the three primary strategies are Word2vec, Glove and BERT. **Word2vec** has two pre-training models, which are the continuous word-bag model (continuous bag of words, CBOW) and the word-skipping model (skip-gram) (Rong, X., 2014). **Glove:** The word-skipping model (skip-gram) and the continuous bag of words model (CBOW) are the pre-training models of the Word2vec. Glove: The total word frequency statistics could be represented by Glove, which emphasises the semantic data of words by modelling the correlation of the words. Words with similar meanings see the words in similar contexts (Pennington, J., et al., 2014).

BERT: The bidirectional encoder representation from transformers is done in both directions, which captures the word context. The advanced versions use the pre-train approach, dealing with another sentence forecast and masked language model, which embeds sentence-level representation and expression in separate ways (Devlin, J., et al., 2018). The architecture is shown in **Figure 9** (Devlin, J., et al., 2018). Skip-gram and CBOW of Word2vec measures are described in **Figure 8** (Mikolov, T., et al., n.d.).

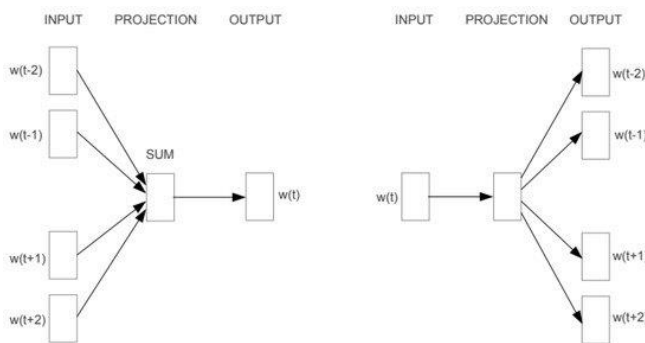


Figure 8 - Word2vec's demonstrate structures

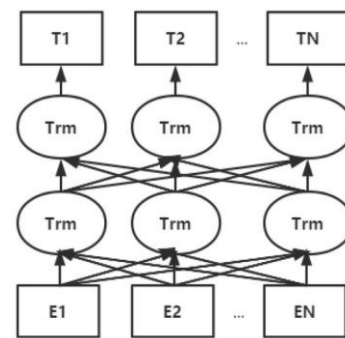


Figure 9 - Bert: Pre-training

The persistent bag-of-words (CBOW) design predicts *the current word based on the setting, and the skip-gram design predicts surrounding words given the current word* (Mikolov, T., et

al., n.d.). BERT's *representations are mutually conditioned on the left and right context in all layers* (Devlin, J., et al., 2018).

Matrix Factorisation Methods: returning to Matrix factorisation, we find LSA (latent semantic analysis), which is used as a strategy for creating low-dimensional word representations. Using low-rank estimations simplifies huge matrices which capture measurable data in the corpus. LSA: The LSA works on the premise of the measurable, comparable levels of word bag vector. (Deerwester, S., et al., 1990), (Kontostathis, A. & Pottenger, W.M., 2006) LSA creates the correspondence between high-dimensional lexicon space and low-dimensional latent semantic space using the particular value decomposition to measure the similarity within the desirable semantic space (Landauer, T.K. & Dumais, S.T., 1997), (Landauer, T.K., et al., 1998). **LDA (Latent Dirichlet Allocation):** a few text are expected for each document in ways that overlap points in the document. This is because the words in each document create the themes. Therefore, all topics will have a distinct distribution from each document, and all words will be affected by discrete distribution from each subject (Blei, D.M., et al., 2003).

Semantic Text Matching: Semantic similarity (Sahami, M. & Heilman, T.D., 2006) determines the similarity between text and document on the basis of their meaning instead of character matching. Referencing the LSA, the query embeds the hierarchical semantic structure, while the extraction of the document is done with deep learning. In this case the content is encoded to extract features so that a new expression is acquired (Li, Q., et al., 2019).

Single Semantic Text Matching: this principally consists of Architecture-I for matching two sentences (ARC-I), the convolutional latent semantic model (CDSSM), Architecture-II of the convolutional matching model (ARC-II) and the deep-structured semantic model (DSSM). **DSSM:** The DNN (Deep Neural Network) links the Title and Query to the low-latitude semantic vector, and the cosine distance computes the distance between the two semantic vectors. Finally, we train the semantic similarity model. This combines the Convolutional Neural Network (CNN) with DNN to compensate to a few degrees for the context loss in the DSSM. (Shen, Y., et al., 2014). The DSSM is presented in **Figure 10** (Shen, Y., et al., 2014). It is divided into three chunks: the feature extraction layer, the embedding layer and the SoftMax output layer.

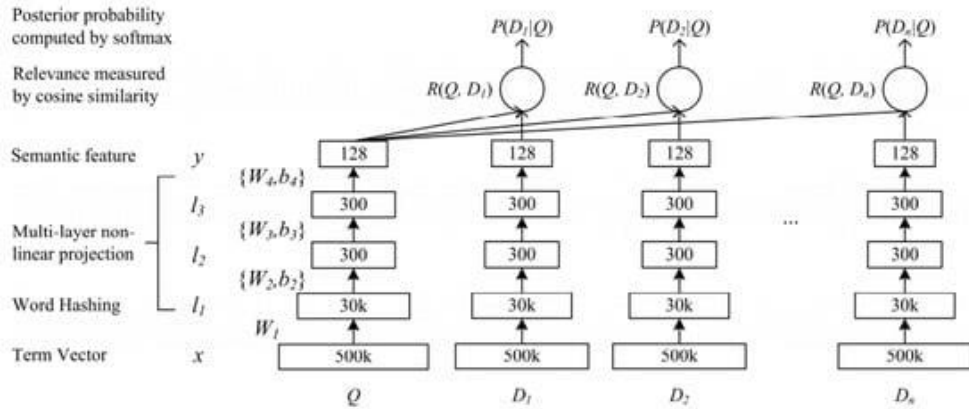


Figure 10 - The model structure of DSSM

Figure 10 Presentations of the DSSM utilise the DNN to map high-dimensional sparse text features into low-dimensional closely packed in a semantic space (Huang, P.S., et al., 2013).

ARC-I: The DSSM models are deficient in retaining context information, doc sequences and queries. The DSSM model includes the CNN module. Hence ARC-II and ARC-I are recommended. ARC-I may be seen as a representation learning-based model, and the ARC-II model has a place in the interactive learning model. Architecture-I (ARC-I) is presented in the **Figure 11**.

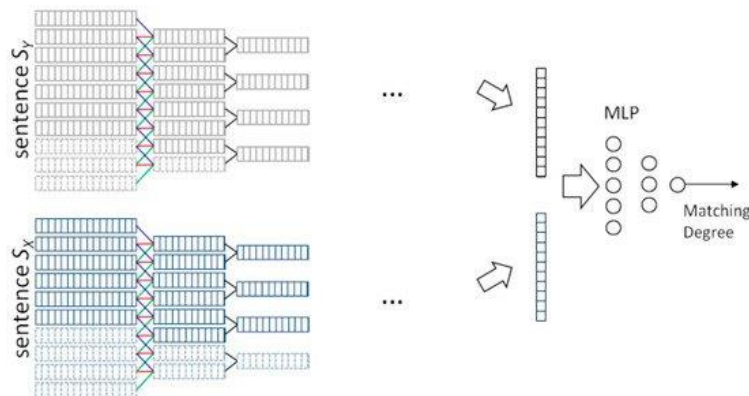


Figure 11 - Architecture-I for matching two sentences (Hu, B., et al., 2014)

Multi-Semantic Document Matching: When complicated sentences are condensed into a unique vector-based on single semantics, they cause a loss of essential local information. On the assumption of single semantics, a single-granularity vector is not enough to deal with a

chunk of text and find it fined. Before matching, it needs multi-semantic expression and does much interactive work to identify a few local similarities and synthesise the matching degree between texts.

MatchPyramid and MV-LSTM (multi-view bi-LSTM) constitute the strategy of the multi-semantic *MV-LSTM* generates positional sentence representations; MV-LSTM utilises the Bi-LSTM (bidirectional long and short-term memory) In order to mirror the text meaning in both directions at this point, particular bi-LSTM can get two hidden vectors for each location (Wan, S., et al., 2016).

MatchPyramid: Driven by CNN in image recognition, to construct a similarity matrix, we calculate the content and then that convolution to take out features. Into image recognition prepares the text matching (Pang, L., et al., 2016). MV-LSTM is illustrated in **Figure 12**, and MatchPyramid is illustrated in **Figure 13**.

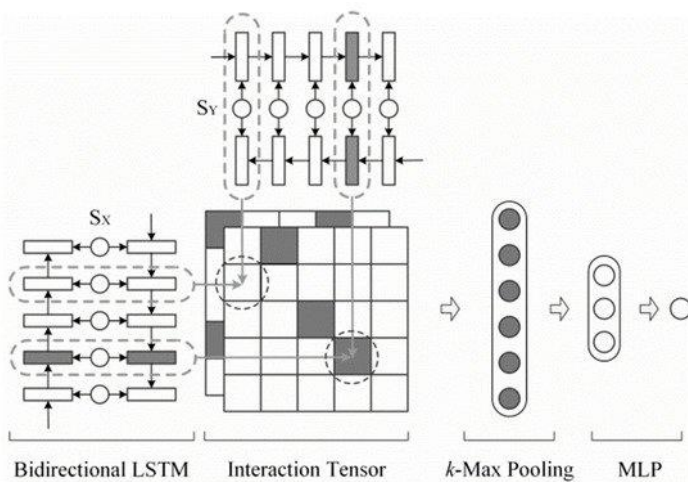


Figure 12 - multi-view bidirectional long and short-term memory

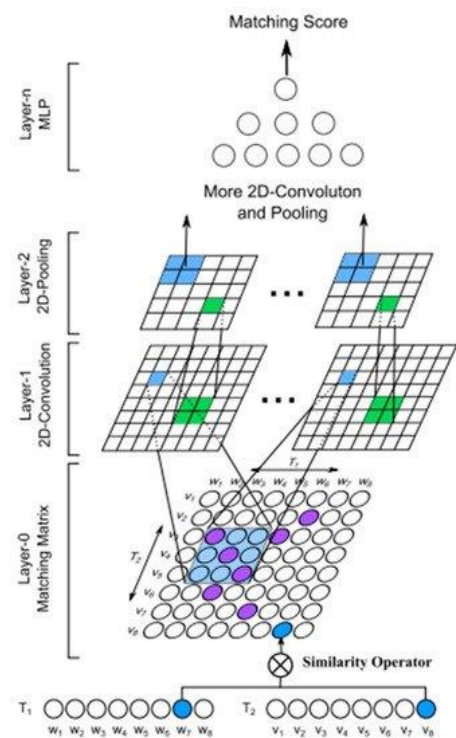


Figure 13 - MatchPyramid on text matching

Figure12 Multi-view bidirectional long and MV-LSTM (Hu, B., et al., 2014). **Figure13** an overview of MatchPyramid on text matching (Hu, B., et al., 2014).

Due to the extensive data training required with supervised text matching and the model training resources, the network becomes complicated (Liu, Z., et al., 2018).

Graph Structure: The industry and the scholarly world have been researching the structure of organised text data or graphs. One advantage of the graph-based method is that the joins between nodes are built up through the edges of graph structures to judge the degree of similarity between nodes. They are generally shown by graph neural network and knowledge graph representations.

Knowledge Graph representation learning: *this extends the connections in the knowledge graph into a persistent low-dimensional vector space through machine learning technology, maintaining the fundamental structure and properties of the first knowledge graph* (Chen, X., et al., 2020).

Graph Neural Network: to organise the hierarchical relationship of the data and infer the GNNs (graph neural networks) it is essential to utilise the model in the presence of many levels of associations and information (Gilmer, J., et al., 2017), (Zhou, J., et al., 2018). The GNNs could be seen as a connector model, which captures the reliance of the graph through the message transmission between the nodes of the graph (Vashishth, S., et al., 2020). Unlike the standard NN, the GNN holds a state that can show any of the information deep meaning related to its neighbourhood (Wu, Z., et al., 2020).

Discussion: The authentic text meaning is considered by the string-based methods; however, these are not flexible in unknown languages and domains. On the other hand, corpus-based methods with a statistical background could be inserted in various languages; however, there is a lack of real text meaning. Excellent performance is offered by methods based on semantic text matching, but then again, these require significant computational resources, and there is a gap in interpretability. Graph-structure methods focus on learning good graph representation. This section clearly shows the advantages and disadvantages of each method, but it is difficult to identify which is the best model; Nevertheless, promising results have been shown by the most used text distance and methods-based text representation compared to independent models. Next, we presented a description of the chatbots.

2.3.6 Chatbots

Chatbots could be defined as agent systems which engage in a conversation with humans using natural language. The first chatbots ELIZA (Weizenbaum, J., 1966) and ALICE (Wallace, R.

S., 2009), experimented with pre-programmed question –answer sequences that fooled people into thinking that they were talking to humans. More recently, the chatbot has been used in real-world applications, such as information retrieval, education and e-commerce. Over the years, the focus changed from achieving perfect human imitation to developing more valuable tools to help people in their tasks using natural language (Shawar, B. A. & Atwell, E., 2007). Research suggests that Internet users use social media to seek assistance more than calling or writing an e-mail (Xu, A., et al., 2017). The virtual assistant chatbot is a candidate for replacing old-style customer service (Brandtzaeg, P. B. & Følstad, A., 2017).

The number of criteria used to classify chatbots may vary. From a technological perspective, dialog systems are divided into two main categories. The high-level dialog system provides an understanding of context and a high level of analytical complexity. The low-level dialog system is a simple pattern-matching algorithm that presents a basic context understanding and a low level of analytical complexity (which means it mimics the conversation rather than understanding it) (Schumaker, R. P., et al., 2007). It is possible to suggest another vital distinction between ‘General Personal Assistant’ (such as Amazon’s Alexa, Apple’s Siri or Google’s Meena, which are mainly voice-driven chatbots) and ‘Specialised Digital Assistant’. The use of this last type of chatbot has increased significantly with the introduction of chatbot platforms such as WhatsApp, Skype, Facebook and LinkedIn which focus on a specific task, for instance, providing specific information or booking a flight; these are mainly text-based chatbots (Meisel, W., 2016). The typical interaction between human and chatbot is based on speech recognition or writing text. The chatbot is trained to understand the user’s input via voice or writing on the basis of a machine learning technique.

The Classification of Chatbot Models (Generative vs Retrieval-Based Models)

Retrieval-based models (low complexity): Necessary knowledge is pre-installed in a repository for responses. This could be done heuristically, for instance, as an ensemble of machine-learning classifiers or through rule-based expression matching. With this model, the system responds to a fixed set of data and does not generate new data (Shang, L., et al., 2015).

Generative models (high complexity): these generate new responses from scratch and do not rely on any pre-defined repository. This model is based on the machine translation technique, instead of an expected translation from one language to another. Generative models ‘translate’ a question or input into an answer, response or output (Yang, Z. & Hu, Z., 2017).

Framework models (moderate-complexity): the initial approach for designing our research model will be a combination of the retrieval-based and the generative model. We intend to have knowledge generated by students and use the generative model with sequence-to-sequence architecture.

Long v. Short Conversations

Long conversations (high-complexity): for long conversations, there is a need to memorise what has been said because there are multiple turns in the conversation. For example, coaching conversations are naturally long threads with various questions. The longer the conversation, the more complex it is to automate (Yang, Z. & Hu, Z., 2017).

There are also ***Short-Text Conversations (low-complexity)***. These conversations are designed to generate a single answer to a particular input, for instance, when the system receives an explicit enquiry (input) from a user and responds with a suitable answer (output).

MCP models (moderate-complexity). We intend to create a standard Coach/Mentor dialogue between the conversation agent and the Candidate, which could be neither short nor long.

Closed v. Open Domain

Open-domain (high-complexity): When a person says something, we never know where it will end up. One typical example is chats on websites such as LinkedIn, Facebook or Twitter. These are classically open domain – they could lead anywhere. There is an endless range of possible subjects, and the fact that a considerable quantity of world knowledge is required to generate a rational response make such conversations highly complex problems (Yang, Z. & Hu, Z., 2017).

Closed domain (low-complexity): there is a specific goal to achieve in closed domain conversations, resulting in a limited input. The best examples of closed domain problems are technical customer support and shopping assistants.

MCP models (moderate-complexity): The Proactive Chatbot research context is a naturally limited conversation domain, as a coach is led into a specific area of conversation. However, it is expected that Students will speak openly about their concerns in relation to that particular topic.

Discussion: In terms of advantages and disadvantages, the first model does not make grammatical mistakes because of its pre-processed repository. However, for the same reason, it is limited to what has been installed, meaning that it is unable to handle other cases. The retrieval-based model cannot refer to contextual object material such as to a person's name which has been given previously in the chat. Generative models can however use items encountered earlier in the input, giving users the illusion that they are chatting to a human. Generative models are complex to train. There is a high probability that they will make grammatical errors (particularly in extended sentences) and they frequently need large quantities of 'training data'.

Common challenges in building chatbots: There are some clear and not-so-clear challenges in building chatbots. Some of these are to be expected in the context of building a proactive chatbot for this research.

Incorporating Context

An example of *linguistic context* we find in extended conversations is that we keep track of the conversation and the contextual information exchanged. To generate sensible answers, the systems should combine both *linguistic context* and *physical context*. The most efficient method is to *embed* the chat into a vector. However, doing this in extended chats is challenging (Yang, Z. & Hu, Z., 2017).

Coherent Personality

When the generative model creates answers, the conversational agent should preferably generate reliable responses to semantically identical inputs. For instance, a user would like to have the same answer to "What do people call you?" and "What is your name?". This might appear to be a simple matter but building the "personality" into models is a complex problem. Even if systems produce linguistically accurate answers, they are not trained to produce answers which are semantically consistent. Frequently, one of the reasons is that they are "trained" on a variety of several sets of data from a number of dissimilar users. A model such as "A Persona-Based Neural Conversation Model" shows clear signs of a personality having been modelled (Yang, Z. & Hu, Z., 2017).

Model Evaluation

Asking whether a conversational agent fulfils its purpose or not is the most practical way to evaluate a model, for example, by testing whether its responses will help to solve the problem in each chat. A study done by Cornell University, “*How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation*”, revealed that none of the regularly used metrics correlated with human judgment (Yang, Z. & Hu, Z., 2017).

Diversity and Intention

The generative model presents an issue with generic responses. For instance, it is known that the responses “I am OK” and “No thanks” work for the same input cases. One reason is how the models are trained, which has to do with data and with the actual training objective/algorithm. *Academics have tried to artificially promote diversity through various objective functions (Sequence-to-sequence neural network models)* (Yang, Z. & Hu, Z., 2017). The framework research direction proposed is a hybrid approach between retrieval-based and generative models. Below is a short overview of chatbot frameworks.

2.3.7 Chatbots frameworks: A short overview

In the sociology of application users, there is a vital distinction between expert users and novice users. The latter require a more user-friendly form of virtual interface assistance that meets certain specifications to work with the novice users’ questions. Accordingly, Sansonnet, Leray and Martin (Sansonnet, J. P., et al., 2006) outline a basic framework for how the virtual assistant can meet the requirements. In the first step, the system should understand the user’s thinking process; the understanding function is “The Dialogical Agent”. The second step is to answer user questions by accessing the knowledge base known as “The Rational Agent”. Finally, the system should be able to earn the user’s trust through the presence of, “The Embodied Agent” (Sansonnet, J. P., et al., 2006).

An inspiring framework evaluation method was proposed by Kuligowska (Kuligowska, K., 2015) to take into account diverse aspects of the way chatbots function. It contains a “Visual Look” and a “Knowledge Base” containing general and specialised information, because a dialogical agent must be able to answer a set of general knowledge questions; but also questions on the topic they specialise in. It should also have as part of the chatbot personality “Language

Skills and Context Sensitiveness” and “Conversational Abilities” to enable it to deal with the context of the conversation, give feedback, and to maintain a harmonious conversation, “Personality Traits” to present a distinct personality with different aspects, emotions and physiology, “Have answers in adverse situations” to deal with the user’s writing errors and answer provocative questions diplomatically), a “Possibility of Rating Chatbot” function (as asking for feedback on communication is a critical function for determining the chatbot’s efficacy) (Kuligowska, K., 2015).

Later systems proposed for chatbots emphasise the prerequisite to develop the chatbot with psychosocial factors and adding the users’ cognitive and social behaviour. The classic machine usability definition is not enough to build a successful chatbot application. Chatbots should build a connection with users, in a way that is both practical and enjoyable. It should therefore be Flexible (able to adapt to the personality of the user), Affective (paying attention to the quality of the relationship), Communicative (keeping things simple for the user) and finally, Autonomous (able to take the initiative in the conversation). Therefore, to be successful and believable, the chatbot requires a body, a personality and a mind. The mind should be developed with cognitive and problem-solving abilities, social capabilities and affective sensitivity (so that it can respond to the user’s emotional state). The chatbot’s personality defines the interaction style of the human-machine, and the mix of these three aspects creates the chatbot’s behaviour (De Angeli, A., et al., 2001).

Discussion: The use of a chatbot in University settings enables greater flexibility of interaction with students, as it can answer pertinent questions and assist students in specific areas such as helping them with their academic subjects. However, even though chatbots are becoming more human-like, scalable and relatively cost-effective for universities, they still have communication limitations in interpreting and contextualising questions from students. Despite these limitations, for universities which intend to assist their students, a chatbot is still the most practical, low-cost option.

2.4 Big Picture of the Proactive Chatbot Framework

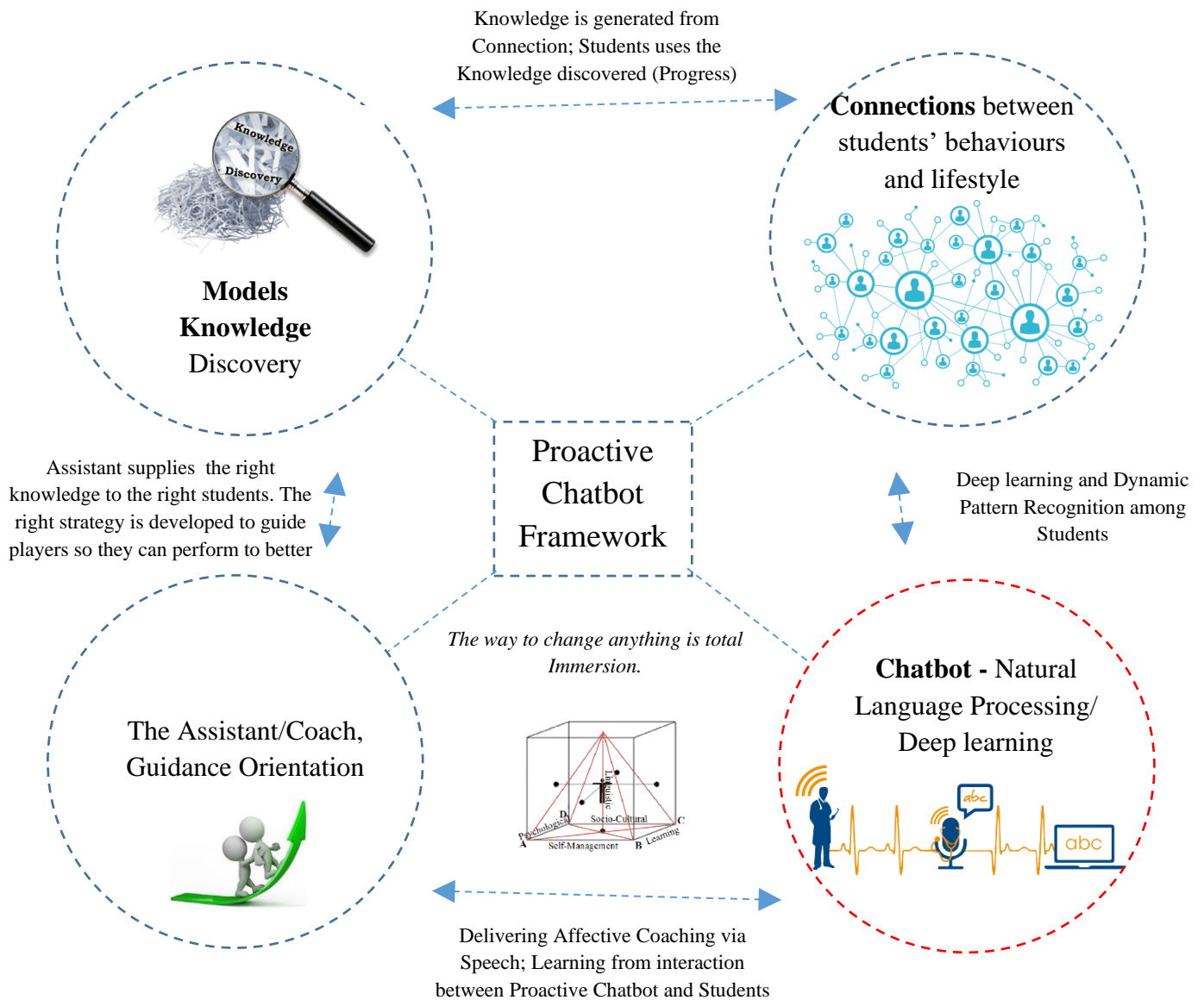


Figure 14 - Proactive Chatbot Diagram, (Almada's Diagram, 2020)

We have now finished this literature review showing the landscape of our proposed “Proactive Chatbot Framework”. A chapter summary is presented below.

2.5 Summary

This chapter reviews the relevant literature, laying the psychological foundation for the Assistant/Coach. It begins with the science behind why and how a practical Assistant/Coach can improve a candidate’s performance. The literature proves that it is hard to change, but it is possible to create a new habit. Moreover, it focuses on five proven principles of modern psychology. These are Principle 1 - the “locus of control” and the “bias toward action”. Some

individuals believe they control their lives (internal locus of control) while others believe life is just happening to them. They are a victim of whatever might happen (external locus of control). Principle 2 - “behavioural flexibility”: our brain never stops creating new connections and learning new things. Principle 3 - do good, be good; to change we should do things, act, and then our mind will follow. Principle 4 - The golden rule of habits: we cannot control how we might feel, but we can always choose how we react or behave. Principle 5 - Activation Energy: the first step is the key to creating any change and reaching our full potential.

The chapter continues by presenting the different assistant/coach styles. The main difference lies in how they approach the instruction they give. The coaches discussed include Richard Bandler, who studies Neuro-Linguistic Programming (NLP), Tony Robbins, a practising psychologist, and Bob Proctor, a thinker who teaches the changing paradigm. They tend to draw on external references, experts in the field or people who have successfully done that activity, and model the technique applied. The group of Assistants presented in our literature review such as Vishen Lakhiani focus on the inner self. They study meditation, focusing on making an outside change by changing what is inside. Wayne Dyer was a spiritual mentor, and Nietzsche’s philosophy teaches us that to find ourselves, to get in touch with our inner genius, we must walk a path no one has walked before, as we are unique, and no one can walk that path on our behalf. Finding ourselves is finding our uniqueness, that unique set of values and things we truly love and which represent us.

The next section briefly shows that an appropriately proactive chatbot must be based on multimodality. To individualise the chatbot’s interaction with students, it draws on 16 distinct personalities developed by Carl Jung, categorised by their preference for wide-ranging behaviours, three areas of preferences and dichotomies: These include Extroverted (E) vs. Introverted (I); Sensing (S) vs. Intuition (N); and Thinking (T) vs. Feeling (F). Dr Isabel B. Myers, developed this further, adding one more field, namely Judging (J) vs. Perceiving (P). The chapter closed by presenting Knowledge Discovery in Databases, which is a combination of technologies for data management including among others machine learning and data warehousing. This technology drives the global development of learning valuable knowledge from big data, in which data mining technology is one element.. The chapter closes by presenting chatbots in Universities, suggesting that using a chatbot can serve as a valuable and effective solution for students because of the chatbot’s conversation synchronicity. It could reduce the University’s costs and staff workloads.

The next section starts by presenting the history and evolution of Artificial Intelligence, which aims to serve as a force for the good of society as a whole. Research suggests that when a machine can mimic human cognitive processes or functions related to human minds, we are in the presence of an intelligent machine. This section then outlines the distinction between Artificial Intelligence, Machine Learning and Deep Learning. Natural Language Processing is also discussed. Artificial Intelligence has many fields, including Natural Language Processing (NLP), which focuses on communications between computers and humans. NLP faces challenges, including empowering computers to derive from people or natural language input, and understanding and generating natural language. NLP allows computers to understand the context and meaning of our words.

The literature review chapter continues with a Deep learning section. Deep Learning was born from Machine learning, or more specifically, from deep neural networks, where the architecture is fully connected, presenting an Input layer, one or many Hidden layers and an Output layer. Then we present a basic formula to calculate the deep learning models. Lastly, we present a brief history of Deep Learning models in recent years. Then the measurement of text similarity is discussed. Finally, the chatbot is presented. Chatbots could be defined as agent systems which engage in conversations with humans using natural language. This section then introduces the history and evolution of the chatbot, outlining the differences between Generative Models (High Complexity) and Retrieval-Based Models (Low Complexity), Long vs. Short Conversations, Closed vs Open Domains and finally outlining common challenges in building chatbots. The next section presents a short overview of Chatbot frameworks. In the next chapter, we present the methodology used in this thesis.

Chapter 3 Methodology

3.1 Introduction

This study comprises two main parts: a proposed academic model and a framework to assist students using Artificial Intelligence. Therefore, the process of collecting and analysing data was based on a mixed methodology of the qualitative and quantitative research to cover the different phases of the research. The data was collected in the Universidade Católica de Angola, Luanda.

Data was collected about students' life, psychology, self-responsibility, sociology, communication, learning and health & wellbeing. The data was collected through students a self-evaluation questionnaire, and was used to identify correlations among the factors by predict the student result using Artificial Intelligence to generate the model. At this point, we recognise an accurate and realistic prediction of students' performance should consider more factors and variables such as lecturer expertise, university conditions, the students' family and other factors that directly affect the students' academic performance. However, our aim is to find the correlation among students' controllable academic factors, which affect their results.

We will use mixed methods, i.e. both quantitative and qualitative, to collect quantitative and qualitative data. For Artificial Intelligence, the well-known Data Mining methodology CRISP-DM (CROSS Industry Standard Process for Data Mining) was used, and the software used to analyse and process the data was SPSS Modeler (IBM Software Business Analytics, 2010). In addition, Scrum and Agile were used. The Schwaber, K & Beedle, M (2001) Software Development method was used to achieve the research aim and objectives quickly (Schwaber & Beedle, 2001).

The interview questions for the students were presented mainly in the form of a questionnaire and sought to identify their best skills (factors that affect students' performance). For ethical purposes in the worse-case scenario, for instance, if a student reports long-term depression or suicidal feelings, a chatbot will advise him/her to go to a doctor or expert immediately. For each question, there is a diagnostic description as appropriate. It should be considered that chatbots do not replace experts on the field or give any medical guaranties.

Differences between Qualitative and Quantitative Research Approaches	
Qualitative	Quantitative
<i>Understand and interpret human perspectives</i>	<i>Comparisons or correlations of population attributes</i>
<i>Less generalizable to populations</i>	<i>Generalization to populations</i>
<i>Rich descriptions</i>	<i>Numerical summaries</i>
<i>Depth</i>	<i>Breadth</i>
<i>Small sample</i>	<i>Large sample</i>
<i>Selection of procedures to establish trust in the findings</i>	<i>Prescribed process to establish validity and reliability</i>

Table 2 - Qualitative vs Quantitative research (British Library, 2015)

These research approaches will be discussed further, starting with qualitative and quantitative research.

3.1.1 Quantitative Research

Quantitative research aims to clarify and analyse data, looking at conceivable consequences and outputs. It differs from qualitative research in various ways, including the aims it seeks to achieve and the methodology and designs it employs. Some of the significant distinctions are presented in the table above.

In interpreting and understanding human phenomena, we usually use all the available populations. For example, by studying the effect of bullying, doing a qualitative study, we will want to approach and observe the victim of bullying and the bully in the field. On the other hand, live quantitative research ends up generalising, using large samples to analyse the data numerically, making comparisons and discovering relationships and patterns in the population studied. For instance, quantitative research to investigate bullying among students could compare students who are considered victims of bullying to students who are not considered victims of bullying, and use an attitudinal survey to quantify the effects on the bully and the victim (British Library, 2015).

3.1.2 Qualitative Research

Einstein said, “*not everything that can be counted counts, and not everything that counts can be counted*”. Qualitative research focuses on the subjectivity of the human experience,

understanding the meaning of the phenomenon in a descriptive and in-depth way rather than focusing on the general. To build the PS2CLH model during the course of this research we read articles, papers and other research on the use of chatbots in Universities, educational conferences, clothing magazines and market research reports, observation interview documents and literature reviews, and we also observed students' behaviour.

The qualitative research carried out for this thesis focuses on students' behaviours and lifestyles, knowing their moods, their thinking, and the choices they make. These data were captured during the observation period (after which they were given a questionnaire to fill in) then find a way to embed the data into the Proactive Chatbot application. The research subjects were encouraged to clarify the reasons for their responses, revealing their true motivation and behavioural triggers (British Library, 2015).

3.1.3 CRISP-DM methodology

To process and transform the data collected into a usable form, we used the Cross-industry standard process CRISP-DM for data mining. We were referencing (IBM Software Business Analytics (2010) *CRISP-DM 1.0: Step-by-step data mining guide*).

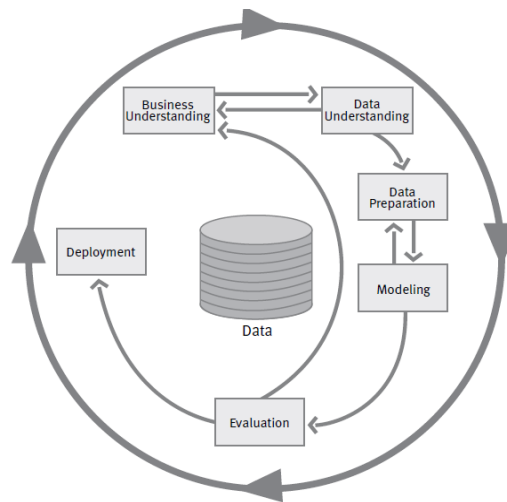


Figure 15 - CRISP-DM reference model (IBM Software Business Analytics, 2010)

The CRISP-DM methodology is divided into six phases for the execution of a data mining system. These phases are as follows

1. *Understanding of the business* (universities in our case): Understanding the goals and company requirements from a business perspective, turning them into a data mining application and subsequently developing a plan to solve the problem.
2. *Understanding the data*: The main aim of this phase is to understand the data.
3. *Data Preparation*: Creation of extraction methods, and cleaning and processing the data for use by data mining algorithms.
4. *Modelling*: The selection of algorithm(s) to be used for an efficient model process. Many algorithms require a clean data sample, as anything else might cause multiple returns to the preparation stage.
5. *Evaluation*: The analyst evaluates many models, finishing the modelling phase. Now, the purpose is to evaluate the models with the understanding of the business, checking that there are no gaps or inconsistencies concerning the principles of the business.
6. *Deployment*: The construction and proof of the model constitute one more step toward making the information produced accessible. Since the development of specific model, it is possible to deploy the model in several ways until the publication of a report for internal use (IBM Software Business Analytics, 2010).

3.1.4 Scrum Methodology

London Metropolitan University regulations ask an MPhil/PhD Student to report every other week. The author should present a report showing the progress of the research. For this reason, it is proposed that Scrum should be used in software research. Agile – Scrum Methodology will therefore be used to develop the proactive chatbot framework.

Scrum is an agile development methodology focused on teamwork, with the team's self-managed and active participation by the end-user. In our case, this will be the research tutor. Agile methodologies have emerged with the purpose of "simplifying procedures" in the software development process, enabling teams to be more adaptable, rapidly responding to constant changes in software projects, and allowing even late changes in requirements or in the project's scope (Scrum Methodology, 2013).

The tutor knows every step toward the main research aim because there are constant delivery features which are already 100% developed. He actively participates in the project, bringing his knowledge to the research, referencing Scrum Methodology and Agile Scrum Methodologies (Scrum Methodology, 2013).

Adapting Scrum to this research is likely to reduce the time required for the research. The team has short daily meetings (Sprint), always at the same time, in which minor problems and small ratifications are presented to be solved rather than allowing them to build into bigger problems later (Hicks, M. & Foster, S., 2010).

In their paper, the research *"Adapting and Using Scrum in a Software Research and Development Laboratory"* present two main goals to explain their success in using Scrum methodology. Firstly, the aim was to produce high-quality research results working collaboratively, and second to help independent researcher's students been capable of working at research labs or home (Lima, I. R., et al., 2012). Below we present an overview of the Research Thesis Methodology.

3.1.5 Research Design or Thesis Methodology

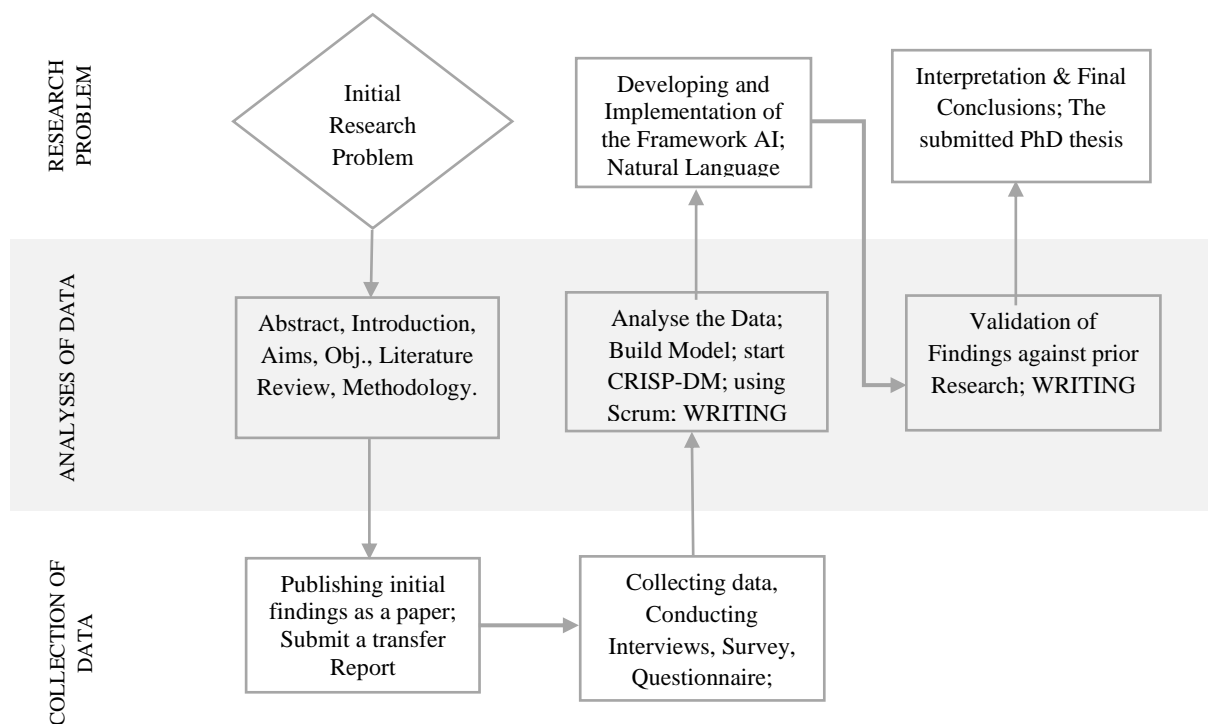


Figure 16 - Phases of the Research

The diagram above begins with the initial research problem. Initially, we need to specify the domain of the research, so we need to explore the existing literature. At this stage, the ideas are vague and abstract, so we try to clarify them by defining the real problem we intend to investigate. The definition of the problem was presented in Chapter 1, and is the lack of efficient, scalable and inexpensive ways to individually assist university students in the areas and factors which affect their academic performance and which they can control. This leads to the research question: *“How to develop a specific model to enhance academic performance that incorporates all the controllable academic factors, and how to further utilise the model in a framework for implementing a readily available student learning assistant tool”*.

Having the initial research problem and question defined, the next step is to build a strategy to tackle the problem and answer the question, which requires developing a clear research abstract, introduction, aims and objectives. In addition, it is essential to see what other researchers have done in the field by going through the literature, finding any gaps and developing a methodology for the research.

The next step is to summarise the findings of the literature and propose a scientific paper at a conference. Writing a scientific essay at this stage provides a basis for an external evaluation of our research idea. The acceptance or rejection of the results presented in the paper submitted show the potential of the research and the level of support for the London Metropolitan University “transfer report”. This is followed by the data collection phase, which uses a web questionnaire and conducts interviews to collect qualitative and quantitative data. We intend to collect data from many students to create a credible model and obtain sufficient evidence to support our research.

The collected data is then analysed and then organised to turn it into information. Next, we use statistics and algorithms to build the model, following the CRISP-DM methodology. The next step in the research design is to build on the previous step and use the Scrum methodology to develop and implement the AI framework. The final stage is to analyse the results and validate the model, and then compare it with existing work and then validate it against current solution, answer the research question and deal with the initial research problem. Then we write the thesis, organising the description of all the steps in chapters.

The conclusions of the thesis are then re-evaluated against the prior research. Finally, the final version of the PhD thesis is submitted.

3.1.6 Research Approach

In this research, we take two primary approaches. The first is exploratory research, which is by nature qualitative and investigates subjects and research questions that have not been studied in depth (George. T., 2022). The second is explanatory research, a method that explores why something occurs when limited information is available, and explores how and why a specific phenomenon occurs and predicts future occurrences (George. T., & Merkus. J., 2022)

Due to the nature of our research, we use a hybrid approach, starting with an exploratory approach. This exploratory approach helps us to identify the controllable factors that affect students' performance. This stage is multidisciplinary, encompassing psychology, sociology, communication, learning, and health and well-being. In the same way, we also explore the artificial intelligence and natural language processing fields, aiming to create a basis for developing a questionnaire for data collection. We also seek to identify and explain the correlation among the controllable factors by creating a model and predicting students' performance, aiming to use the model in a framework for implementing a readily available student learning assistant tool.

3.2 Methodology to find the correlation among the PS2CLH model factors

This section presents an overview of the steps or methods used to find the correlation among the PS2CLH model factors by predicting students' academic results to create students' profiles. In this chapter, we are presenting "WHAT" we have done to answer the Research Question. The following Chapters 4 and 5 present "HOW" and "WHY" we developed this solution, so further details will be offered in those chapters.

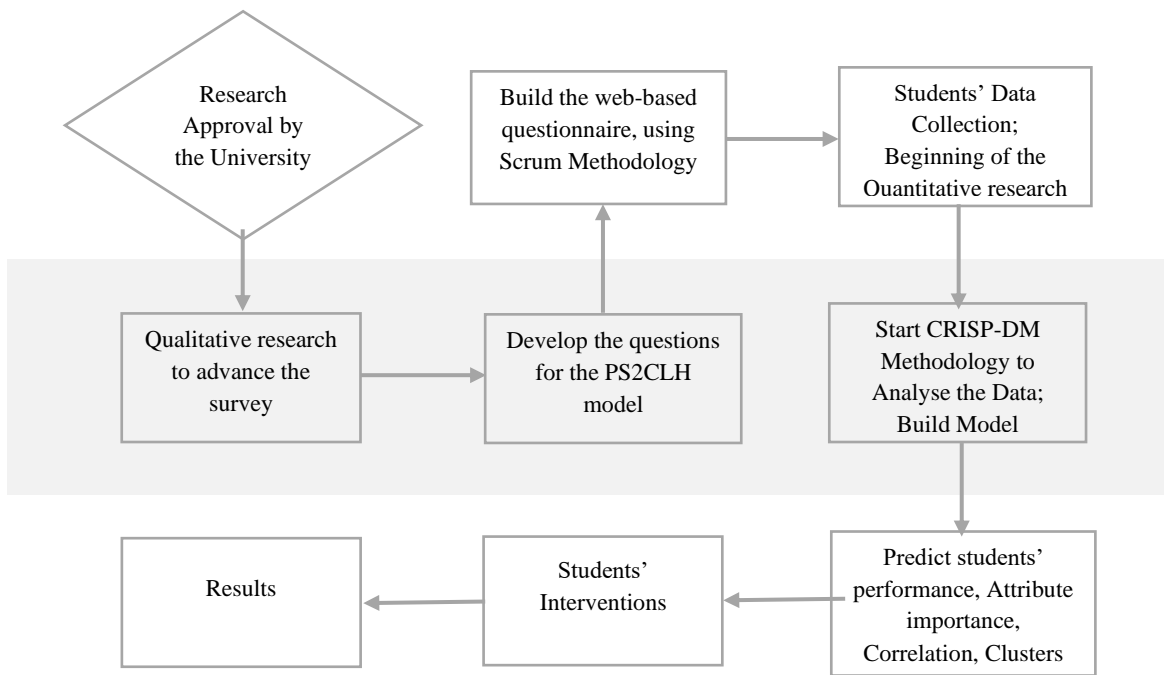


Figure 17 - Methodology to find the correlation among the PS2CLH model factors

Project Approval by the University: The first action to be taken was to seek the University's ethical approval and approval for data collection and access to laboratories. The University Directorate had to have some involvement in the research, as both students and professors would give more credibility to the project and want to participate. The other reason why it is essential for University management to be part of the project is that more attention will be paid to problems that arise during the research. With the approval and support of the University, we moved on to the second phase of the research, which is qualitative research.

Qualitative research to develop the survey: Before formulating the questions for a questionnaire, a priority is to carry out qualitative research in the country where the University in question is located, which should involve searching books, articles, magazines and studies related to our research. Those works are designed to assist students to improve their academic performance. Qualitative research should focus on issues relevant to the University and its students. Interviews should be conducted with students, University staff and professors, focusing on students' behaviour both inside and outside the University. We will be looking specifically at the factors which are controllable by students, which affect their academic performance in the areas of personality, sense of responsibility, social class, communication, learning and health and well-being. After this phase, we move on to formulating the questions on the basis of the results of the qualitative research.

Building the questions for the PS2CLH model: Questions for the PS2CLH model should be clear and straightforward so that students do not need outside help. They should also be intuitive so that the student can make a self-assessment when answering the questionnaire. Given that the questions fall into six areas (psychology, self-responsibility, sociology, communication, learning and health and well-being), the number of questions for each area should be between 4 and 7. The selection of these questions must be on the basis of the results of the qualitative research, and particularly questions that have the greatest impact on the students' academic performance. The number of times a subject is mentioned shows its importance. We can also get a sense of the questions that most impact students by looking at the pattern of responses from students and lecturers. After selecting the questions, we will build the web-based questionnaire.

Building the web-based questionnaire using Scrum methodology: We used the Scrum methodology to develop the web-based questionnaire. The questions were to be multiple-choice, and the radio button was used for student responses. The questionnaire was designed so that the value of each question depends on the student's answer. We also developed a database using SQL to store the data produced by student responses. Once the questionnaire was completed, we proceeded to the stage of collecting data from students.

Data Collection from Students; Beginning of Quantitative Research: Due to the lack of studies in this field in Angola, the aim was to start with a large number of questions and then reduce these to the more important ones. Two experiments were therefore conducted. The fundamental difference between these experiments was the number of questions and the fact that students intervened in the second experiment. The number of questions in each category in the first questionnaire will be 16 Psychology, 13 Self-management, 15 Sociology, 9 Communication, 17 Learning and 5 multiple choice questions, with a 5-point Likert scale response. The aim will be to conduct the research with 600 students from the Universidade Católica de Angola, with the University's permission, from March to May, between 07:00 and 18:00. The students will be defined as those students who are studying at the University on the day of the survey. Participants will be given 30 minutes to complete the form. The second questionnaire will consist of 7 Psychology, 8 Self-responsibility, 12 Sociology, 7 Communication, 8 Learning, 6 Health & well-being, and 5 multiple choice, with a 5-point Likert scale for responses. The aim will be to conduct the research with 500 students from the Universidade Católica de Angola with the university's permission, from September 2nd to

November 28th between 07:00 and 22:00. Students are defined as those who are studying at the University on the day of the survey. Participants will be given 20 minutes to complete the form. After outlining the data collection process, we will move on to data analysis and the construction of the predictive model.

Using CRISP-DM Methodology to Analyse the Data and Build the Model: We will perform the data analysis using the CRISP-DM methodology with student responses. We will use the advanced statistics tool SPSS Modeler for data analysis and predictive model creation. The data from the database will be converted to an Excel file. The different types of fields will be scaled from 1 to 5. At the same time, a second experiment will be conducted.

Predicting students' performance, Attribute importance, Correlation and Clusters: After the creation of the model, we will programme it and add it to the application so that as soon as the student fills in the questionnaire, the system automatically produce a forecast of his academic result using the PS2CLH model as a reference. In future, the predictions of student results can be shared in the lecturer's profile (with the student's authorisation).

Student Interventions: With the knowledge produced by data processing and data analysis, we see in the cluster a clear direction for our interventions. According to Oxford Languages, 'A cluster is a group of similar things or people positioned or occurring closely together'. (<https://languages.oup.com/>). In our case, we have seven clusters, and a cluster of students is a group of students with a similar number of academic controllable factors that affect their performance. The interventions aim to assist students by reducing the number of factors affecting their performance, helping them to overcome their limitations and develop new study habits so that they can go to the highest clusters, which are the 6th and 7th. Having described the interventions, we move on to the evaluation of the results.

Results: The results of the PS2CLH model. In this way we close the description of this methodology to find the correlation among the PS2CLH model factors and start on the method to build the proposed chatbot framework. The development of the proposed framework is now presented.

3.3 The method for building the framework of the proactive chatbot for students

This section presents the method for building the proactive chatbot for students based on the PS2CLH model. In this Chapter (3), we have focused exclusively on “WHAT” we have done to answer the research question. The following two Chapters (4 and 5) present “HOW” and “WHY” this solution was chosen; a more detailed explanation of the framework will therefore be offered in Chapter 5.

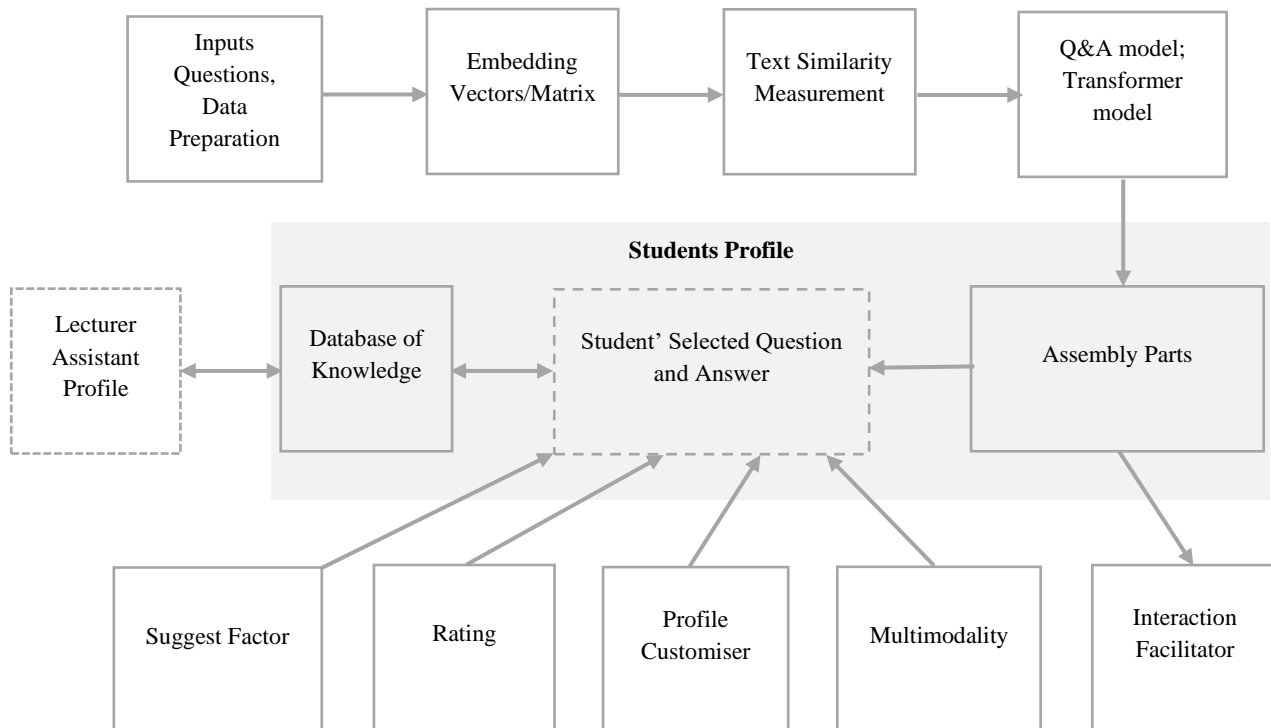


Figure 18 - Method used to build the framework designed for a proactive chatbot for students

Inputs/Questions, Data Preparation: the chatbot framework starts by receiving students’ questions in writing or orally, and these are then transformed into a matrix of vectors using embedding and vectors. These steps transform human language into computer language. This chatbot’s framework can receive students’ inputs orally and in writing. In both cases, the framework implements syntax and semantics, which are the broad categories into which NLP techniques fall.

Embedding, Vector/Matrix: for a machine to understand human language, it must be converted from text to numbers. There are different ways to represent a word. Word embeddings are vector representations of a specific word.

Text Similarity Measurement: there are multiple ways to determine text similarity in a document. However, in this research we are going to use Cosine similarity measurement. Cosine Similarity is simple to implement, and it quantifies the similarity between vectors or matrices by measuring it. The similarity between vectors is the resulting cosine of the angle between vectors (Gunawan, D., et al., 2018). We will explore this topic further in Chapter 5.

Q&A model; Transformer model: Having evaluated the student's question using cosine similarity, we will create a vector with the results of the ten most similar questions in the knowledge data. Each question is linked to an answer in the repository. Then we apply the transformer model question - answer to the ten most similar answers to predict the right answer. Transformer models are based on the attention mechanism which helped enhance the performance of neural networks machine attention (Vaswani, A., et al., 2017). We will explore this topic further in chapter 5.

Assembly Parts: The model then goes through those ten answers and chooses the answer according to the student's initial question. At this point, the system selects three possible equal questions in the knowledge data as the student's initial question, and the Q&A model gives the first potential one, and the other two questions are the second and third most likely. The fourth option is "none", which means that none of the above three questions is the question that the student would like to know the answer to. Each of the above four options is represented by a number. When a student chooses a number or option different from "none", the chatbot sends the answer to the selected option.

Interaction Facilitator: If a student chooses "none" as an answer, the system sends that question to the Lecturer's or Assistant's profile. When they answer the student's question, the system saves the question into the knowledge base. As soon as the student enters the chatbot, he/she will receive the answer.

Profile Customiser: When completing the questionnaire, the student can select the type of personality they would like the chatbot to have. In the questionnaire, there are four pairs of radio buttons, categorised by their preference for wide-ranging behaviours, and three areas of preferences and dichotomies: Extroverted (E) vs. Introverted (I), Sensing (S) vs. Intuitive (N) and Thinking (T) vs. Feeling (F). Dr Isabel B. Myers, continuing Jung's work, added one more field as a fourth antagonism pair; namely Judging (J) vs. Perceiving (P) (Briggs, M., 1980).

The result of the permutations of the four dichotomies are 16 distinct personalities. This feature plays a vital role in building the proactive chatbot.

Multimodality: Using the proactive chatbot and applying multimodality to students' learning process can retain students' attention and explain content in different ways, using text, image, video and audio to assist students and effectively improve their learning experience.

Rating: When the student receives the answer to their question, the chatbot sends a rating scale from 1 to 5 where 1 indicates that the student is not satisfied with the answer and 5 that the student is satisfied. The chatbot sends the student's feedback to the knowledge base so that the chatbot knows the best and worst answers. This student rating will count, so when other students ask a similar question, the system will give questions a weight, which we call bias. This way, in the future, lecturers and assistants will be able to delete or improve the worst answers in their profiles. After the chatbot has had many interactions with students, all the conversations are saved, and the chatbot transformer model is improved by retraining it with this new data.

Suggest Factor: If the student chooses one of the three questions sent by the chatbot, the chatbot goes to the knowledge base and sends the answer associated with that question to the chat. In this methodology, we mentioned that after predicting student results, we correlate the variables to analyse how they interact. To create a correlation table, each variable will have a vector of the three most correlated variables. For example, suppose the question is a question from the PS2CLH model. In that case, the chatbot will proactively go to the table of correlations between the variables and suggest other correlated factors that the student should also pay attention.

Knowledge Database: The knowledge database contains the questions and answers and all important information related to them. The Lecturer and the Assistant can also add a new subject and add questions and answers to the JSON file.

Profiles: These are the profiles of the student, the lecturer and the assistant, and these are presented in Chapter 5.

3.4 Summary

This chapter starts by outlining the general methodologies used in this thesis, which are: Quantitative research, Qualitative research, CRISP-DM for the data mining methodology and Scrum Methodology. It then presents the methodology for finding the correlation among the PS2CLH model factors. It finally introduces a method for building the proactive chatbot framework to assist students, using the results of the previous methodology. In a nutshell, we started by investigating the literature to build the PS2CLH model that combines the perspectives of personality, sense of responsibility taking care of oneself, social class, communication, learning and health and well-being to facilitate a student-controllable learning factor model. Then we developed a methodology to determine the correlation among the PS2CLH model factors and created a method to build a framework designed for the proactive chatbot for students. Finally, we built an application that incorporates a proactive chatbot that could potentially assist students. The next chapter proposes the PS2CLH model.

Chapter 4 The PS2CLH model: Results and Discussion

4.1 Introduction

The factors influencing a student's academic performance have been a focus of research for decades. Consequently, there is an overwhelming number of studies in this field. Four researchers have presented the most influential work.

The first of these is Professor John Hattie, one of the world's leading education researchers in this field. His ongoing research, *Visible Learning* (Hattie, J., 2009), focuses on evaluating learning and teaching techniques, models of measurement and performance indicators. In 2018, Hattie's *Visible Learning* research synthesised findings of 1,500 meta-analyses of 90,000 studies (Hattie, J., 2018).

Secondly, Rossi and Montgomery's model focuses mainly on students' societal context, suggesting two distinct scenarios. Firstly, the community environment and the quality of the home, and secondly the quality of the school, such as classroom conditions, the curriculum and incentives for students (Akama, E., 2017). Thirdly, a research group led by Dunlosky from Kent State University in 2013 summarised ten years of literature on the possible enhancement of student accomplishment in different conditions (Ericsson, A. & Pool, R., 2016). Lastly, the 'Chemers, Hu and Garcia model' is a longitudinal study carried out by Martin M. Chemers, Litzte Hu and Ben F. Garcia at the University of California. They investigated the effects of optimism and academic self-efficacy on students' achievement, commitment to continuing in school, health and stress (Chemers, M., et al., 2001).

These studies present a broad range of research highlighting numerous learning factors affecting students' achievement. Many of these factors are outside students' control. Even though they are aware of them, they may not address issues associated with them on their own. For instance, students cannot choose where they are born, and they may not be able to change other people's decisions. However, they can control learning factors such as their attitude, their sense of responsibility, their psychology, behaviour, self-responsibility skills, and most cases, their physical health. Furthermore, students have responsibility for their communication and how they want to study and learn. However, there is a gap in the literature where those

perspectives are associated and find the correlations of the students' controllable factors (Akama, E., 2017) .

This thesis proposes the PS2CLH model as a basis for finding the correlation among the controllable factors by predicting students' academic results to create the students' profile. The PS2CLH model is a student-controllable learning factor model that combines the perspectives of Psychology, Self-responsibility, Sociology, Communication, Learning and Health & wellbeing (PS2CLH). The research first required ethical approval from the university to experiment with students, as the proposed model used qualitative methods to identify underlying factors effecting academic achievement and selected controllable factors. The factors which were identified as influencing students' performance were then used to build the questions for the PS2CLH model. We then built the web-based questionnaire using Scrum methodology. The data was collected through a self-evaluation web-based questionnaire. The past performance of each student and factors affecting it were then quantified. The CRISP-DM Methodology was used to analyse the data and build the model. The final step was to predict students' performance, evaluate the importance of the attributes and determine the correlation among the factors and clusters of students. It is important to remember that making a realistic prediction of students' performance requires several elements and variables, such as lecturers' expertise, university conditions/quality, students' family and other factors. However, this research focuses on finding the correlation among students' controllable academic factors.

This study investigates the impact of students' controllable factors on student achievement. Therefore, the focus of the proposed PS2CLH model is on factors that students can control so that they are aware of how such factors influence their achievements and can then take action to address these issues independently (or with a chatbot assistant or via mentorship programmes).

4.2 Proposed PS2CLH's model

PS2CLH model is a student-controllable learning factor model for enhancing students' ability to control their performance, which combines the perspectives of Psychology, Self-responsibility, Sociology, Communication, Learning and Health & wellbeing. In this research, soft skills are referred as 'self-responsibility'. This research acknowledges other perspectives such as religion, spirituality, positive thinking and the law of attraction, among others. However, the proposed model excludes them as there has been no scientific study showing the correlation of those perspectives and students' academic performance.

The proposed PS2CLH's model is based on an abstract umbrella concept of perspectives. Knowing that each perspective covers a large spectrum of learning factors which affect students' performance, the model will be adapted to the reality of each country or university. They will select perspectives in such a way only the most influential learning factors are included, since each University face different challenges.

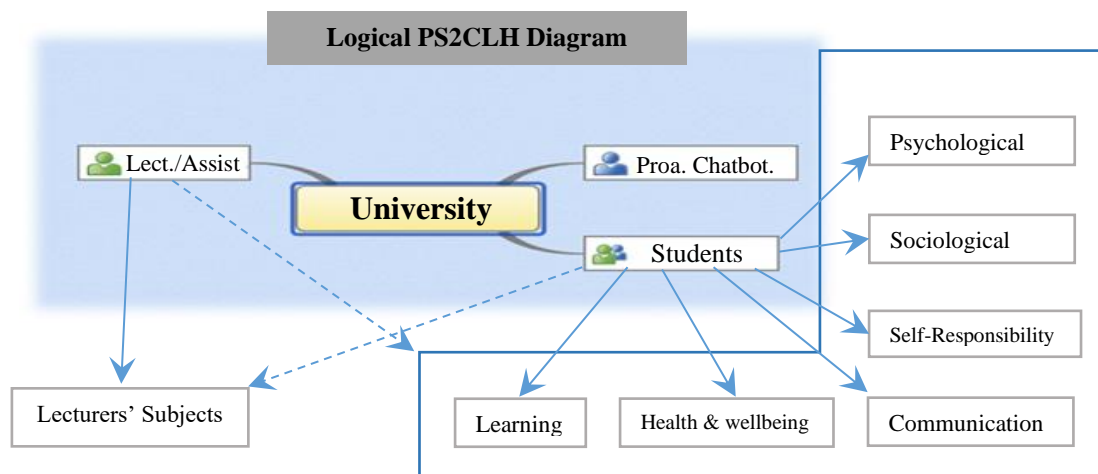


Figure 19 - Logical Diagram, PS2CLH, 2020

Figure 19 presents a diagram of the PS2CLH model, showing three stakeholders, namely the chatbot assistant, lecturers/assistant and students together with six issues which directly affect students' results.

The model was inspired by the child development and early learning field (Landry, S. H., 2014). This field develops children's critical skills through interactive play in a safe and engaging environment, covering the domains of child development including cognitive development; general learning competencies; socioemotional development, physical

development and health (Committee on the Science of Children Birth to Age 8: Deepening and Broadening the Foundation for Success, 2015). The categorisation of child development and early learning stems from a variety of sources.

There is therefore no single best categorical organisation. Indeed, it is essential to recognise that the perspectives shown in PS2CLH Fig. 17 are not easily separable. For instance, general cognitive processes such as persistence and engagement also relate to learning competencies (Committee on the Science of Children Birth to Age 8: Deepening and Broadening the Foundation for Success, 2015). Nevertheless, PS2CLH identifies the main factors affecting students' achievement which are under students' control in their daily life, and recognises that they are interactive and mutually reinforcing rather than hierarchical (Committee on the Science of Children Birth to Age 8: Deepening and Broadening the Foundation for Success, 2015). Therefore, for future students' representation, the diagram merges six perspectives into three pairs.

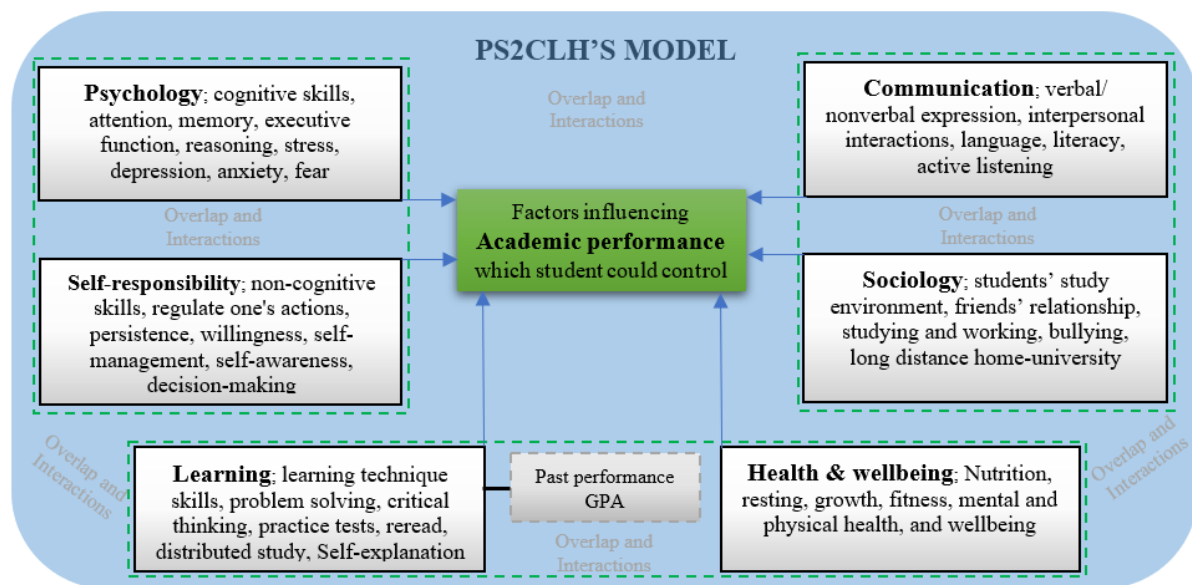


Figure 20 - PS2CLH perspectives

The "P" in PS2CLH stands for Psychology and the "S" for Self-responsibility. These two perspectives relate to mental skills, where Psychology represents hard skills or cognitive skills and Self-responsibility soft skills. In addition, both perspectives relate to students' internal state, in that one controls the psychological state and the other relates with what to do with the feelings or emotions such as willingness, passion, and persistence. They are incorporated in the model

because they are directly related to student performance, and most of the factors affecting students' performance are under students' control. It should be bear in mind that there are also psychological factors which are out of students' control such as genetic factors and IQ.

Likewise, the diagram incorporates Sociology ("S") and Communication ("C"), so the social perspective covers elements such as social interaction, relationships with family and friends' and the study environment. Communication is essential for human interaction, allowing one to express and understand others; in this diagram, communication covers understanding and linguistic elements. Excellent communication is vital in the learning process, and it seems that it could be under students' control.

The "L" in the diagram indicates learning and the "H" health & wellbeing, which to a large extent is mental and physical health and wellbeing. It is the students' responsibility to stay healthy, and this affects their learning. The diagram also covers learning strategies and study elements. The model itself is supported by previous research. Some of the scientific literature to support the proposed PS2CLH's model is presented below.

4.3 Psychology and Self-responsibility

Students' psychology has a direct impact on their performance. According to Hattie's list of factors that affect students' academic performance, a wide range of studies support the idea that psychological learning disabilities such as autism, dyslexia and ADHD negatively affect learners' academic performance. On the other hand, considerable research has proved that psychological interventions for students with learning needs positively impact their achievement. A student psychology is described as having cognitive skills, which contrast with non-cognitive or soft skills. Again, in this research, soft skills are referred as 'self-responsibility'.

The term 'self-responsibility' is associated with non-cognitive skills since it is related to "personal choice and learners' willingness", in contrast to genetic or cognitive skills. 'Self-responsibility' skills are behaviours, strategies and attitudes that underpin victories in life, such as resilience, beliefs and self-control. The concept of 'non-cognitive skills' was coined by Gintis and Bowles (Bowles, S. & Gintis, H., 1976) to call attention to the other factors not those factors measured by cognitive test scores. Bowles and Gintis emphasise the role of

perseverance, motivation and attitudes in addition to academic skills and IQ as factors contributing to accomplishment. Others offer additional support for this view, confirming the vital role of non-cognitive skills such as students' attitudes and determination in influencing social behaviour and students' results and health (Bowles, S. & Gintis, H., 1976).

To determine academic and employment outcomes, non-cognitive skills are increasingly considered to be as significant as or even more significant than IQ or cognitive skills. Indeed, growing attention is now being paid in the United Kingdom by policymakers to how non-cognitive skills could be developed in young people and children (Lesli, G. M. & Ingrid, S., 2013) The well-known scientist Michio Kaku said that when they looked at all the different theories about what makes a successful person they realised that almost all of them were wrong because it has been verified, for instance, that high IQ does not determine of a person's success (Kaku, M., 2018).

When Kaku asked which particular psychological test correlated with success in life, he found that the marshmallows test predicts people's success. The marshmallow experiment, which was created by Walter Mischel, who studied delayed gratification in young children, emphasises self-control in human growth. The experiment consists of asking a child if he/she wants a marshmallow at that moment or two marshmallows an hour from then, and the children that wanted a marshmallow immediately tended to be those who wanted shortcuts, or those who did not want to put in hard work (Mischel, W., 2015) Resisting the marshmallow and the success of self-control are the focus of studies such as "Grit", the power of passion and self-control by Angela Duckworth (Duckworth, A., 2017).

'Grit' and Self-control are two of the non-cognitive skills which have a strong correlation with outcomes. However, these skills seem more closely correlated with a steady personality than soft skills (Lesli, G. M. & Ingrid, S., 2013). On the other hand, interventions developed by Wilson in his book titled "REDIRECT" (aiming to changing the stories we live), showed long term positive outcomes for students (Wilson, T., 2013). Moreover, these academics have suggested that employment outcomes and education would benefit from investment in and development of these non-cognitive factors and help close the gap between disadvantaged and advantaged young people (Lesli, G. M. & Ingrid, S., 2013).

To summarise these psychological and self-responsibility perspectives, the psychological/cognitive skills represent students' internal state, and the self-responsibility/non-

cognitive skills are related to their decision-making. The next section presents an examination of sociology and communication perspectives.

4.4 Sociology and Communication

Communication and Social relationships play an essential role in students' performance. According to Noble (Noble, J. P., et al., 2006), there is a strong correlation between students' academic results in secondary school and their academic activities, awareness of the study strategies they use, parental guidance, family income and their parents' level of education among other factors. However, students' social relationships (friends and family) and their homes seem to be forgotten by most of the literature. For instance, the environment at school or university (Laiqa, R., et al., 2011) supports the idea that school facilities affect the education process. Rossi and Montgomery's model also reinforces this idea, focusing mainly on student's social context, which leads to two distinct factors - the quality of the school climate and curriculum, and the quality of the home and community environment (Chemers, M., et al., 2001). Laiqa, et al. (2011) also argue that the environment directly effects students' academic performance (Laiqa, R., et al., 2011).

Therefore, the architecture of the home is important. The shape, colour, texture, scale, proportion and quality of illumination link to the quality of the environment, and these factors impact human and cultural behaviour. Although most of the research focuses on schools rather than residences, they draw attention to the fact that building conditions associated with personal comfort affect students' performance. Lawson and Bacolod support this argument, emphasising the importance of the supply of essential services - for instance, claiming that "the better luminosity in learning environment improves the concentration of lecturers and students" (Laiqa, R., et al., 2011). It should be remembered that people's environment is dynamic and constructed out of family and social relations.

Family structure has a direct impact on student's achievement. Bankston and Caldas' research states that non-single-headed families (i.e. two-parent families) are six times more likely to be wealthy than single-headed families (Bankston, C. L. & Caldas, S. J., 1998). Consequently, students in single-headed family environments will have to live without a father or mother figure and with financial difficulties. Mulkey et al. (1992) also claim that family structure

influences school performance. Furthermore, Bankston and Caldas (Bankston, C. L. & Caldas, S. J., 1998) reinforce the idea that students' family structure impacts educational success, not only socioeconomic status.

Social communication skills also play an essential role in students' achievement. That means that differences between the language used at home and that used on campus can cause problems. Moreover, it raises a double awareness of the process of adapting to a different language environment. On the other hand, according to Abdullah (Abdullah, A., 2005), students who have good English and excellent communication skills do better. Furthermore, Williams & Burden (Williams, M. & Burden, R. L., 1997) discovered that the spoken language in the classroom gives students the confidence to discuss, use the new terminology to communicate and to experiment with different ways of conveying meaning, as well as dealing with failures and successes. In brief, the social perspective encompasses the student's external environment, and communication is the student's understanding of the bridge between external and internal dialogue. We will now turn to learning and health and wellbeing.

4.5 Learning and Health & wellbeing

Learning techniques and students' health and wellbeing directly affect their academic performance. An outstanding monograph, "Improving students' learning with effective learning techniques" (Kent State University, 2013), presents ten years of literature indicating that these factors can enhance student accomplishment across various environments. The study focuses on practical learning techniques in a research group led by Dunlosky (Dunlosky, J., et al., 2013). Active learning strategies led to positive results in students' academic performance. Ericsson is considered to be the research leader in the study of what makes people great in what they do. He coined the term "expertise", inspired the 10,000 hours rule and created the deliberate practice technique. For 30 years this has been a powerful method for helping children to develop expertise in practical learning. In his recent book, "PEAK", Ericsson calls attention to purposeful practice, divided into the four components of: 1. Having a clear goal, 2. Intense Focus, 3. Immediate feedback and 4. Get out of their comfort zone (Ericsson, A. & Pool, R., 2016).

Health & wellbeing is the other factor that has a significant impact on students' performance. According to Singh, there is a correlation between physical activity and scores on various

subjects. She argues that “*there are, first, physiological explanations, like more blood flow, and so more oxygen to the brain. Second, being physically active means that there are more hormones produced like endorphins. Also, endorphins make the stress level lower and improve the mood, which means better performance*” (Singh, A., 2012). Likewise, students involved in organised sports are more focused in the classroom. However, differences among the observational studies lead Singh to declare that it is impossible to establish correlations between the amount or kind of activity and the level of academic enhancement (Singh, A., 2012). In a nutshell, the learning and health & wellbeing perspective represent the student’s strategy to deal with learning techniques and action toward a personal healthy life.

To test the effectiveness of the model proposed, we conducted two experiments to test it with Angolan students.

4.6 The Experiments: Setup, Results and Discussion

The Research Questions are:

1. How can the impact of the assistant on the student’s academic life be investigated?
2. How can a structure be built to assist students at universities that combines factors perspectives such as Psychology, Self-responsibility, Sociology and more, and which incorporates the university focus on those student-controllable learning factors which most affect students’ academic performance to support students in dealing with them?
3. Is there a scalable and cost-efficient way to deal with student-controllable learning factors in students’ academic subjects to facilitate student - lecturer interaction at universities using new technologies such as Artificial Intelligence, Natural Language Processing and Deep Learning?
4. How can an application be built which incorporates an AI student learning assistant tool that could assist students in dealing with students-controllable factors?

To answer these questions, two experiments were carried out with students using the PS2CLH model. At the end of each session, we presented the results and made a disclaimer.

The first experiment was carried out at the beginning of 2018, from March to May, at the Universidade Católica de Angola. All that was done on this occasion was to collect the students' data. The researcher was not present. Nevertheless, we took the first step towards determining correlations among the PS2CLH model factors. It should be noted that the health and well-being area was not included in this first experiment, and only added in the second experiment. We started with a qualitative survey to identify the academic factors that most affect students, and then developed the web-based questionnaire. Once the student data had been collected, the quantitative research was completed using SPSS Modeler statistics and data mining software. The same software was used to develop the prediction model, the attribute importance and the correlation among the variables.

The second experiment was carried out at the end of 2018, from September to November, at the Universidade Católica de Angola. The health and wellness area were added for this experiment, and the methodology for finding the correlations among the PS2CLH model factors was implemented fully. Approval was first obtained from the London Metropolitan University and the Universidade Católica de Angola. We reviewed the results and experience of the first experiment, which made us add health and well-being to the questionnaire and reduce the number of factors in each area. We then built the web-based questionnaire and collected data from the students, and then analysed the responses using SPSS Modeler statistics and data mining software. Using the same software, we developed the prediction model and determined the importance of the attributes and the correlation among the variables.

4.6.1 First Experiment, without Intervention

As previously mentioned, in this first experiment, our aim was simply to find the correlation among the PS2CLH model factors and to start the qualitative research.

Qualitative research to develop the survey: Due to the lack of published research papers in Angola on the factors that most affect the academic performance of Angolan University students, we researched international publications describing studies which had been carried out in this area, and found four significant references.

We interviewed students over the phone, trying to find out what they thought about the academic problems they were experiencing. We also conducted interviews with specialists in

Angola in psychology, sociology, communication, pedagogy and nutrition. Finally, we collated the information and identified the variables presented in the table below.

Experiment 1: P2SCL variables in the Angolan Context

Psychology	State of stress, Depression, A generalised anxiety or fear, Restlessness, Excessive use of alcohol, Sleep problems, Feeling that there is a void in my life which influences my hours of study, I feel that I undervalue myself, I feel that I have some psychological problems that I cannot specify, Autism, Chronic fatigue syndrome, Pervasive development disorder, Sensory , Processing-disorder
Self-management	Set priorities, Establish and achieve personal goals, Time management, Aim for excellence in everything you do, Management of stress related to studying, Coping strategies, Procrastination, Immediacy, Resourcefulness, Accept responsibility, Accept change, On average how many hours studying per day
Sociology	Violence and/or gangs, Cheating in tests, Studying and working, Bullying, Sexual harassment, Family income, Sensual images, Discouragement and negativity, Long distance, Existence in my family of certain beliefs and habits, Living conditions, Lack of electricity, water and sanitation, Lack of a public transport network, Structures and services offered by the University, On average how many hours you play/distraction per day
Communication	Interpersonal communication, Speak fluently, Understand the lecturer in the classroom, Understanding and interpretation of reading, Expression, Verbal dyspraxia, Grammar and vocabulary, Stuttering or typology of disfluencies, Nonverbal expression, Problems specific to learning that influence my communication, Total communication
Learning	Last year’s grades, Dyscalculia, Attention problem, Preparation of a questionnaire, Highlighting and underlining, Use of images for text comprehension, Practicing interleaving, Practice tests, Reread, Distributed study, Self-explanation, Preparation and presentation, Set a plan to review, Prepare summaries, Dyslexia, Dysgraphia, Dysorthographia

Table 3 - Experiment 1 Variables identified in the Angolan Context

The table above presents the variables or factors affecting Angolan students’ academic performance. When students fill in the web-based questionnaire, there is a value for each answer, and the sum of these makes the coordinate; when they complete it, it automatically calculates the cluster. Thus, we limited our choices to simple techniques that students could implement without any assistance (e.g., without requiring advanced technology or complex tests (*tests that take hours to execute*) that would have to be prepared by experts).

Developing the questions for the PS2CLH model: Initially, we used the most frequently cited questions in the interaction with students and experts. The first questionnaire consisted of 16

Psychology, 13 Self-management, 15 Sociology, 9 Communication, and 17 Learning questions (see Appendix). Each question had five multiple choice answers, with a 5-point Likert scale response. Due to the author’s lack of experience, this first experiment had more questions than the second one (see “Table 3 - Experiment 1 Variables identified in the Angolan Context”. However, we had positive feedback from 15 students, confirming that the questions were straightforward. This also enabled us to develop a web-based questionnaire.

Building the web-based questionnaire using Scrum Methodology: We used the Scrum methodology to develop the web-based questionnaire. The questions are multiple-choice, with a 5-point Likert scale response; we used the radio button for student responses. The questionnaire was constructed in such a way that each question has a value which depends on the student’s answer. The results of the first questionnaire are presented below.

Figure 21 - P2SCL form wizard

Each factor or question had an explanation of what it means. The students selected the best option in accordance with their personal reality. This completed the data collection phase of our research.

The quantitative research: The data was collected in Luanda, the capital of Angola, at the ICT laboratory of the Universidade Católica de Angola, and was grouped into five categories: Psychological, Sociological, Self-Management, Communication and Learning. Respondents had to react to each statement by choosing from ‘Strongly Disagree’, ‘Tend to Disagree’, ‘Do not know’, ‘Tend to Agree’, and ‘Strongly Agree’. Initially, the population sample included 600

students from different courses and years. However, after the cleaning process, we had around 554 students aged between 20 and 30 years old. After this exercise, the web-based questionnaire was built, and we developed a query to extract the data from the MySQL Server to a Microsoft Excel document.

Start CRISP-DM Methodology to Analyse the Data and Build the Model

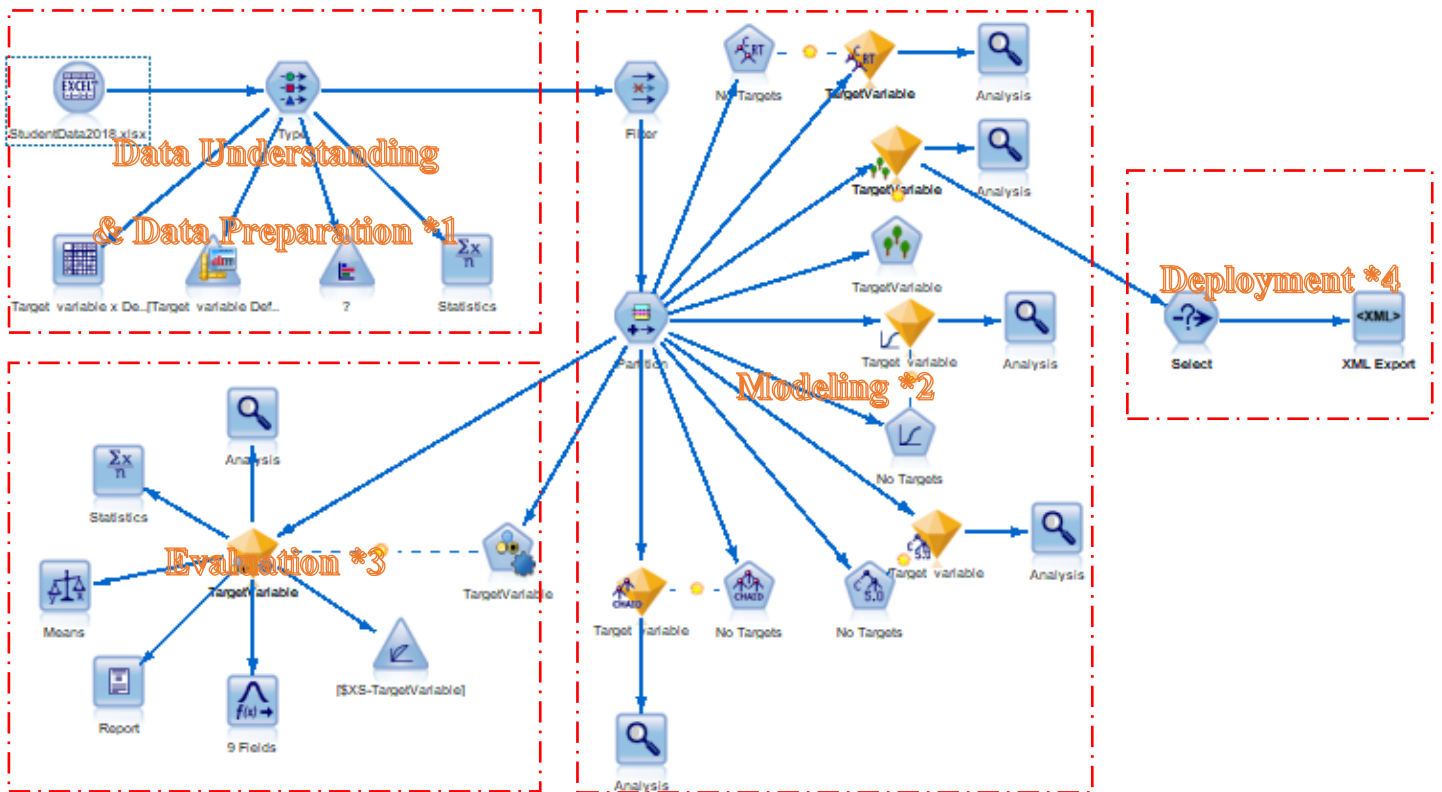


Figure 22 - SPSS Modeler CRISP-DM Prediction diagram

In 2018 the SPSS Modeler software was among the best freely available tools students' for statistics and data mining. It had more than 30 models for forecasting. The best thing was that it was free for London Metropolitan University students, and that is why it was chosen for data analysis in this research. Figure 22 above presents the phases in the construction of the predictive model.

The CRISP-DM methodology starts with understanding the business and data (universities in our case), understanding the goals and company requirements from a business perspective, turning these into a data mining application, and then developing a plan for solving the problem. A better understanding of the problem we intended to solve was achieved through qualitative research and analysing the data this produced. After extracting the data in an Excel

file, we entered the data types for each field in the file to better understand the data. The questions were in scalar form from 1 to 5. The target variable was the results of the student’s last academic year, converted into a binary data format. Data preparation involved extracting, cleaning and processing the data for use by data mining algorithms. The next step was data cleaning, and this was done by deleting incomplete records and those that had only one answer. We also deleted the extremes. Initially there were 600 records, but after cleaning we were left with 554.

As we were familiar with how the education system works in Angola, we understood and prepared (ETL – extract, transform and load) the data, then created the models and evaluated them. After finding a likely result, the model was deployed on the basis of that result. Figure 23 presents the best data partition for this specific data after testing other combinations: Training 50%, Test 25%, Validation 25%.

The screenshot shows a configuration window for data partitioning. It includes the following elements:

- Partition field:** A text box containing the word "Partition".
- Partitions:** Two radio buttons: "Train and test" (unselected) and "Train, test and validation" (selected).
- Training partition size:** A spinner box set to "50".
- Testing partition size:** A spinner box set to "25".
- Validation partition size:** A spinner box set to "25".
- Labels and Values:**
 - Training: Label "Training", Value "1_Training"
 - Testing: Label "Testing", Value "2_Testing"
 - Validation: Label "Validation", Value "3_Validation"
- Total size:** A label indicating "100%".

Figure 23 - Partition table

After data partitioning, we trained, tested and validated the data. We evaluated the partitions and applied statistics and algorithms to determine the correlations between the variables as a function of the target variable. Then the tool presents the best models results for our data. Modelling: This modelling phase select algorithm to be used for an efficient modelling process. Many algorithms require a clean data sample, causing multiple returns to the data preparation stage. Figure 24 below shows the results for overall accuracy (%) for the validation dataset, which took less than one minute to process. The empty “No. Fields Used” column indicates that the model used the most that number of variables to predict.

Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
<input checked="" type="checkbox"/>		Random Trees 1	< 1	90.741	9
<input checked="" type="checkbox"/>		XGBoost Tree 1	< 1	85.6	9
<input checked="" type="checkbox"/>		Logistic regression 1	< 1	84.8	9
<input checked="" type="checkbox"/>		LSVM 1	< 1	84.8	9
<input checked="" type="checkbox"/>		Tree-AS 1	< 1	84.8	1

Figure 24 - Models results

Discussion: The analyst evaluated many models before completing the modelling phase. The aim now was to evaluate the models against the vision of the objectives. We did that by checking that there were no gaps or inconsistencies concerning the principles of the research objectives. Figure 24 above presents the results of the five best models from 20 different types of models. The best is Random Trees 1, offering 90.74% Overall Accuracy, and using 9 fields (presented in the next section) among 66 fields (from Table 3 above); The second best is XGBoost Tree1, with 85.6% Overall Accuracy.

It seems that students spend too much time on entertainment. The responses “aim for excellence in everything you do” plus “establishing and achieving personal goals” suggest that the time that students spend on distractions such as social networks, Facebook, YouTube and Twitter are game-changing for top students. Responses to ‘1. Having a clear goal’, ‘2. Intense Focus’, ‘3. Immediate feedback’, and ‘4. Get out of your comfort zone,’ seem to be in harmony with the literature reviewed in relation to the marshmallow factor or delaying instant gratification and deliberate practice. After selecting the best model, we used it to forecast student results.

Predicting students’ performance, Attribute importance and Correlation: The current tool presented some limitations, it cannot explain some procedures, so we are first going to explain how Person Correlations and Attribute Importance are defined. A predictor variable is a variable that is used to predict some other variable or outcome. The importance of a Variable is calculated by the sum of the decrease in error when split by a variable. Relative importance is variable importance divided by the highest variable importance value so that values are bounded between 0 and 1 (Chauhan, A., 2017). The predictor importance chart helps one do this by indicating the relative importance of each predictor in estimating the model. Since the

values are relative, the sum of the values for all predictors in the display is 1.0. Predictor importance does not relate to model accuracy, but only to the importance of each predictor in making a prediction, not whether or not the prediction is accurate. Variable importance evaluation functions can be separated into two groups: those that use the model information and those that do not. The advantage of using a model-based approach is that it is more closely tied to the model performance and that it may be able to incorporate the correlation structure between the predictors into the importance calculation, regardless of how the importance is calculated (Chauhan, A., 2017)..

Pearson's correlation coefficient is the statistical test that measures the statistical relationship or association between two continuous variables. It is regarded as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association or correlation, as well as the direction of the relationship.

Properties

Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and 0 indicates that no relationship exists.

Pure number: A pure number is independent of the unit of measurement. For example, if one variable's unit of measurement is inches and quintals are used for the second variable, even then, Pearson's correlation coefficient value does not change.

Symmetric: The correlation of the coefficient between two variables is symmetric. This means that between X and Y or Y and X, the coefficient value will remain the same.

Degree of correlation:

Perfect: If the value is near ± 1 , then it said to be a perfect correlation: as one variable increases, the other variable also tends to increase (if positive) or decrease (if negative).

High degree: If the coefficient value lies between ± 0.50 and ± 1 , then it is said to be a strong correlation.

Moderate degree: If the value lies between ± 0.30 and ± 0.49 , then it is said to be a medium correlation.

Low degree: When the value lies below $+ .29$, then it is said to be a small correlation.

No correlation: When the value is zero.

(Statistics Solutions, 2020)

The Pearson correlation coefficient measures a linear relation and can be highly sensitive to outliers. In such cases one prefers the Spearman correlation, which is a robust measure of association (Profillidis, V.A. & Botzoris, G.N., 2018).

$$r_{XY} = \frac{n \cdot \sum_{i=1}^n x_i \cdot y_i - (\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{\sqrt{n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \cdot \sqrt{n \cdot \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}}$$

Equation 12 – Pearson Correlation coefficient

The Pearson correlation coefficient (also known as the “product-moment correlation coefficient”) is a measure of the linear association between two variables X and Y. It has a value between -1 and 1 where:

-1 indicates a perfectly negative linear correlation between two variables

0 indicates no linear correlation between two variables

1 indicates a perfectly positive linear correlation between two variables.

(Profillidis, V.A. & Botzoris, G.N., 2018)

One of the findings was that we could reduce the number of variables from 66 to nine to build the models, which would have a reasonably high prediction accuracy level. By improving those nine variables, we could build a blueprint to guide or mentor a student to obtain better academic results, with an accuracy of approximately 91%. This is presented at figure below. The variable ‘having the average hours students play/distractions per day’, has the most significant predictor

importance, followed by the variables “Aim for excellence in everything you do”, “Establish and achieve personal goals”, “Practice Tests”.

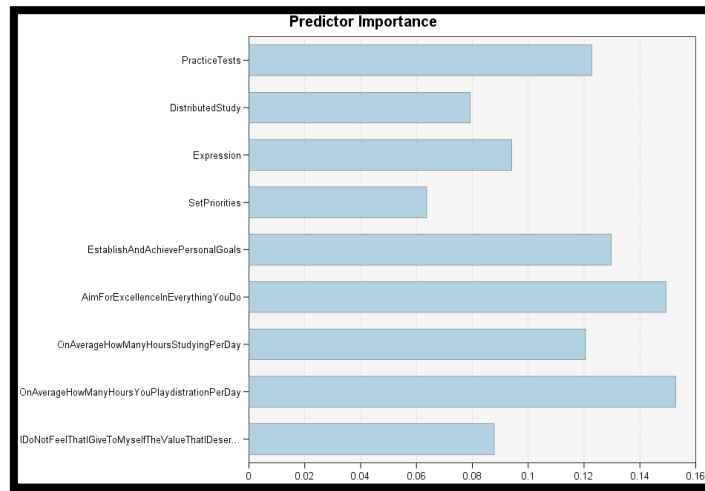


Figure 25 - Predictor Importance Random trees 1 model

The first experiment gave the following results regarding predictor importance variables. “Having the average hours’ students play”/”Distractions per day”, the most important predictors are followed by the variables "Aim for excellence in everything you do", "Establish and achieve personal goals" and "Practice Tests" among others. The results for predictor importance in the second experiment were similar to those in the first experiment. The correlation between the variables was determined in both the first and the second experiments. After completing the questionnaire, the students had coordinates based on the answers given. The two experiments revealed a pattern in the type of student. We will develop this topic further in discussing the next experiment, where we explain how we built the clusters.

Establishing the correlation among the variables allows us to add the proactivity function to the chatbot, because it will suggest the correlated factors related to the current students’ factors. Below are three variables correlated with the target variable.

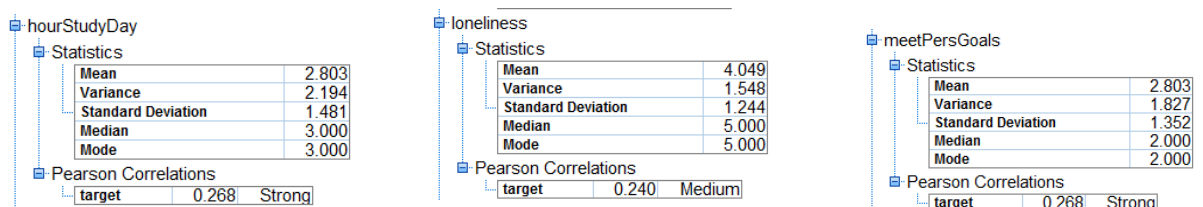


Figure 26 - Variables' Correlations

4.6.2 Second Experiment, with Intervention

In describing this second experiment, we will further discuss the phases that were not covered above, thus further developing those that were not referenced and avoiding redundancy in explaining the phases. In addition, we will explain and discuss the usability of creating clustering and 3D representation.

This second experiment was different from the first experiment. In this second experiment, the researcher will travel to Luanda, Angola, to collect data and interact with students. When he comes back to London, the interventions will continue with the selected group of 25 students from the total experimental population of 500 through WhatsApp, phone calls, local lecturers and assistants to help other students.

Phase 1: (Research Ethical approval) To be a credible and official scientific experiment, the experiment needed the initial approval of the Universidade Católica de Angola in Luanda, Angola. Angola is a developing Southern African country which gained independence in 1975 and then lived in a state of civil war until 2002. The war left many psychological and social traumas which affected the Education System directly; as a result, the level of students' academic performance is very low (United Nations Children's Fund, UNICEF, 2011). There have been no scientific studies on the relationship between factors affecting students' achievement. Therefore, the variables selected reflect common assumptions about the factors that most affect students' performance. They also reflect the Angolan reality.

Variables were chosen which do not require intervention by an expert or anybody else.

Research ethics play a vital role in our research (please find attached a document from London Metropolitan University). Confidentiality was therefore important in this study, and all student participants must be respected. Consequently, students were able to withdraw and refuse to participate in the project at any time (Research & Enterprise Development Centre, 2014). Students have the freedom to be anonymous. Angolan Universities do not have any clear research ethics approval procedures. We therefore went through the following steps:

1. we obtained written confirmation that such a framework does not exist in the country and written permission for the proposed research from the university.
2. we applied for a level of research ethics approval similar to that required for UK-based research (Ellison, G., 2013).

Once the research was approved, we started preparing the experiment by examining the University rules and deciding which data were more relevant in the Angolan context.

Qualitative research to develop the survey: The required qualitative research was carried out in the first experiment, and we used the results of that investigation as a basis for this second experiment. The experience we had acquired enabled us to improve the research. We were examining specifically the factors controllable by students, which affect their academic performance in terms of psychology, self-responsibility, sociology, communication, learning, and student health and well-being. This was the basis for developing the questions. Table 4 below shows how this is different from the first experiment (as shown in Table 3), in that it has a reduced number of variables.

Experiment 2: P2SCL Variables in the Angolan Context

Perspectives	PS2CLH variables in the Angolan Context
Psychology	Stress; depression; anxiety or fear; Disturbance of the mode of being; Belief in witchery; Low self-esteem; Unspecified psychological problems
Self-responsibility	Set priorities; Establish and achieve personal goals; Time management; Aim for excellence; Procrastination; Immediacy; Accept change; Hours of study per day
Sociology	Studying and working; Bullying; Family income; Sensual images; Discouragement and negativity; Long distance; Certain negative beliefs and habits; Bad living conditions of habitability, Lack of electricity, water or sanitation; Lack of public transport; hours of play/distraction per day
Communication	Linguistic fluency; Understanding the lecturer in the classroom; Understanding and interpretation of reading; Expressing oneself; Grammar and vocabulary; Stuttering or typology of disfluencies; Learning problems which impact my communication
Learning	Preparation of a questionnaire; Highlighting and underlining; Practice tests; Reread; Distributed study; Self-explanation; Prepare summaries; Problems with calculus and mathematics
Health & wellbeing	Regular physical activity and exercise; Feeling mentally healthy; Plenty of energy during study time; Eating healthily; Rest body and mind; Sleep problems.

Table 4 - Experiment 2 Variables used in the Angolan Context

Developing the questions for the PS2CLH model: After we had completed the first experiment, we concluded that we should reduce the number of variables. Furthermore, the variables should be factors that did not need external or expert evaluation.

Building the web-based questionnaire using Scrum Methodology: We used the Scrum methodology to develop the web-based questionnaire. The questions were multiple-choice with

5-point Likert scale responses; the radio button was used for student responses. The questionnaire was constructed in such a way that the value of each question depended on the student's answer.

Collecting Student Data – the beginning of quantitative research: Our second questionnaire had questions on 7 Psychology, 8 Self-responsibility, 12 Sociology, 7 Communication, 8 Learning, and 6 Health & Well-being, each with Likert scale multiple choice responses. The aim was to research 500 students from the Universidade Católica de Angola with the University's permission from September 2nd to November 28th between 07:00 and 22:00. 'Student' was defined as a student who was studying at the University on the day of the survey. Participants were given 20 minutes to complete the form. 432 completed records were included in the data analysis.

Using CRISPDM Methodology to Analyse the Data and Build the Model

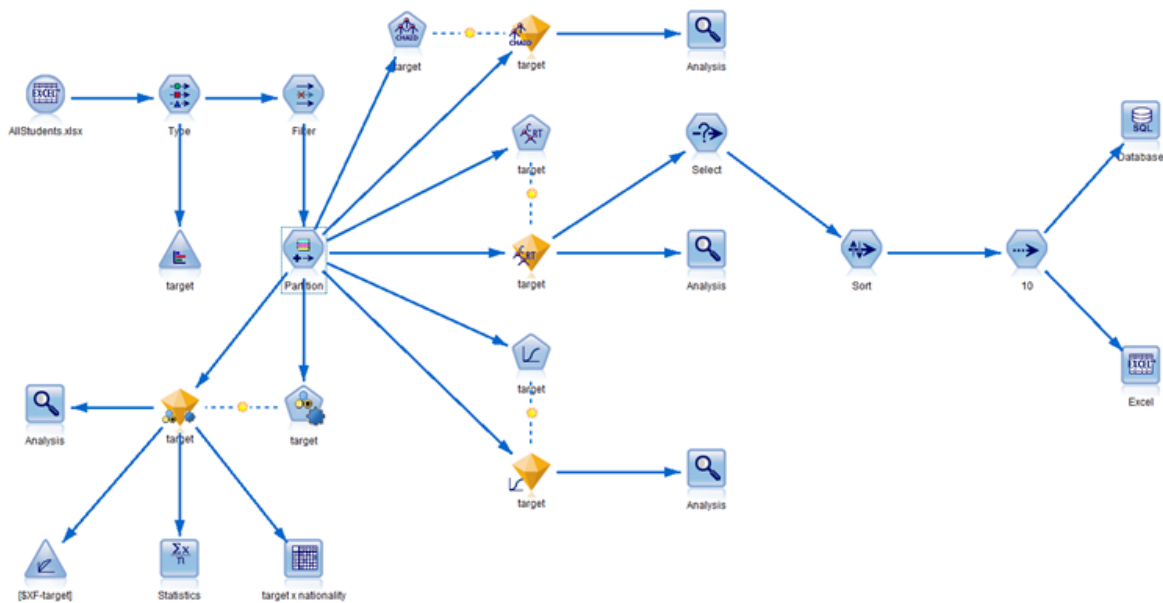


Figure 27 – Process for Predicting Students' Performance

Above, we present the process for predicting students' performance. It starts by collecting data from an Excel document, classifying the types of variables, and cleaning the data. Then the data is partitioned into training, validation and test data and different models are applied. Next, we evaluate their performance and select the best one. Finally, we save the processed report and the prediction model into a database and an Excel document.

Phase 2: (Experiment preparation, Understanding the Business and Data) Our understanding of the business and the data was based on the previous experiment. Therefore, we started our second experiment having as a reference point the variables of the first experiment. We therefore reduced the number of variables which had less impact on the prediction model and kept those with the highest score for predictive importance. We also added a new area, namely Health & wellbeing, and changed Self-Management to Self-Responsibility.

We then informed the students about the research, organising a conference with those who were interested in the research. We explained the project and told them why we needed their participation. They spread the message to their classmates. The data collection day was set, and the web questionnaire was installed in the ICT laboratory. Due to the large number of students who were interested in the research, we organised a team to direct the students to the venue for the data collection and help them if needed in filling in the questionnaire. We also developed an online application to collect the students' data using PS2CLH variables. During the data collection process, the team helped students with lab networking and technical issues for filling in the web questionnaire.

To collect data on students' controllable factors which affect their performance using a multiple-choice self-evaluation questionnaire is not the ideal scientific method. There is a dilemma in relation to the number of variables and testing whether the student is affected by that problem. Psychology has a test for each problem, which takes a considerable time to complete. Because of the number of PS2CLH model variables, such a test is not practical. Students will not stay more than two hours to fill in a questionnaire. and the diagnoses of factors for the model does not require external specialist tests. The test we used therefore shows the symptoms related to each factor, which makes the process practical and efficient. According to the positive results of the PS2CLH model, the variables selected are directly correlated with students' results, which leads us to suggest that the selected areas and variables are the right choice for our study.

The target sample population was around 540 students from different courses, aged between 20 and 25 years. Based on our extensive knowledge about Universidade Católica de Angola students, we prepared a venue for collecting the data. We selected one class to collect data from. 25 students were selected from a group of 50 students to work on some interventions with the researcher. The purpose was to guide/assist those students for three months, and then

compare the final academic results of the 25 selected students who received assistance with those of 25 students who did not get any external assistance. We created a WhatsApp group to assist those 25 students for three months.

We discussed the problems with all 25 students and the many reasons for them, which showed what the best sort of intervention should be. First, we looked for a common denominator; but it was a real challenge to find a common problem by talking to the students. Then, having prepared the ICT laboratory for our experiment in the ‘understanding business and data’ phase, we started collecting the data by asking the students to fill in the questionnaire.

Phase 3: (Filling in the Questionnaire) This phase will further help understand the impact of the academic factors that affect students’ performance and make them aware that their daily choices directly impact their academic results. In this phase, we had to identify a typical academic factor which students encountered. We found that they spent a lot of time on distractions such as social networks, games, TV series, listening to music and so on. Therefore, we tried to show them how much time they spent on distractions, and the aim was to calculate the amount of time they spent on non-productive activities.

We asked students, on average, how many hours they spent per day on distractions. They spent 6 to 9 hours per day enjoying themselves. We then asked each student individually to use a calculator and calculate how much time they spent per year on distractions. Taking the average hours they said they spent per day, which was 7.5 hours per day, and multiplying that by 7 days ($7.5 * 7$) gives 52.5 hours per week. ($52.5 * 4$ weeks = 210) hours per month. ($210 * 12$ months = 2,520) hours per year. ($2520 / 24$ hours = 105) days per year. ($105 / 30 = 3.5$) months per year. On average, then, they spent 2,520 hours, or 3.5 months per year, on non-productive activities. This result shows 24 hours awake per day of non-productive activity during 3.5 months. Here we confess that we applied some motivational tactics to get their attention. They were shocked at what this showed, and became interested in changing their addictions and improving their academic results.

To achieve our target, we had to motivate the students. We gave them an incentive – something they liked. We bought the best chocolate we could get. Even though it was expensive, we could afford it. We told them that if they brought another student, they could have more chocolate. This was the sweet phase of the experiment. The euphoria and enthusiasm were palpable. With a team of about seven volunteer students to support other students in completing the form, we

finished this phase in two weeks. After that, the researcher had to go back to London to continue other work on the research. We used a machine learning mechanism to process and model the data.

Phase 4: (Data Processing and Modelling) This research used two approaches to select algorithms for an efficient model process and to model the data: The first was IBM® SPSS®Modeler, which is a predictive analytics platform designed to bring predictive intelligence to individuals, decision, groups, systems and the enterprise. The second was the R language programming, used for processing and modelling the data.

As we mentioned before, the population sample comprised around 540 students from different courses and aged between 20 and 25 years. We processed all the data by applying machine learning models and algorithms.

```
In [1]: import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import precision_score, recall_score, accuracy_score, f1_score
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
# Instantiate model with 1000 decision trees
from imblearn.over_sampling import SMOTE

studentData = pd.read_excel('datosAlun2018_19.xls', 'Table1', index_col=None, na_value=['NA'])

studentData.info()
print(studentData)
```

stress	540	non-null	int64
feelDepressed	540	non-null	int64
anxietyFear	540	non-null	int64
disturbanceModeBeing	540	non-null	int64
loneness	540	non-null	int64
lowSelfsteam	540	non-null	int64
unspecifiedPsychProb	540	non-null	int64
setPriorities	540	non-null	int64
establishAchievePersonalGoals	540	non-null	int64
timeManagement	540	non-null	int64
aimExcellence	540	non-null	int64
procrastination	540	non-null	int64
immediacy	540	non-null	int64
acceptChange	540	non-null	int64
hoursStudyPerDay	540	non-null	int64
studyingWorking	540	non-null	int64
bullying	540	non-null	int64
familyIncome	540	non-null	int64
AdictSensualImages	540	non-null	int64
discouragementNeerativity	540	non-null	int64

Figure 28 – Importing required libraries

The figure above, shows the importing required libraries and reading the data from `datosAluno2018_19.xls` file into a data frame.

```
In [2]: studentData.set_index('id',inplace=True)
studentData.head(5)

Out[2]:
```

	targetVariable	stress	feelDepressed	anxietyFear	disturbanceModeBeing	loneness	lowSelfsteam	unspecifiedPsychProb	setPriorities	establishAchievePerso
id										
1	0	2	3	1	2	2	1	3	1	
2	0	2	1	1	1	1	1	1	1	
3	1	5	5	5	4	1	2	4	5	
4	0	1	1	2	5	5	5	5	5	
5	0	5	5	1	5	5	5	5	1	

5 rows x 48 columns

```
In [3]: studentData_2 = pd.get_dummies(studentData,drop_first=True)
studentData_2.head(5)

Out[3]:
```

	targetVariable	stress	feelDepressed	anxietyFear	disturbanceModeBeing	loneness	lowSelfsteam	unspecifiedPsychProb	setPriorities	establishAchievePerso
id										
1	0	2	3	1	2	2	1	3	1	
2	0	2	1	1	1	1	1	1	1	
3	1	5	5	5	4	1	2	4	5	
4	0	1	1	2	5	5	5	5	5	
5	0	5	5	1	5	5	5	5	1	

5 rows x 48 columns

Figure 29 – Displaying the dataset

Figure 29, display our dataset information. Here we remove the first level to get k-1 dummies out of k categorical levels by setting the drop_first parameter to True.

```
In [4]: studentData.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 540 entries, 3 to 540
Data columns (total 48 columns):
targetVariable          540 non-null int64
stress                  540 non-null int64
feelDepressed           540 non-null int64
anxietyFear             540 non-null int64
disturbanceModeBeing    540 non-null int64
loneness                540 non-null int64
lowSelfsteam            540 non-null int64
unspecifiedPsychProb    540 non-null int64
setPriorities           540 non-null int64
establishAchievePersonalGoals 540 non-null int64
timeManagement          540 non-null int64
aimExcellence           540 non-null int64
procrastination         540 non-null int64
immediacy               540 non-null int64
acceptChange            540 non-null int64
hoursStudyPerDay        540 non-null int64
studyingworking         540 non-null int64
bullying                540 non-null int64
familyIncome            540 non-null int64
AdictSensualImages      540 non-null int64
discouragementNegativity 540 non-null int64
longDistance            540 non-null int64
certainNegativeBeliefsHabit 540 non-null int64
badConditionsHabitability 540 non-null int64
lackElectricity         540 non-null int64
lackPublicTransport     540 non-null int64
universityCondictions   540 non-null int64
hoursPlayDistractionPerDay 540 non-null int64
fluencyInLanguage       540 non-null int64
understandLecturerClassroom 540 non-null int64
understandingInterpretationOfReading 540 non-null int64
expressingYourself      540 non-null int64
grammarVocabulary       540 non-null int64
stutteringTypologyDisfluencies 540 non-null int64
learningProbImpactCommunication 540 non-null int64
preparationQuestionnaire 540 non-null int64
highlightingUnderlining 540 non-null int64
practiceTests           540 non-null int64
reread                  540 non-null int64
distributedStudy        540 non-null int64
selfExplanation        540 non-null int64
prepareSummaries        540 non-null int64
sleepProblems           540 non-null int64
feelMentallyHealthy     540 non-null int64
```

Figure 30 – Dataset Information

Displaying the dataset information, the number of variables used.



Figure 31 – Target variable

The above figure shows the target variable distribution. Checking on the target variable distribution to check if the dataset is balanced; in our case, this is imbalanced data.



Figure 32 – Training, test and confusion matrix

We are applying StandardScaler to standardize the features (to deal with the imbalanced dataset). Then we split the dataset into training and testing. We used the SMOTE function during the data training to further improve the model in an imbalanced dataset. Finally, we plot the Confusion Matrix that shows promising results.

```

In [9]: clf_cw = LogisticRegression(random_state=0, class_weight='imbalanced').fit(X_train, y_train)
y_pred = clf_cw.predict(X_test)

from sklearn.metrics import roc_auc_score

print("ROC_AUC_score",roc_auc_score(y_test, y_pred))
print("Accuracy",accuracy_score(y_test, y_pred))
print("Precision", precision_score(y_test, y_pred))
print("Recall",recall_score(y_test, y_pred))
print("F1 score",f1_score(y_test, y_pred))

ROC_AUC_score 0.8845108695652174
Accuracy 0.9351851851851852
Precision 0.7647058823529411
Recall 0.8125
F1 score 0.787878787878788

```

Figure 33 – ROC AUC score, Accuracy, Precision, Recall and F1 score

We were running logistics regression on the imbalanced training dataset. Then we used performance metrics to evaluate the model’s efficiency. First, we used the ROC AUC score, meaning the area under the Receiver Operating Characteristic curve or AUROC (or AUC/ROC). It is commonly used to evaluate classification models.), precision, recall and F1 score (vary from 0 to 1, where 1 is the perfect model) are all used to measure the accuracy of a model.

We can see that we have a very high ROC AUC score of 0.88, an excellent accuracy of 0.935, and a precision = 0.76. The recall and F1-score are also excellent, close to 0.79. This result demonstrates that we have a good model, a correlation between the factors or variables, and a clear pattern that shows the difference between the best students and the regular ones.

Phase 5: (Model Evaluation) As in the first experiment, we evaluated a range of models, and “Random Forest” demonstrated the best results. The purpose was to evaluate the models in relation to the vision of the research objective, checking that there were no gaps or inconsistencies concerning the business principles. The approach to processing the data is presented below.

Taking the other approach using Python language, instead of using the SPSS Modeler, we have a population data of 540 records with 83 positive values “1” and 457 negative values “0”. We divided the data into training and test data. We applied the SMOTE function to balance the data. Then we applied the “Logistics Regression” classification function to build the model.

Having built the model, we tested it. The accuracy was $\approx 0.94\%$.

Below is the prediction importance or the weights of the variables predicting students’ performance. This selection is an essential step in our design, because it will suggest which

variables students should pay attention to. It will be a crucial feature of the chatbot when it suggests or recommends specific controllable factors.

Figure 34 below shows the impact of the variables on the target variable. We have the top students and regular students, and the average between the two types of students. We define top students as students with results last year equal to or greater than $\geq 14/20$ values (“1”), and regular students as those with results less than $\leq 14/20$ values (“0”).

Modelling using SPSS Modeler: The Selection(s) algorithm(s) to be used should offer an efficient modelling process. Many algorithms require a clean data sample, which ends up causing multiple returns to the data preparation stage. Below is the architecture used to process data into IBM SPSS Modeler.

The balanced partitions split between results and percentages for the available data were: 60% training, 20% testing and 20% validation. The test model uses the past performance grade point average (GPA) as the target variable, correlated with all other variables selected on the implementation of the PS2CLH model in Angola.

'Partition'	Testing		Training	
Correct	64	94.12%	121	88.32%
Wrong	4	5.88%	16	11.68%
Total	68		137	

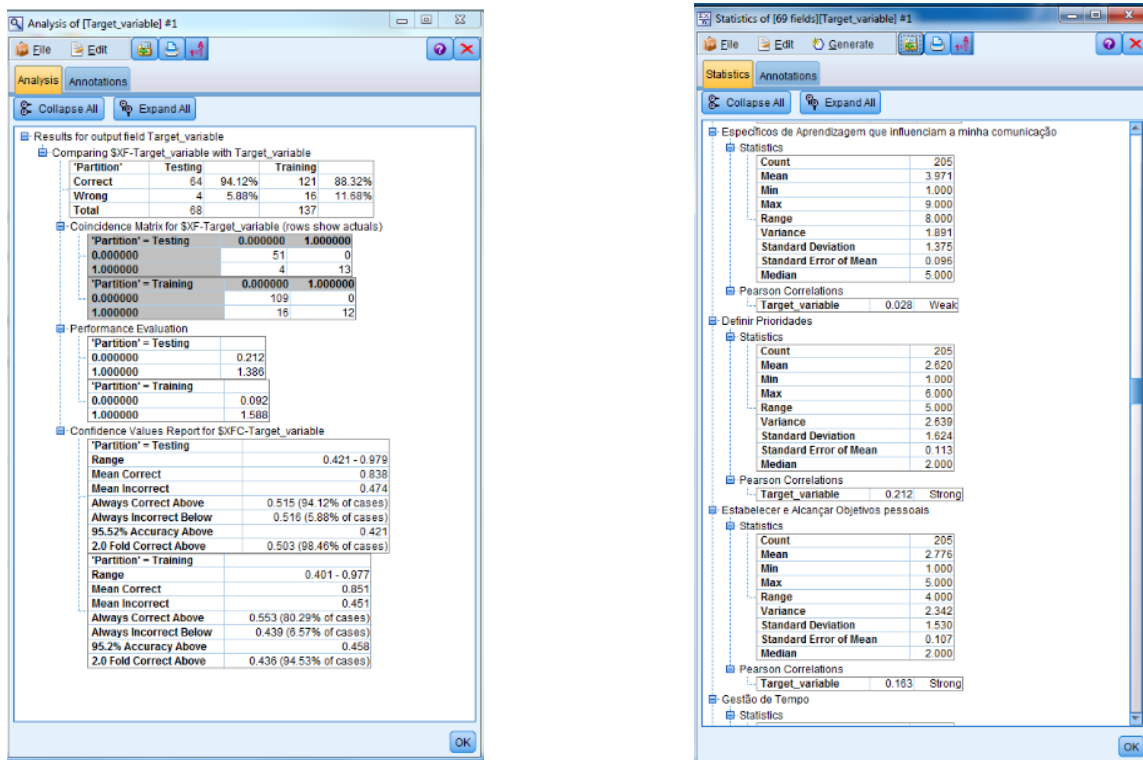


Figure 34 - Models best result

We can see the Results for the Target_variable in the first windows, which presents the Testing and Training partitions. With this model, the “Correct Testing” was 94.12%. The other windows show the correlation among the variables.

Predicting students’ performance, Attribute importance, Correlation and Clusters: As a detailed presentation was given of the first experiment using IBM® SPSS® Modeler and the results, we will show the final result, remembering that our focus is on creating an automated assistant process for students. Consequently, we have three models: the CHAID model, the Logistic Regression model and the CRT model. The model with the best results was the CRT model (94.12% accuracy).

Phase 6: (Interventions) The intervention context and actors were: the Universidade Católica de Angola in Luanda, the capital of Angola, the intervention duration was three months, as we mentioned before, and the population sample was around 540 students from different courses aged between 20 and 25 years old. To ensure a controllable intervention, this researcher chose a class of 50 students and selected 25 students to work on the interventions. Due to the distance, the assistance interactions between researcher and students, including guidance and follow-up, were carried out online. As a result, we recruited 25 committed students, most of whom participated in the interventions until the end.

The University contacted the researcher and asked him to expand the intervention to the other approximately 500 students who had filled in the questionnaire. Even if at that stage of the research the researcher did not have a clear idea on how to build a more significant intervention, the solution was to work with more assistants. In order to reduce costs, we selected the best of the original volunteer students and asked them to volunteer again follow the researcher’s instructions. Each group of assistants was given a specialist/lecturer in the area. We built four groups (Psychology and Self-responsibility), (Sociology and Communication), Learning and Health & Wellbeing. We also built a web profile for students’ interventions, and it showed their factors ordered in terms of importance according to our findings. They share those students’ problems with their assistants. The focus was on the five most essential predictor factors. Nonetheless, the research focus was on 25 students.

Looking at Angola, at the time, we did not find any significant research work dealing with these academic problems, nor any intervention which had relevant results. Firstly, we

acknowledge that we were working with students who were distracted by entertainment, given the amount of time they spent on this non-productive activity, making it a habitual problem. One psychological solution for a bad habit is replacing it with a healthier and more productive habit. In our Literature review, we discussed Principle 4 - The golden rule of habits. Most researchers believe that habits come from one golden rule. We can never change the things that act as a trigger for us, we cannot control how we might feel, but we can always choose how we react or behave (how we behave is a choice); the best way to stop a habit is to replace it with a new behaviour (Duhigg, C. & Ruben, G, 2014).

Therefore, to start our interventions, we needed to redirect students to more productive habits. To accomplish that we had to analyse students individually. We looked at those 25 students' problems. They had different daily problems and home conditions. Rather than looking at the factors with the highest scores, we tried individually to go through their Psychology, Social, Self-responsibility, Communication, Learning and Health & Wellbeing factors, which affect students' performance as stated in their PS2CLH answers. In relation to referencing Mihaly Csikszentmihalyi's graphic states that low skills and high challenges lead to worry and anxiety, whereas low challenges and high skills will generate boredom and relaxation.

We used the state of flow for our intervention when a person's very high skills meet a significant challenge. This is the happiest moment of their lives, doing what they love to do, during their most complex challenges. However, they need to take the first step to face their challenge. Dr Mihaly studied human behaviour and human performance. He used "activation energy", a term from chemistry; for the initial chemical reaction which uses a tremendous amount of energy. That is why it is hard for people to start doing what they should do (start studying hard, go to the gym and work as hard as they can, go to work and do their best) - that feeling of how hard it is to get started is what Mihaly was talking about. The first step is the key to creating any change and reaching our full potential (Csikszentmihalyi, M., 2008).

Nevertheless, the research focuses on the most influential factors that impact students' performance. Therefore, "Practice tests" based on practical learning techniques were one of the main focuses during the intervention. The interventions were most influenced by the works of Angela Duckworth and Dr Ericsson. Looking at the time students spent on distractions, it was clear that they had a problem with delayed gratification, which made us gravitate towards Duckworth's work on self-control in human growth, and particularly studies such as "Grit" the

power of passion and self-control (Duckworth, A., 2017). In addition, Ericsson’s active learning strategies presented positive results in students’ academic performance. However, as we can see below deliberate practice is hard.

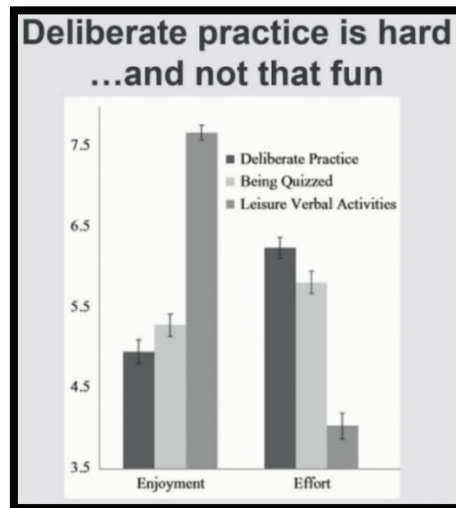


Figure 35 - Deliberate practice: enjoyment vs effort

In his book, “PEAK” (Ericsson, A. & Pool, R., 2016) Ericsson calls attention to purposeful practice, divided it into four components:

1. Having a clear goal, 2. Intense Focus, 3. Immediate feedback and 4. Get out of your comfort zone

The intervention developed on this basis, with each student having to build specific, concrete and clear goals. They had six subjects; and for each subject they wrote the result they were aiming for, and had to see it every day over the three months of the experiment when they woke up each morning. Studying was the highest priority on the agenda, so they also had the intense focus of study and practice, with immediate feedback, and the challenge increased as soon as they found themselves comfortable studying.

In addition, we evaluated the models to get the best results and diagnosed the students’ controllable factors which most affected their performance. With those results it was easy to see the variables that had the greatest impact on students. Of the PS2CLH variables, “Set priorities”, “Practice tests” and “Establish and achieve personal goals” were among the most impactful, and when we confirmed that, we knew we had discovered the common denominator among the students’ problems. Therefore, the interventions were built around these three factors: “Set priorities”, “Practice tests” and “Establish and achieve personal goals”.

The practical part of the intervention was to create a student activity plan on which they had to tick to record whether each activity had been completed or not every day, just as they would if they were setting priorities and trying to establish and achieve their daily personal goals. An example is given below.

Student Activity Plan

Activity Done? - yes ✓ no ✗		Morning									
Year: 2018		Activity Plan									
Weeks: 26-August à 1-September		Student name									
Hours	Monday - 26	Done?	Tuesday - 27	Done?	Wednesday - 28	Done?	Thursday - 29	Done?	Friday - 30	Done?	
5h-6h	Exercises and Breakfast	✓	Exercises and Breakfast	✓	Exercises and Breakfast	✓	Exercises and Breakfast	✓	Exercises and Breakfast	✓	
7:30 - 12:30	Class	✓	Class	✓	Class	✓	Class	✓	Class	✓	
13h - 14h	Lance	✓	Almoçar	✓	Almoçar	✓	Almoçar	✓	Almoçar	✓	
14:30-15:30	Studying Subject 1	✓	Studying Subject 2	X	Studying Subject 3	✓	Studying Subject 4	X	Studying Subject 5	✓	
10 min	rest	✓	rest	✓	rest	✓	rest	X	rest	X	
15:40-16:40	Studying Subject 1	X	Studying Subject 2	✓	Studying Subject 3	✓	Studying Subject 4	✓	Studying Subject 6	✓	
10 min	rest	✓	rest	✓	rest	✓	rest	✓	rest	✓	
17:50-18:50	Studying Subject 2	✓	Studying Subject 3	X	Studying Subject 4	X	Studying Subject 5	✓	Studying Subject 6	✓	
19h - 20h	Student' choices	✓	Student' choices	✓	Student' choices	X	Student' choices	X	Student' choices	✓	
20h à 22h	Dinner and distractions	✓	Dinner and distractions	X	Dinner and distractions	✓	Dinner and distractions	✓	Dinner and distractions	✓	
22h à 5h	Analyse the day 30min and Sleep	X	Analyse the day 30min and Sleep	✓	Analyse the day 30min and Sleep	X	Analyse the day 30min and Sleep	✓	Analyse the day 30min and Sleep	✓	

Figure 36 - Activity plan prototype (English)

In the two weeks, they had ticked around 40% of their planned activities. Then we evaluate why they were not doing what they had said they could do. We finally realised that they had set unrealistic targets, and we had to direct them back to their primary goal. This was a difficult phase for them, but they did not give up, and after two months, 17 of 25 students had started to tick between 60% and 80% of their activities.

They were embracing the difficulty of self-discovery in the words of Nietzsche: “*No price is too high to pay for the privilege of owning yourself.*” (Nietzsche, F. W., 2014)

This intervention focuses on 25 students whose results are presented in the next section, “Results”.

Phase 7: (Results) Among the numerous machine learning models tested, the “RandomForest model” produced the best result. The prediction presented by the model, based on the PS2CLH’s perspectives, was 94% accurate. It showed that the selected variables from PS2CLH’s proposed model directly correlate with the target variable or the student’s academic performance. Below we present the students’ visual 3D representation.

Students’ visual 3D Representation: In recent years, there has been a growth in the number of studies measuring and representing students’ learning and their performances. However, there

is a lack of research on representing, measuring and monitoring the controllable factors which affect students' performance. From an assistant's or mentor's point of view, it is essential to measure, visually represent and keep track of the student's performance alongside factors that affect their academic achievement.

It would be natural at this point to choose a particular type of visualisation for student clusters. Why represent students visually? And why 3D, not 2D, 4D or 5D?

Answering the first question, we can argue that our goal is to help students throughout their studies, and for that, we need to know how much impact the student assistance has had. Therefore, we measure the student's development at the level of clusters. It is necessary to have an initial reference point indicating the student's starting point and thus create a history of their trajectory in their academic life.

The second question has to do with the most efficient way to represent the students in the cluster, so we think the 3D representation is ideal for our data. If we used 2D, we would lose information given the number of areas we have. That is, we would have to group 3 areas to have one coordinate x or y. In addition, considering the number of students at the university, there would be an overlap in the 2D representation, which is ineffective. 3D and 4D representations would be too complex to allow the data to be read. Therefore, we conclude that 3D representations are ideal for the data and the areas we have.

This research proposes a visual representation and measure of a student-controllable learning factor that affects their performance, based on the academic model that combines psychology, Self-responsibility, Sociology, Communication, Learning and Health & wellbeing (PS2CLH). We associate psychology & self-responsibility (coordinate/axes X), social class and communication (axes Y), learning and health and wellbeing (axes Z). This results in student representation on a point in three-dimensional space 3D. Consequently, it will be possible to represent students in different clusters, effectively monitor the issues they have and understand the patterns of the PS2CLH model, leading to better academic performance and understanding of the students' behaviours in different clusters.

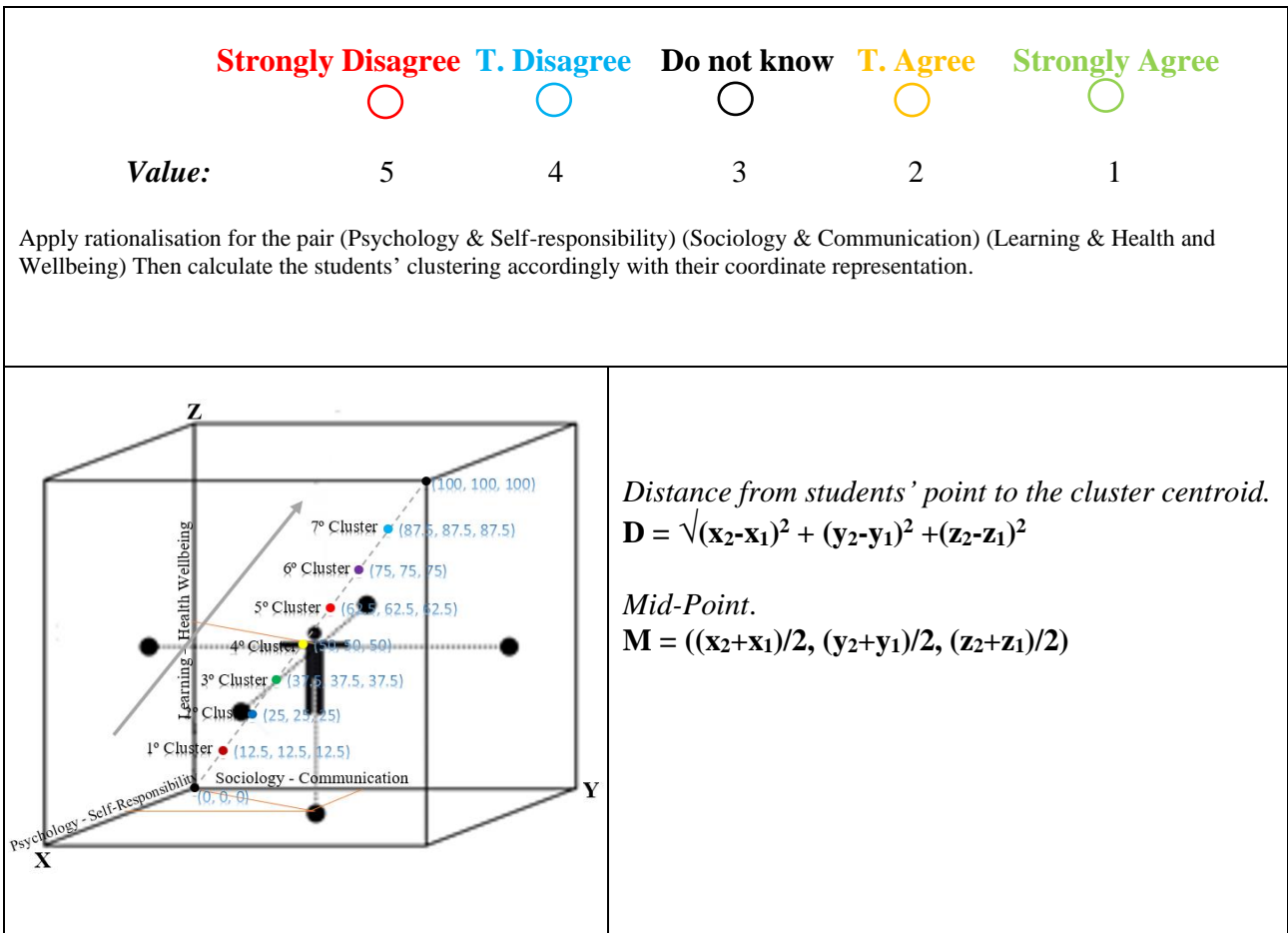


Figure 37 - PS2CLH Visual 3D representation

When students fill in the questionnaire, there is a value for each answer. The sum of these values makes the coordinate, and when they finish filling in, it automatically calculates their clustering. In the questionnaire, each question has a weight. According to the answer, this weight is attached to the question in such a way that each area weight will be the sum weight of questions in the six areas: Psychology, Self-Responsibility, Sociology, Communication, Learning and Health-wellbeing. Each pair of two areas represent a coordinate: PS coordinate X, SC coordinate Y and LH coordinate Z.

To monitor students' evolution or growth, we need to represent and observe the initial state, then monitor their evolution through clusters. Representing in 3D the factors that affect students' performance allows the system to know the distance between students, leading us to build clusters of students—clustering students into groups according to the student controllable learner model and the data from the students' questionnaires.

Given a clear representation of different clusters, the chatbot can adopt different behaviours in relation to a particular cluster of students. This makes it possible to pay attention to the students who need the most attention and have a clear notion of where each student stands and the direction and necessary steps each student needs to take to get to the desired cluster. Results show that the best students are located in clusters six and seven, i.e. clusters with fewer problems or factors that affect their performances. Therefore, the goal is to work with the individual student during the academic year to tackle their problems so that they move towards clusters six and seven.

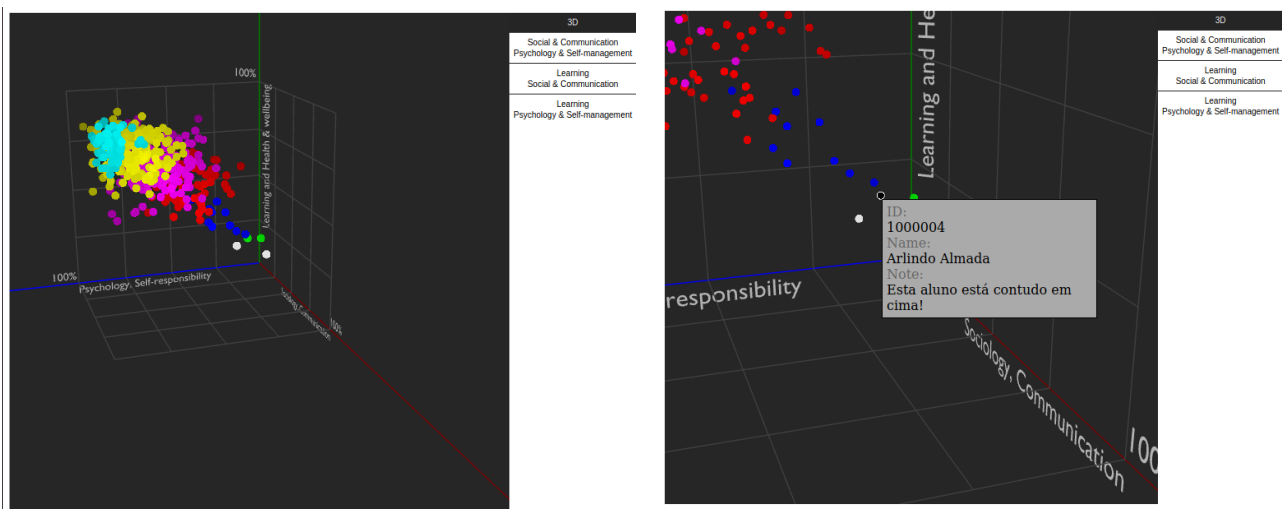


Figure 38 - Students represented in the PS2CLH Visual 3D representation

The clusters are represented by different colours; each point represents one student.

Visually representing students using the PS2CLH model in a 3D visualisation based on these students' controllable factors has the following implications: the representation of students will make it possible to visualise the students' clusters, thus showing those who will need more help. This view will also show the patterns and the cluster of the best students, which can guide university decision-makers to act proactively. In addition, the visual image of the factors that affect students' performance allows lecturers to have a visual image of their students' academic controllable factors which affect their performance.

4.7 Summary

Numerous factors that influence students' academic performance involve issues beyond the individuals' control, such as national policies, government initiatives and university resources, among many others. Even if students are aware of these factors, addressing them may not be feasible. Identifying causes within students' control could improve students' understanding of these factors and enable students to deal with related issues independently. This thesis proposes a student-controllable learning factor model that combines the perspectives of Psychology, Self-responsibility, Sociology, Communication, Learning and Health & Wellbeing (PS2CLH). The proposed model used qualitative methods to identify underlying factors affecting academic achievement and selected controllable factors. This research reports on the outcomes of the employment of the PS2CLH model to predict student performance. Initially, data is collected through a self-evaluation web-based questionnaire. Each student's past performance and factors affecting this are then quantified. This study reveals the impact of students' controllable factors on student achievement.

In the first experiment, we focused primarily on the CRISP-DM methodology. The best model was the Random Trees 1, with 90.74% Overall Accuracy, using nine of 66 variables. The second experiment focused on areas which we did not explore in the first experiment. The test results indicated that the proposed PS2CLH model offered 94% accuracy in predicting student performance.

The two experiments presented similar results in terms of the crucial variables. "Establishing and achieving personal goals" and "practice test" were higher than "stress", "learning room", and "grammar and vocabulary" among other factors. This research raised participant students' awareness of PS2CLH perspectives, which helped them manage academic performance factors more effectively. In the second experiment, most of the 25 students enhanced their academic performance by addressing these critical factors. However, due to the limitations of the current sample data, the PS2CLH model will be further monitored for various applications.

Chapter 5 The proposed Proactive Chatbot Framework Designed to Assist Students based on the PS2CLH model, Test and Results

5.1 Introduction

Nowadays, universities are using more technologies to interact with students. In fact, education has been slowly incorporating new ways to convey the message to teach students into education (Becker, W. E. & Watts, M. C., 1996). This pattern started decades ago with the spread of electronic e-mails and the web (Goffe, W. L. & Sosin, K., 2005). At the same time, distance learning required more innovative and efficient technologies to deal with natural learning challenges which presented the need to devise more effective tools to establish interaction between computers, lecturers and students. Researchers had long tried to develop such tools, but were not very successful. However, in the last decade, 2010 to 2020, there was a significant evolution in computer hardware, which made possible AI machine learning and advances in Deep Learning to make tools such as chatbots more usable (Müller, V. C. & Bostrom, N., 2016) (Holmes, W., et al., 2019). There were also advances in applications in other sectors such as Health, Customer Service, Security and Leisure. Practical applications have been pointed out in the generic field of education (Seldon, A., 2018).

The last few years have presented further new challenges to lecturers and students at universities. For instance, the pandemic forced students to study from home, meaning that they faced new problems in their learning, which affected their results. This caused a great need for new tools, but there are still issues to tackle and improvements to be made. The traditional support system is failing to tackle this new set of learning environment rollbacks.

This chapter proposes a new framework for chatbot assistants, which is integrated with students' learning profiles and is able to improve student interaction and helps to address the issues presented above. This chapter presents the framework for the new chatbot assistant, including the test and results. It starts with an outline of the fundamental connection between the PS2CLH model and the proactive chatbot framework.

5.2 The Fundamental connection between the PS2CLH's model and the proactive chatbot framework

The purpose of the PS2CLH model is to help students become aware of how controllable factors influence their achievements so that they can address relevant issues independently. The fact is that there are numerous controllable learning factors which affect students' achievement, such as their attitude, psychology, behaviour and self-responsibility. In most cases, these include their physical health. Students are also responsible for the way they communicate and how they want to study and learn and more.

To go back to the PS2CLH model, the coefficient of PS2CLH factors was calculated. Therefore, it is appropriate to apply the PS2CLH model to build the learning profile of new students, as this could be used as a knowledge base to improve the effectiveness of a chatbot.

When they interact with the chatbot, the profile could provide additional suggestions to assist students with subject answers. This will also make the chatbot more proactive, enabling individual students to have a clear view of the factors that most affect them.

5.3 The proposed Framework design for the Proactive Chatbot for Students based on the PS2CLH model

This research aims to build a framework to help students deal with their controllable academic factors in terms of the PS2CLH model which affect their performance, and to guide them to the next step for success. This thesis sets out the proactive chatbot framework. A computer programme framework could be defined as an abstract platform which serves as a foundation for creating programme applications.

The proactive chatbot framework is divided into two parts. The first part of the framework is the wide-ranging extended chatbot. It was inspired by the extending language representation BERT (Bidirectional Encoder Representations from Transformers) model based on machine learning techniques for natural language processing. This first part deals with the initial interaction with students' input/questions, while the second part is the Educational Chatbot Ecosystem, which is an educational ecosystem that supports and enhances the AI ability of the chatbot, facilitating and improving the student-lecturer/assistant interaction.

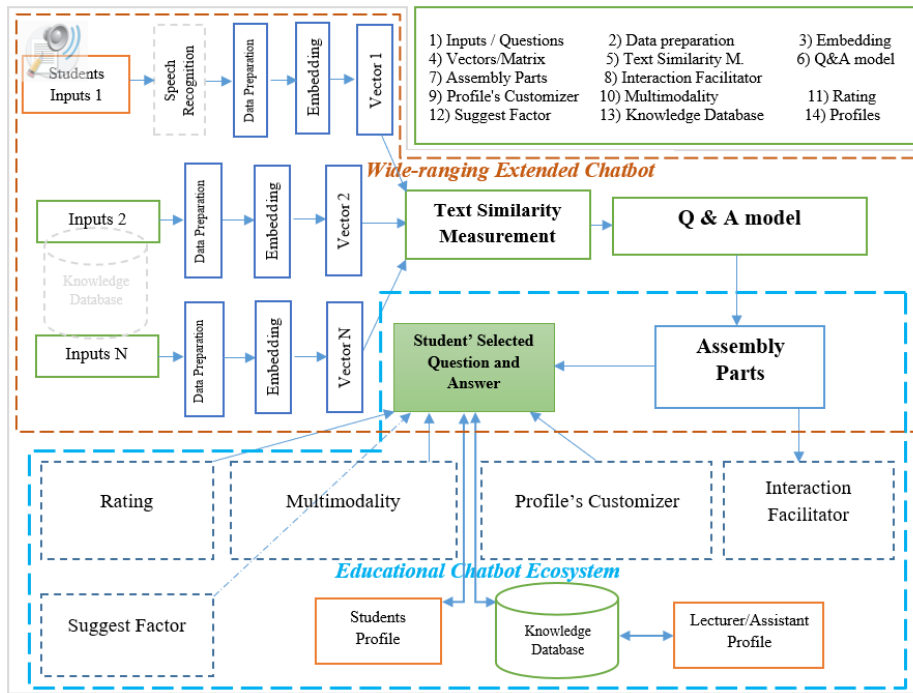


Figure 39 - Framework for the Proposed Chatbot

In summary, the framework has a layered structure composed of Inputs; the chatbot starts by receiving the students' questions and then transforms them into a matrix of vectors. These vectors first prepare and transform the data, and a text similarity measure is developed, together with a Q & A attention model, a knowledge database and components supporting student - lecturer/assistant interactions.

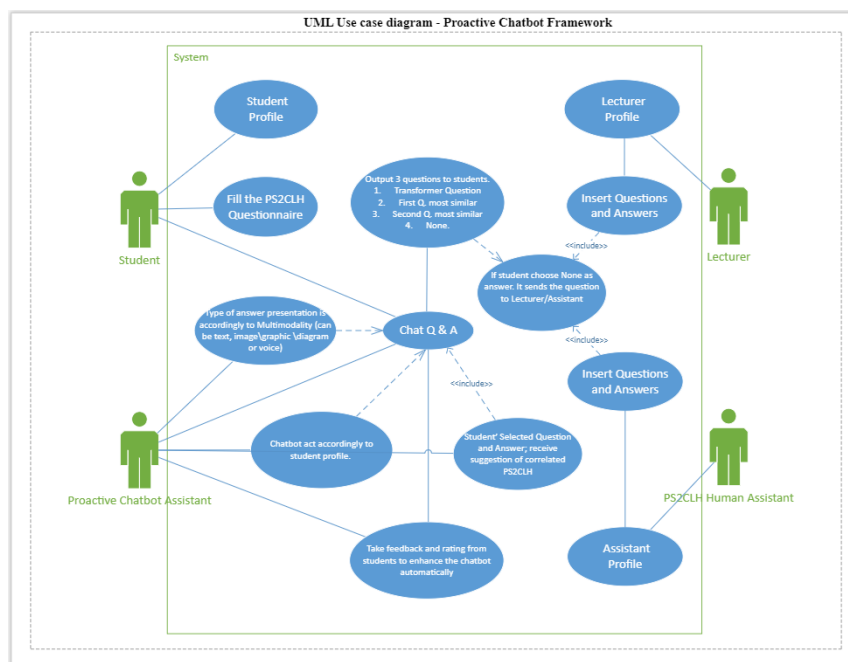


Figure 40 - Use Case Diagram Proactive Chatbot Framework

The Use Case Diagram shows the interactions between the proactive chatbot components with students and the lecturer and human assistant interactions. Students first create their profile and fill in the PS2CLH questionnaire, and the lecturers and human assistant also create their profile. Then, on the basis of the students' profiles, questions are formulated for the proactive chatbot. To help the chatbot succeed, we present four possible question options for students to choose from. At the same time, four chatbot components interact in this conversation with the student. First, the chatbot acts according to the students' profiles, and second, the content of the answer is presented to the students as multimodality content, and then the chatbot takes feedback ratings from the students. Then, if it was a PS2CLH question, the proactive chatbot suggests correlated controllable factors to students. Finally, if none of the questions satisfy the students, the chatbot sends the question to the Lecturer or Assistant, who can add questions and answers from their profile, and answer the unanswered questions from the students - chatbot chat.

As we mentioned, the framework is divided into two parts, and the first part is presented below.

5.4 Wide-ranging Extended Chatbot

This first part covers the initial interaction with students' input/questions. The chatbot is referred to as the 'wide-ranging extended chatbot'. It was inspired by Extending BERT, and is presented below.

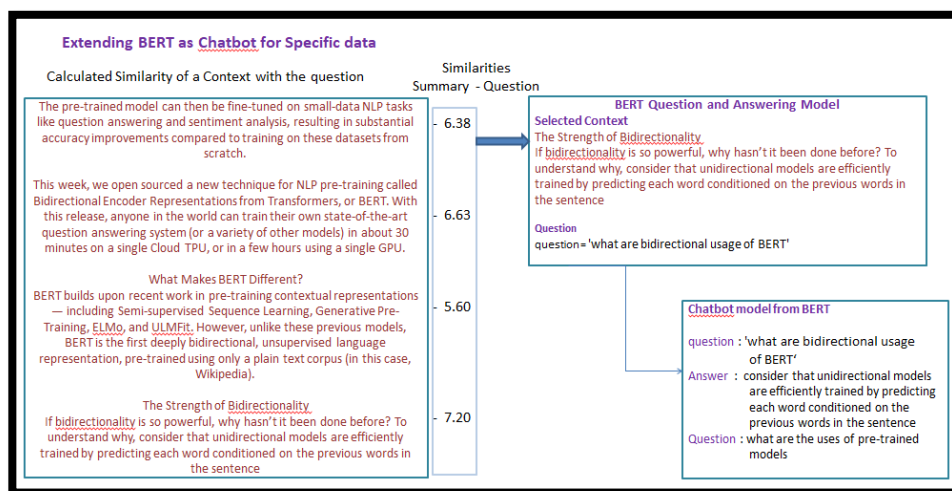


Figure 41 - Extending BERT

The typical BERT-based Question and Answer system works well for questions in the form of a short summary. However, it does not understand more than ten paragraphs or 512 tokens well. This limitation prevents it from obtaining good results with a large corpus. It is trained

on the Stanford Question Answering Dataset SQuAD (*a reading comprehension dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles*) (SQuAD, 2020), which has fewer summary paragraphs and questions related. The first part of the framework, the wide-ranging extended chatbot, is presented in the figure below.

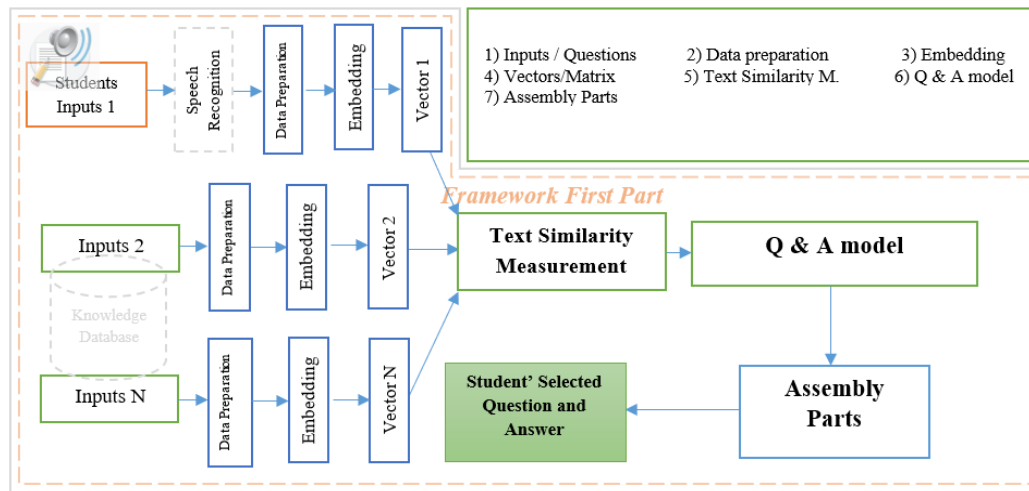


Figure 42 - Framework (first part)

Each component of the Wide-ranging Extended Chatbot is described below, remembering that it is a framework, which means that it can be adapted for use in any country or university. Moreover, components can be added to or removed from the framework. Below, we introduce the components of the first part, starting with Inputs.

5.4.1 Inputs/Questions, Data Preparation

The Inputs component is the connection point between the framework and students. The Inputs start by receiving students' questions via writing or voice. We decided to receive writing and voice because they are simple, and are the most common forms of communication. Writing inputs are the norm from keyboards. Speech recognition captures voice inputs, for which we used the google speech API, which has all the necessary functions to convert text to voice and voice to text.

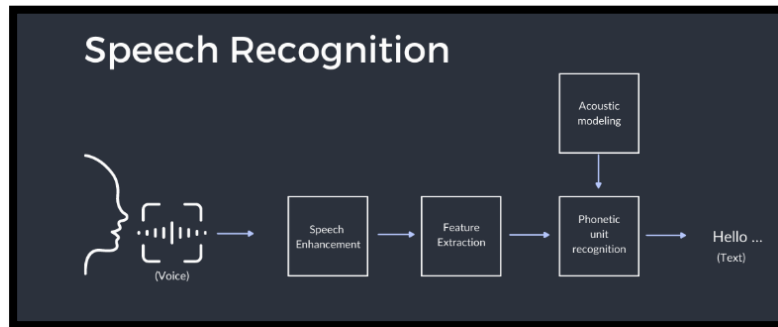


Figure 43 - Speech Recognition

Briefly, after the student’s voice is captured via a microphone, it goes to speech enhancement to improve the sound quality and then extract the feature. Next, acoustic modelling is applied to phonetic unit recognition, which converts the oral input into written text (Google Cloud, 2021). This speech API was our choice because it is efficient and straightforward, and one of the best free available speech recognition tools. Then, having dealt with the questions, we go to data preparation.

Data preparation: This component is essential for cleaning the data. The framework implements syntax and semantics. Syntax and semantics are the broad categories under which NLP techniques fall. It starts by tokenising the students’ questions. Tokenisation can be split into two categories - sentence tokenisation and word tokenisation. In contrast, sentence tokenisation separates a paragraph into distinct sentences, while word tokenisation separates a sentence into different words. Tokenisation enables the computer to learn the potential meanings and purpose of each word. Next comes stemming, which is the process of reducing a word to its root or stem. This is done by chopping universal prefixes and suffixes such as *-es*, *-s*, *-ing*, and *-ed*. Stemming is a powerful technique, but is simply a crude chopping operation based solely on common prefixes and suffixes, so it sometimes cuts necessary components off a root and changes the word’s original meaning (Roman, M., et al., 2021). Lemmatisation reduces a word to its root form by analysing the word morphologically. Let us look at what this means. If we have the words *am*, *are*, and *is*, the root form for these words is *be*, which can be observed through lemmatisation. However, stemming would not have been able to figure this out, as chopping any letters from these words would not have outputted *be*. After this cleaning process the next step is based on word embeddings.

5.4.2 Embedding, Vector/Matrix

Embedding was one of the essential breakthroughs which led to the current very impressive performance of deep learning methods. Generally speaking, word embeddings are vector representations of a specific word. Having said that, how do we produce them? More importantly, how do they capture context? There are different ways of representing a word. Word embedding is one of the foremost representations of document vocabulary. It can capture the context of a word in a text, identify semantic and syntactic similarities, connections with other words, and more. Word embeddings make it possible to give equal representation to words with similar meanings (Almeida, F. & Xexéo, G., 2019). In our framework we used **TF-IDF** (Term Frequency-Inverse Document Frequency) for this. *TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.*

As mentioned above, the different types of word embeddings can be generally classified as Frequency-based Embedding or Prediction-based Embedding (Roman, M., et al., 2021).

The table below shows the most popular word embedding methods, one similarity metric and the execution time, combining the cosine similarity with different embeddings.

Method	Similarity Metric	Target Year	Execution time (sec)	Most similar years
TF-IDF	Cosine	2008	295	[2010, 2009, 2007, 2013, 2011, 2004, 2015, 2012, 2014, 2016]
Word2Vec	Cosine	2008	105	[2014, 2015, 2005, 1997, 2012, 2008, 2003, 2001, 2010, 1991]
Doc2Vect	Cosine	2008	147	[2009, 2011, 2010, 2013, 2002, 2003, 2015, 2012, 2014, 2007]
Transformers bert-base-nli-mean-tokens	Cosine	2008	106	[2011, 2016, 2015, 2014, 2013, 2012, 2009, 2010, 2019, 2020]
TransformersT- Systems- onsite/cross- en-de-roberta- sentence- transformer	Cosine	2008	162	[2020, 2018, 2017, 2019, 2016, 2015, 2014, 2013, 2012, 2011]

Figure 44 - Document similarity experiment (Neto, J., 2021)

The best combination of NLP algorithms to determine document similarity for this experiment was Word2Vec-Cosine with 105 execution time (sec). However, this does not mean that this is the best choice for every dataset.

Strengths and Weaknesses: The ability to respond to text and speech for students' questions is a strength of the application. However, the way it processes writing and speech was designed only for regular students. The reality is that there are many universities worldwide with students with learning difficulties/disabilities who will not be able to use the application. The data preparation goes through many processes (in our case, sentence tokenisation, word tokenisation, stemming and lemmatisation), which can increase the time the chatbot needs to respond to the student.

In contrast, the data cleaning phase reduces the number of words in the question, which helps in processing the following stages. Finally, the processing of word embeddings runs the risk of the word not being contextualised and the question losing its meaning. Despite this limitation, it is essential for the machine to understand the student's question.

Embeddings generate vectors from text, and the combination of these vectors can generate a matrix, and these steps transform human language into computer language.

Matrix similar: Going back to Matrix factorisation, we find LSA (latent semantic analysis), which is used as a strategy for creating low-dimensional word representations. Using low-rank estimations simplifies huge matrices that catch measurable data in the corpus.

The next section describes text similarity measurement, the Q&A model and outputs of students' questions.

5.4.3 Text Similarity Measurement

This component measures text similarity. Measuring the semantic similarity between two text fragments has been one of the most challenging tasks in natural language processing. Various methods have been proposed to measure semantic similarity over the years.

For this framework, we chose Cosine Similarity as the measurement of text similarity. Cosine Similarity establishes the similarity between the student question and questions in the dataset (Gunawan, D., et al., 2018). Cosine Similarity examines the matrix with the student question and then goes through the questions in the knowledge database using term frequency to select the 10 most similar student questions from the knowledge database. Let A and B be two vectors for comparison as an example.

As a practical example, we have: Cosine similarity between two term-frequency vectors, we have \mathbf{x} = (*the vector of: **How much sleep do I really need?***) and \mathbf{y} = (*the vector of: **What is the amount of sleep I require?***), which are the first two term-frequency vectors, which converted into machine language become $\mathbf{x} = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ and $\mathbf{y} = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$. How similar are \mathbf{x} and \mathbf{y} ? The cosine similarity between the two vectors is determined below:

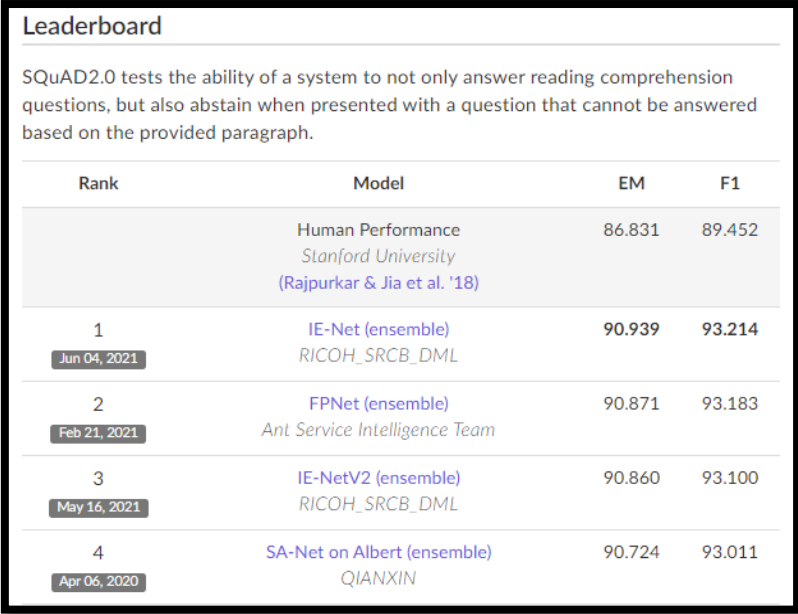
$$\begin{aligned} \mathbf{x}^t \cdot \mathbf{y} &= 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 \\ &\quad + 0 \times 0 + 0 \times 1 = 25 \\ \|\mathbf{x}\| &= \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48 \\ \|\mathbf{y}\| &= \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12 \\ \text{sim}(\mathbf{x}, \mathbf{y}) &= 0.94 \end{aligned}$$

Equation 13 - *Cosine similarity calculation*

One of the chatbot's main innovations is the data structure and the similarity algorithm. The data structure used in this research saves the questions students ask, which means that the chatbot learns the different formats the students use to ask specific questions. Consequently, there is no need to fine-tune or retrain the model to enhance the chatbot's understanding of the questions. Instead, the chatbot saves the question each time it is asked. Then, the algorithm will select the question with the highest similarity score for that particular answer, which means that the system will learn the student language without fine-tuning the student language question - answer model. The model for fine-tuning our model using the pre-train data is presented below.

5.4.4 Q&A model, Transformer model

There are several Q&A models, and the state-of-the-art models can be found in the SQuAD benchmark dataset. For this research, we will use the BERT model, which stands for BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. We fine-tuned the BERT model on the SQuAD dataset benchmark (Vaswani, A., et al., 2017). The Transformer model is a deep learning technique model introduced in 2017 at the time of the innovative paper “Attention is all you need” (Vaswani, A., et al., 2017) that utilises attention as a mechanism. The attention mechanism allows the chatbot to understand the context of the students’ questions. Below, we present the state-of-the-art Q&A models.



The image shows a screenshot of the SQuAD2.0 Leaderboard. At the top, it says 'Leaderboard' and provides a brief description: 'SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.' Below this is a table with four columns: Rank, Model, EM, and F1. The table lists several models, with the top one being Human Performance from Stanford University. The next four entries are ranked 1 through 4, each with a date badge indicating when they were achieved.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) RICOH_SRCB_DML	90.939	93.214
2 Feb 21, 2021	FPNet (ensemble) Ant Service Intelligence Team	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) RICOH_SRCB_DML	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) QIANXIN	90.724	93.011

Figure 45 - SQuAD Leaderboard Nov 2021

What these models have in common is that they use an attention mechanism like the transformer model.

The transformer model is applied to the top 10 most similar questions selected. As a free tool and referential for deep learning research projects, Colab Google is one of the best available, and this is the tool we used to generate the Question Answering model for our chatbot. Below is an example using the model:

```
# Run our example through the model.
outputs = model(torch.tensor([input_ids]), # The tokens representing our input text.
                token_type_ids=torch.tensor([segment_ids]), # The segment IDs to differentiate question from
                answer_text
                return_dict=True)
```



```
start_scores = outputs.start_logits
end_scores = outputs.end_logits
```

We then develop the model using a huggingface transformers library because of its simplicity and efficiency and the fact that it is free. Then we fine-tune the model. Previously, we downloaded the Pre-trained BERT model.

```
from transformers import BertForQuestionAnswering
model = BertForQuestionAnswering.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')
```

Strengths and Weaknesses: Briefly, the transformer model overcomes the long-term dependency problem. It also deals with vanish and explosion problems found by previous models. Another advantage is that it also solves the context problem. However, it still has the vanilla transformer model problem. The transformer model still has computational cost problems.

5.4.5 Assembly parts

This component is the linking component between the first part of the framework and the second part, resulting from the student's question. We have three possible questions as chatbot answers. Even though adding more than one possible answer breaks the natural flow of a conversation, it increases the chances of the chatbot giving a correct answer and reduces the possibility of the student saying that this was not the question he/she would like to see answered. This approach is appropriate for universities that want to provide more assertive assistance to their students. We show the interactions between chatbot's object using Sequence diagrams in Figure below.

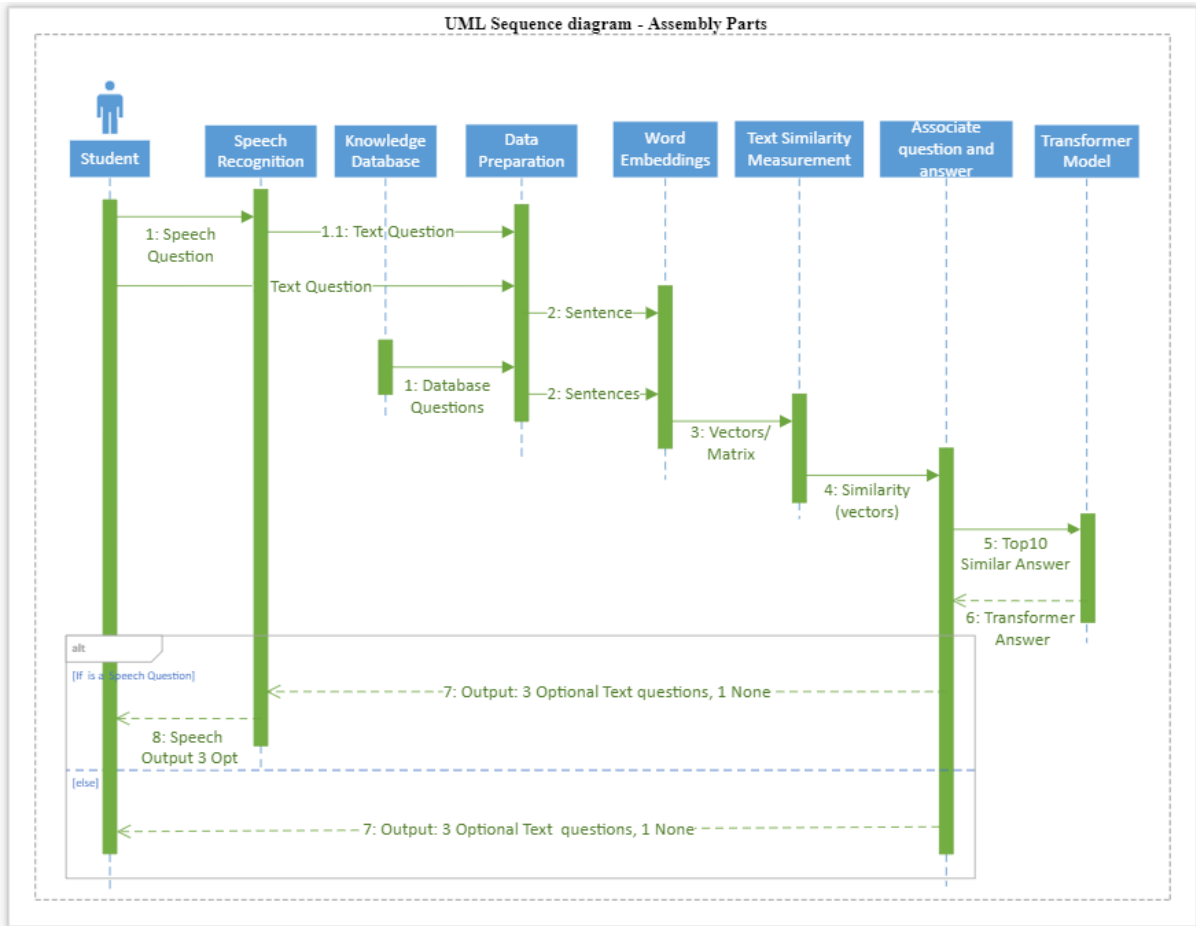


Figure 46 - UML Assembly Parts

In this diagram, we present the proactive chatbot's answer to the student's question. We have represented seven objects and the actor, who is the student. The student asks the question via writing or speech. If it is via speech, the speech recognition object transforms the question from speech to text. In parallel, the questions in the knowledge database are transferred to the data preparation, where they are processed and sent to word embeddings to be transformed from sentences to vectors or matrices. Next, the text similarity of the questions is calculated, and then the ten questions are selected for the next step. This selection is based on more similar questions on the data knowledge to the students' question that it is associated with a respective answer, and the transformer model searches for the answer within those ten selected questions, on the answers related to ten questions. Once the answer is found, it passes to the previous object, which selects three answers, the first being the Transformer model's answer and the next two the most similar questions. Finally, the student's question is answered via speech or writing.

In that case, this question-and-answer pair will be the first option among the three questions sent by the chatbot. The second and third options are the two most similar questions of the ten, excluding the question selected by the Transformer model. Finally, if the chatbot finds questions similar to the student's question but does not think it was the student's question, it gives the student the option of choosing 'None'.

Strengths and Weaknesses: The weakness of this component is that it breaks the natural flow of the conversation between the student and the proactive chatbot, because when a question is asked, one answer is expected, not four options. Nevertheless, this component gives the proactive chatbot more chances to answer the student's question correctly.

5.5 The Educational Chatbot Ecosystem

The second part framework is the ecosystem that allows the student to have more personalised assistance and interact more with the proactive chatbot. It is presented in the figure below.

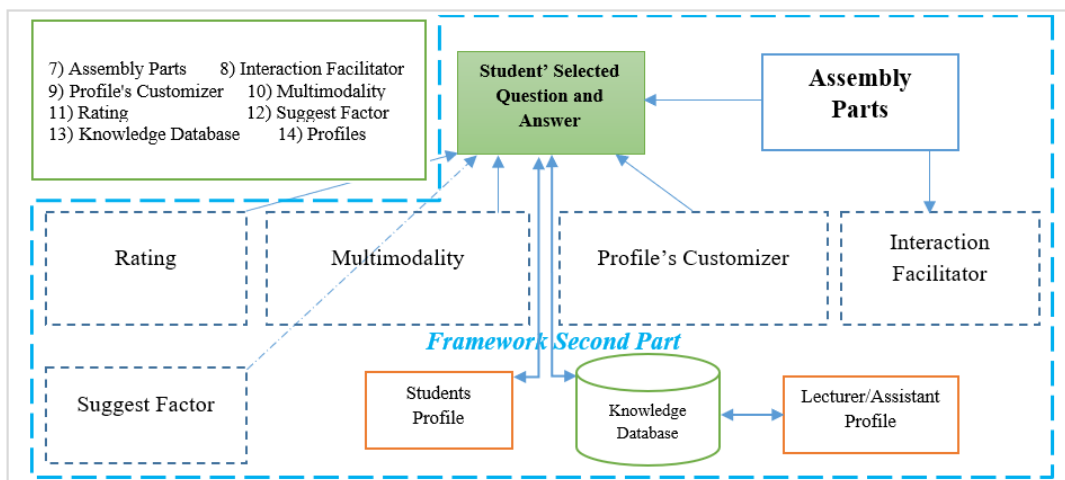
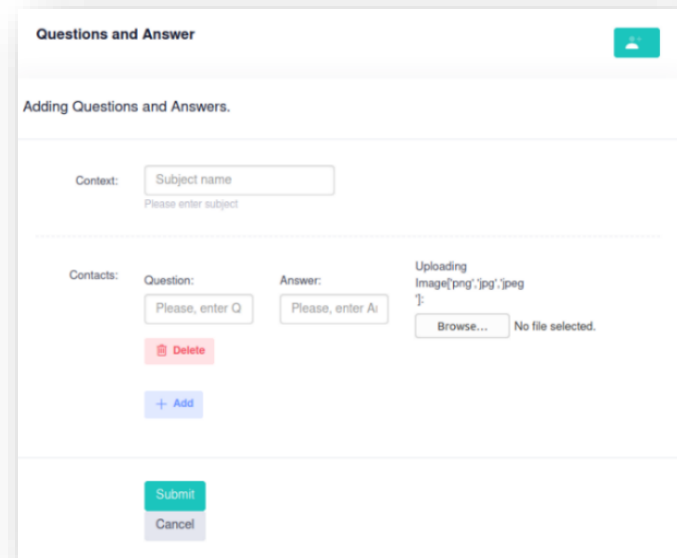


Figure 47 - The second part framework

This phase starts when the student selects the proposed questions given for the previous stage. Then, in the following sections, we present the components of the second part of the proposed framework.

5.5.1 Interaction Facilitator

This component facilitates the interaction between the student and the lecturer or human assistant. One of the most glaring problems of the current chatbot is the usability issue. When the user does not find the answer to his question, he may stop using the chatbot, which impacts its usability. This issue becomes more accentuated when dealing with students who do not have much patience.



The screenshot shows a web interface titled "Questions and Answer" with a cloud icon in the top right corner. Below the title is the heading "Adding Questions and Answers." The interface is divided into several sections:

- Context:** A text input field containing "Subject name" with a placeholder "Please enter subject" below it.
- Contacts:** A section with three input fields: "Question:" (placeholder "Please, enter Q"), "Answer:" (placeholder "Please, enter A"), and "Uploading Image" (placeholder "Image['png', 'jpg', 'jpeg']" and "No file selected." with a "Browse..." button).
- Actions:** A red "Delete" button, a blue "+ Add" button, a green "Submit" button, and a grey "Cancel" button.

Figure 48 - Assistant adding Q&A

When the student selects the 'None' option, the chatbot sends the question to the lecturer or the human assistant. When they answer the question, it is saved in the knowledge database to be available for all subsequent users and to improve it. Besides, lecturers and human assistants can also add new questions - answers and new subjects from their profiles.

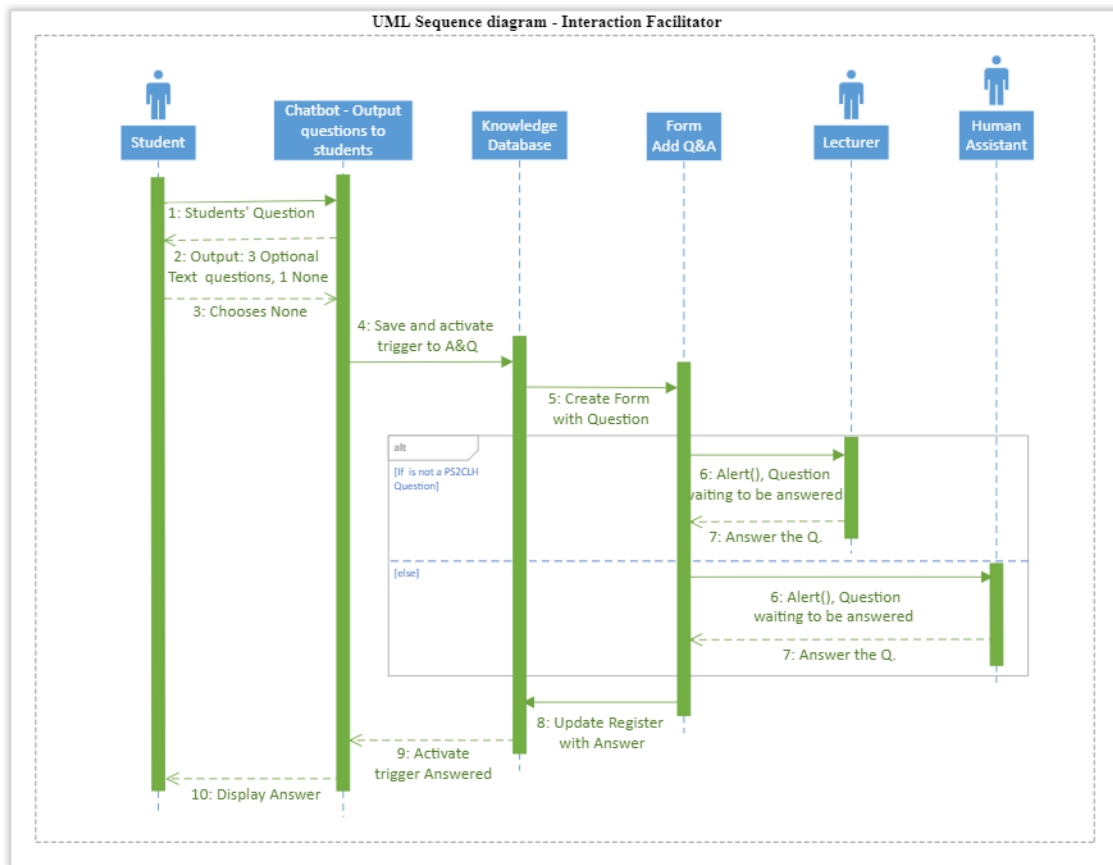


Figure 49 - UML Interaction Facilitator

We have three actors - the student, the lecturer and the human assistant. The student asks the question and the proactive chatbot gives four options. When the student chooses the 'None' option, the question is saved in the knowledge base, which creates a form with the question and alerts the lecturer or human assistant that there is a question from the student which has not been answered by a proactive chatbot. When the question is answered, the knowledge database is updated, and a trigger is activated for the question; when the student is again in front of the proactive chatbot, the unanswered question appears with the answer given by the lecturer or human assistant.

Strengths and Weaknesses: The weakness of this component is that the lecturer often does not have time to answer the questions. If the lecturer gets too many questions, he/she will likely leave the student unanswered, or he/she might answer it too late when the student has already investigated and found the answer. The strength is that it creates an indirect interaction, so that the student can feel more comfortable asking the lecturer or the human assistant. The following section describes how the chatbot responds to students' profiles.

5.5.2 The Profile Customiser

This framework component customises the student profile, so that the proactive chatbot works according to each student's profile. This is defined by the PS2CLH model and the selected proactive chatbot persona, thus organising the controllable factors that affect student performance in order of importance and making the proactive chatbot act in accordance with the student's profile.

There is general agreement that students pay more attention and are more engaged in front of a person they desire as a teacher. It is a fact that there are preferences for different teaching styles and attitudes in classes. We have therefore replicated this in the framework, in addition to the student profile being developed by completing the PS2CLH questionnaire, which will make the student focus on the factors that present problems. That is, the student profile includes the controllable factors that most affect that particular student and are organised in such a way that the student focuses on those that have the most impact on their academic performance.

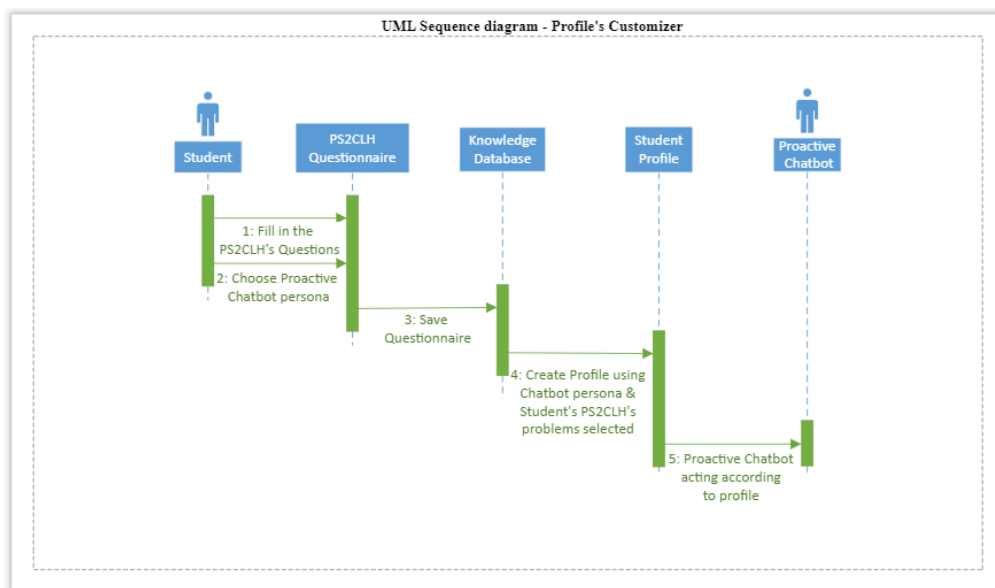


Figure 50 - UML Profile's Customizer

This sequence diagram is simplified because we still have the entire PS2CLH model building process and the proactive chatbot persona to work on. However, we have two actors and three main objects. First, the student fills in the PS2CLH questionnaire, in which he chooses the persona he/she would like or prefer not to have for the assistant. The questionnaire is then saved in the database, so when creating the student profile, we go to the database to select the problems that the student would like to work on and the type of interaction chosen by the

student. Finally, when the student interacts with the proactive chatbot, the latter will have the properties selected by the student, thus acting according to the student's profile.

The persona chosen for the chatbot by the student is based on four pairs (Introversion vs Extraversion, Sensing vs Intuition, Thinking vs Feeling and Judging vs Perceiving). The chatbot will act according to the persona chosen by the student. However, this approach still has many representational limitations in the proposed framework.

Strengths and Weaknesses: One of the weaknesses is the complexity of representing the 16 different personality types in terms of programming. The other problem is the fact that people generally like variation in their dealings with other people. For example, an assistant needs to know when to be more logical than emotional or vice versa. As things stand, , the proactive chatbot will not have the flexibility to behave in a more extroverted way once it has been programmed to have introverted attitudes (or responses). The key point is that the people tend to be habitual in the way they behave. Customising the student profile can give the student more control.

Next, we discuss the component responsible for the presentation in multimodal form of the content of the chatbot's responses.

5.5.3 Multimodality

In this component, we present the diversity of the ways chatbot responses are presented. The chatbot displays the answer according to the student's learning style, using text, image\graphic\diagram, video or voice. A university has several faculties with different courses. Course contents vary. In this variance, we find several disciplines such as history, geography, marketing, computing, music, archaeology, graphics, images, and so on.. In addition, there are things that we naturally learn by listening, seeing or reading. Therefore, by adding this component to the chatbot, we hope to spur curiosity in learning and the same involvement in the way students deal with the chatbot so that we can offer assistance in the various subjects and factors that affect student performance.

In the registration phase, we collect student data to create their profile and predict their academic performance. Students can create their profile during their registration using a self-evaluation questionnaire, selecting their PS2CLH problems. Then, the proactive chatbot

personalises the answer when it is possible. The multimodal personalisation of the answer works in the Knowledge database. When we save the question and answer, the image or graphic is saved, as well as the video link if applicable. Finally, the proactive chatbot will turn on the speech recognition functionality if the student selects the speech mode.

Strengths and Weaknesses: Not all content can be represented in the best or the expected way. Sometimes, for example, multimedia content can distract the student. In contrast, some contents are more easily represented using multimedia resources.

5.5.4 Rating

This is the student's performance evaluation component of the proactive chatbot, the feedback on the interaction between the proactive chatbot and the student. What is the motivation for giving the proactive chatbot feedback and a rating for the answers it provides? The title answers this question. The primary purpose of getting feedback from students is for the chatbot to understand how helpful its responses have been so that the course lecturer can improve the responses and the human assistant can improve the PS2CLH responses, aiming at the overall qualitative improvement of the chatbot responses. The other motivation, for getting a rating of the answers, will make the algorithm developed in the proactive chatbot give more weight to the most relevant questions and pay less attention to the questions with a lower rating, so that it automatically improves the answers given by the proactive chatbot in the future.

Once the chatbot has finished answering, the system asks the student to give feedback, rating the chatbot's answer on a scale of 1 - 5 where 1 is unsatisfactory and 5 is very good. This will help to improve enhance the chatbot's responses, and results in the students' questions being saved only when the rating is equal to or greater than 3. This procedure will help to ensure that it is the questions most closely related to the answers which are saved. Yet ratings below 3 will also be saved but the poor questions will not be, allowing the lecturer or assistant to improve the response in the knowledge database for the future.

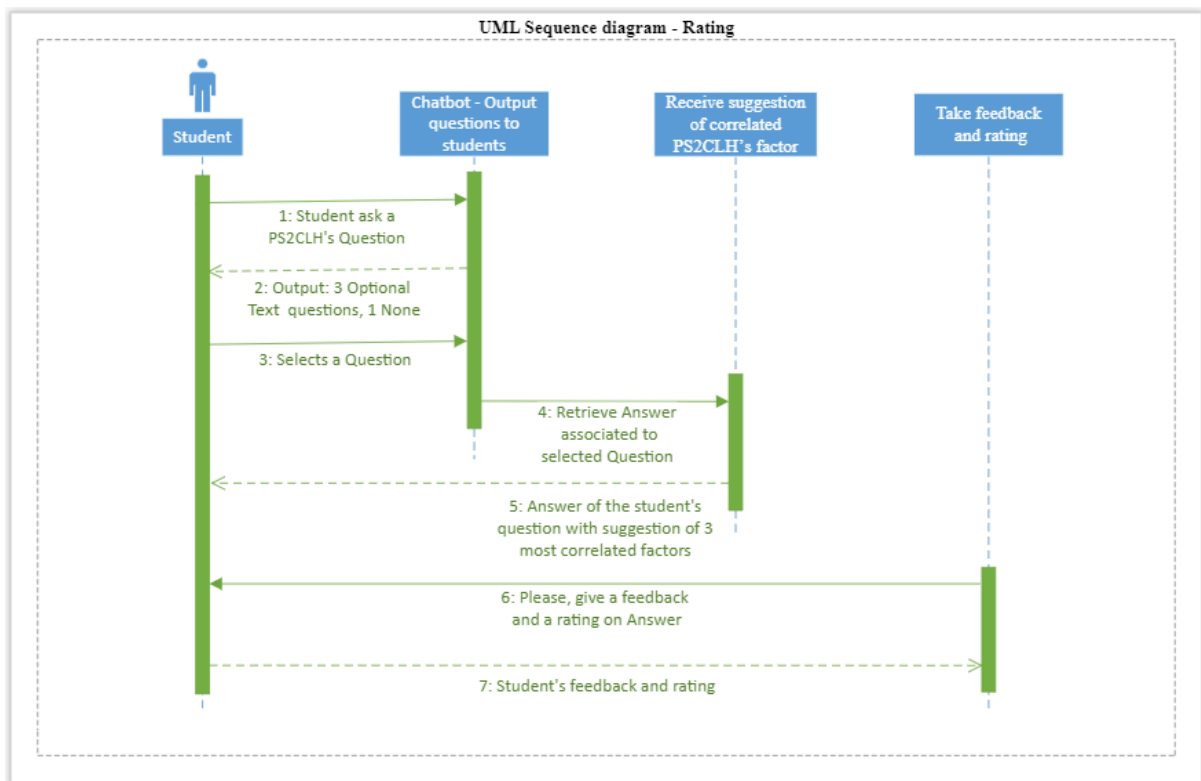


Figure 51 - UML Rating

When the student receives the answer to their question, the chatbot also sends a rating on a scale of 1 to 5 where 1 means the student is not satisfied with the answer and 5 that the student is satisfied and gives the answer five stars. The chatbot sends student feedback to the knowledge base so that it will know the best and worst answers. This enables lecturers and assistants to delete or improve the worst answers in their profiles. After the proactive chatbot has had many interactions with students, all the conversations are saved, and the chatbot Transformer model is improved by retraining the model with all this new data.

Strengths and Weaknesses: Usually, people do not give the rating, and this makes for a less natural conversation. However, the vital point is that the proactive chatbot will improve the Q&A using the structure of the knowledge database.

5.5.5 Suggest Factor

This component is what makes the proactive chatbot have a more proactive approach. Generally, universities, student assistants and even typical students are unaware of the existence and impact of controllable student factors that affect their students' academic achievements. It is almost humanly impossible for a regular assistant to know the correlated

factors for each student in a university of ten or twenty thousand students. The number of controllable facts is high and complex, as it varies between countries and between universities. Often the student can ask about a factor, and the average chatbot would respond reactively, unlike the proactive chatbot, which proactively suggests to the student factors that may have a more significant impact on the student's results. It may redirect the student to look at the root (what generated the problem the student is facing) of the problem and not just its consequences.

If the student's question is related to the PS2CLH model, proactively, the chatbot suggests they should work on the correlated factors which are affecting their performance. Correlating the variables or factors that affect students' results will allow the chatbot to make individual suggestions, giving a specific answer - for instance, if a particular student has a problem with stress and wants to know more about dealing with it during the interaction between chatbot and student. In the example below, the student first asks about sleep problems. Then, the suggestion is about stress and anxiety, time management, and establishing and achieving personal goals.

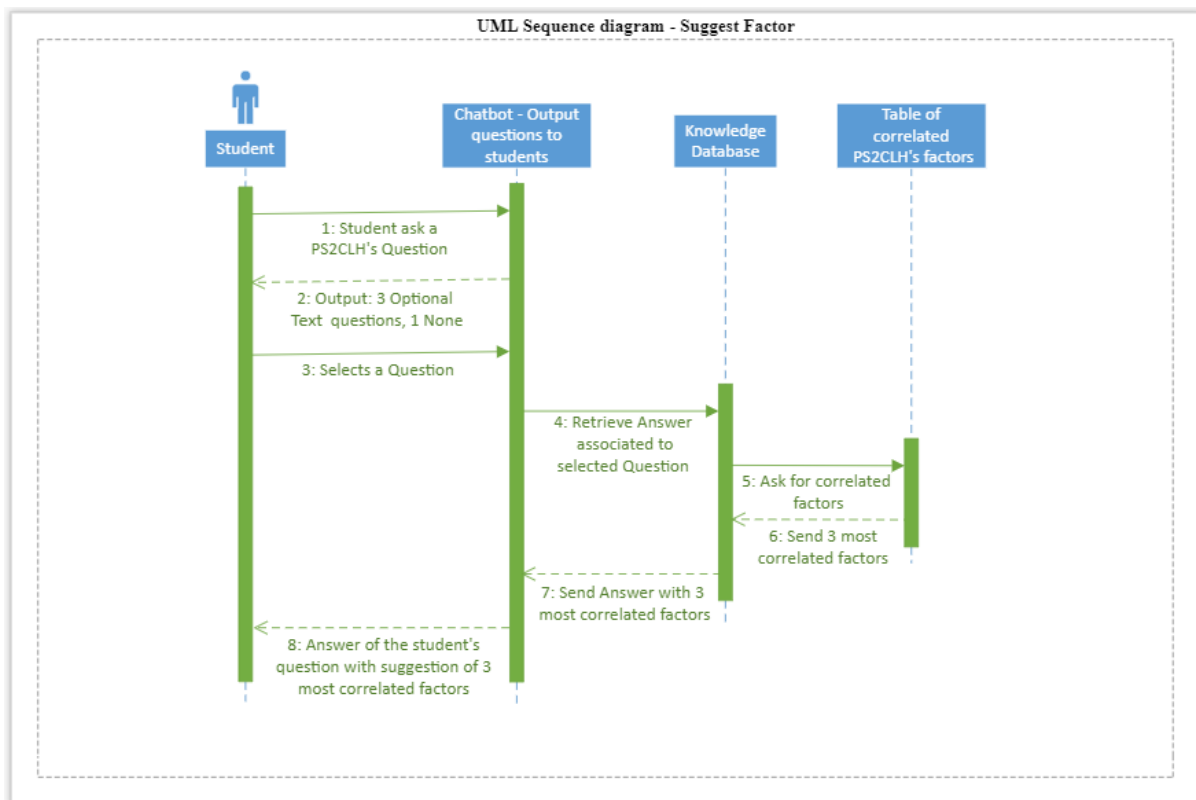


Figure 52 - UML Suggest Factor

As in the previous diagrams, the student is the actor, and in this case, we have three objects. The student asks the question, and the proactive chatbot responds with four options. When the student selects one of the three suggested questions, the answer is retrieved in the knowledge

base. After the question-and-answer association, the table of the object is requested. Then, the correlated PS2CLH's factors are also requested, the system responds by sending the three most closely correlated factors. Finally, it sends the answer with the three most correlated factors to the proactive chatbot and then it answers the student's question associated with the three most correlated factors.

It uses weightings to quantify the correlation among the factors. The system will know which factors impact the most on the other factors. Therefore, the suggestion will be according to the level of correlation among the factors. Thus, it leads to building clusters of correlated factors.

Strengths and Weaknesses: The weakness is that this can distract the student from working on the factor they asked about. However, on the other hand, the student will have more assistance and more knowledge related to the factor he is concerned about, thus opening the possibility for the student to find the root or actual cause of the problem he is facing.

5.5.6 Knowledge Database

The file JSON has the proactive chatbot's knowledge database of each subject, which has the questions and answers from the PS2CLH model and the lecturer's subjects. The structure of the dataset.json file is presented in Figure 47 below.

```
{
  "context": "sleepproblem",
  "qas": [
    {
      "id": "2",
      "is_impossible": false,
      "question": "Does the light from smartphone stop you from sleeping effectively?",
      "answers": "Dr Ong, In general, the light from your screen can affect your circadian rhythms. Using light",
      "rating": [4,3,5],
      "urlImg": "/static/basic_app/profileStudent/dist/assets/media/users/300_12.jpg",
      "urlVideo": "https://www.youtube.com/watch?v=j5S18LyI7k8",
      "studentQ": [
        {
          "idSQ": "1",
          "stdQ": "Does the light from pc negative affects my sleep?",
          "ratingSQ": [4]
        },
        {
          "idSQ": "2",
          "stdQ": "How does the light from my smartphne affects my sleep?",
          "ratingSQ": [5]
        }
      ]
    }
  ]
},
```

Figure 53 - Chatbot's JSON file structure

The “context” represents the student's subject, “qas” has the questions and answers. Students may ask questions in different ways; therefore, as we can see in the figure above, one answer can be used for many questions, giving the proactive chatbot the flexibility to improve the accuracy of the answer. However, by increasing the number of questions associated with an answer, we will have a more extensive knowledge base, which increases processing time, so we limit the number of questions associated with an answer to just three. We are seeking to make the proactive chatbot as efficient and as human-like as possible. The aim is to minimise the chatbot's response time and to improve the accuracy of the answers. The current proactive chatbot assists students in their subjects by interacting with them.

Strengths and Weaknesses: The weak point is the growing number of questions associated with an answer. In a large data set, this will negatively affect the response time, impairing the proactive chatbot's efficiency, causing the student to wait longer for the chatbot's response. The vital point is that it will enable machine learning of the students' terms and language,

making algorithms which are similar, so they work better after several interactions with students, thus improving the effectiveness of the proactive chatbot.

Below we present the main tables of the framework. Not all the tables and fields are presented, because they are not the main focus of the research.

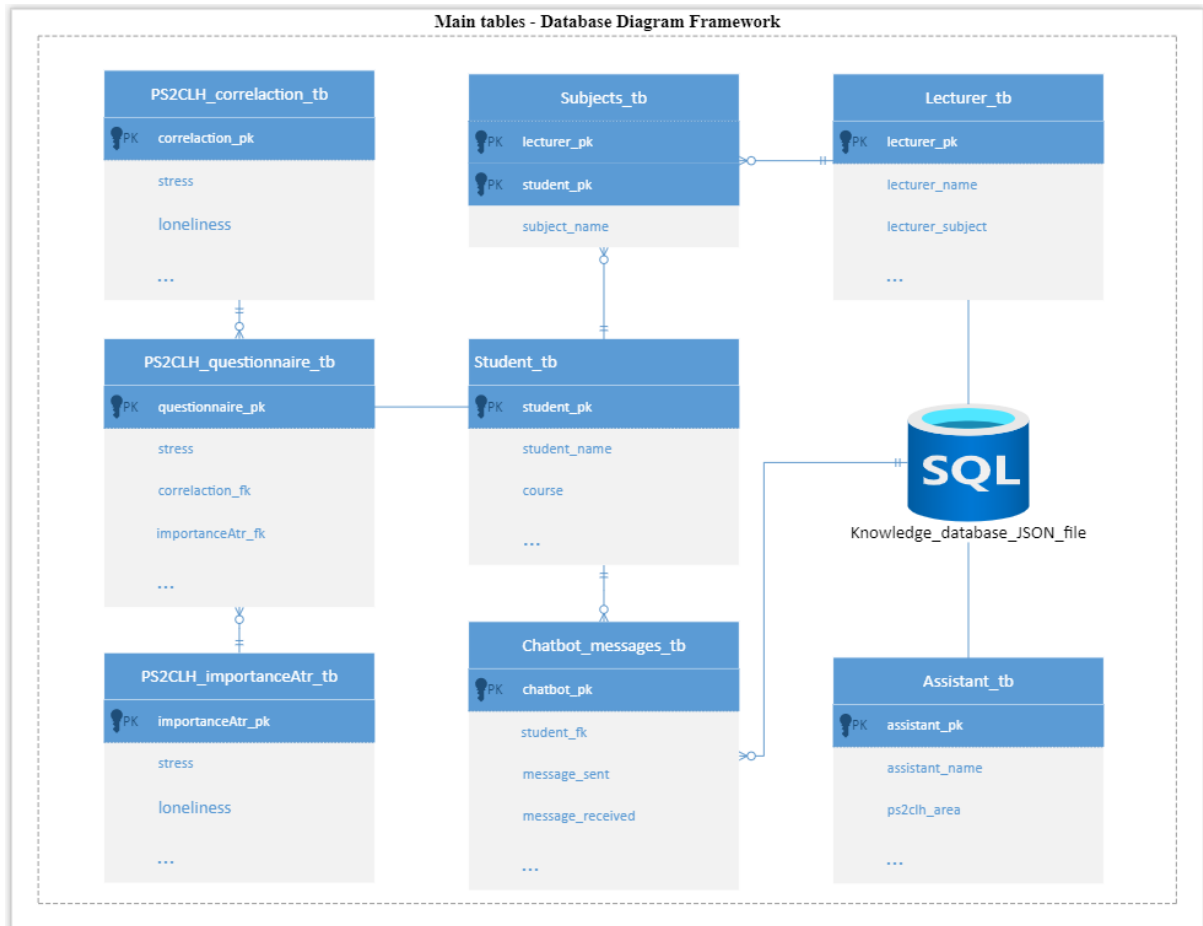


Figure 54 - Database Diagram

A student has a PS2CLH questionnaire. The PS2CLH_importanceAtr_tb table and the PS2CLH_correlation_tb table have a one-to-many relationship with the PS2CLH_questionnaire_tb, thus having the correlation_fk and importance_fk foreign keys. The knowledge base is linked to the chatbot, the lecturer and the assistant, and these last two (human) entities can add questions and answers. Finally, a lecturer can add multiple subjects, and a student can have multiple subjects.

Below are profiles of students, lecturers and assistants.

5.5.7 Profiles

This is the profiles component, where we present profiles of the students, the lecturer and the assistant. Below are the tools and programme language used to develop the Proactive Chatbot Framework. For this project, we used an environment called “chatchannels2” as a virtual machine. We use Django version 2.2, which works best with Python 3.6 into Anaconda. The local running application address is `http:127.0.0.1:8000/`. The web development framework uses Django Channels, which adds more complexity to the application but improves the system’s security and confidentiality. We create a channel for each subject which is visible only for students registered on that subject. Channels also allow a community to be created where students can interact and discuss the topic and create a one-to-one channel which allows private interaction with the system.

Using the Ubuntu OS, we developed the Django project, and created two applications. The first one is the social network with the students’ profiles, the assistants’ profiles and the lecturers’ profiles. The second application has the chatbot app, which uses Django channels. The database is `db.sqlite3`, and this saves the social interaction. In addition, we used a JSON file to save the chatbot conversations.

The application starts with the site’s manager, three sites (Student Profile, Lecturer Profile and Assistant Profile) and the Admin site. For the Student Profile site, the user has to sign up to be able to log in.

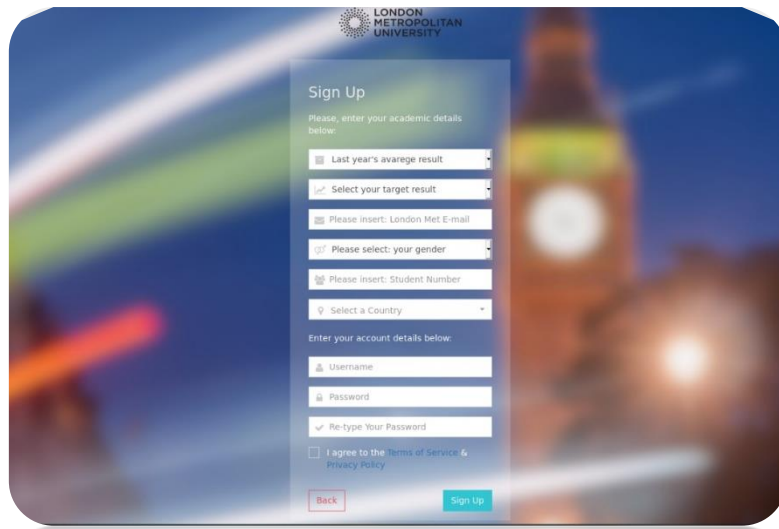


Figure 55 - Student's Sign Up

It asks for the standard sign up requirements and the “Last year average result” field that is used to for the students’ result prediction model. Then we write the target result for the current year, which intends to give students a clear picture of the desired result.

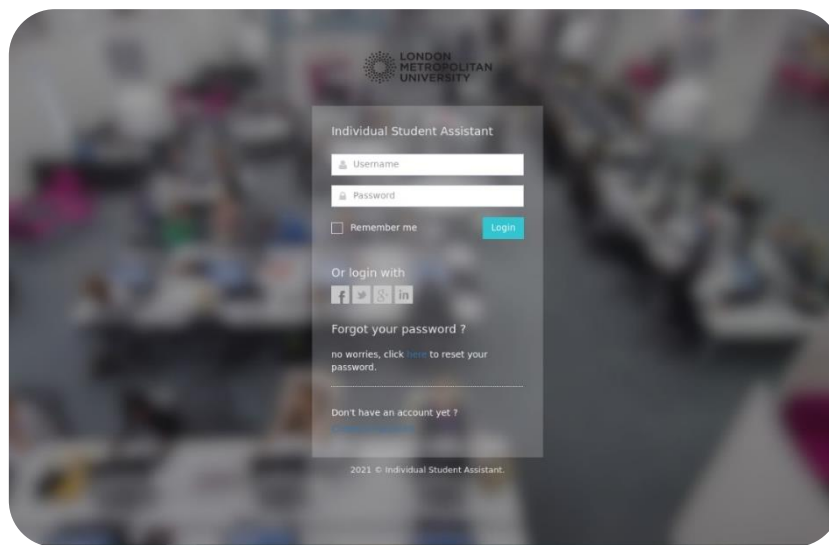


Figure 56 - Student's Login

This is a standard login page, where students enter their details to register, log in to complete the PS2CLH questionnaire, choose one of the 16 personality types for our chatbot, and fill in the learning style. The PS2CLH questionnaire is shown below, starting with the Psychology area.

Figure 57 - Students' Questionnaire

This latest version of the questionnaire was developed in the light of the most frequently selected factors in our last data collection exercise in 2020 (see Appendix), and this is not specific to one particular country. There are five questions for each area, each of which has symptoms and a link to redirect students so that they can find out more about the problem.

The questions in the questionnaire are presented as a radio button, allowing students to choose one of the five options. Below each question are the symptoms related to that question.

Carl Jung presented the theory of psychological types, which has the same purpose, individualise people by their different functions and attitudes of consciousness. Categorised by their preference of wide-ranging behaviours, three areas of preferences and dichotomies: Extroverted (E) vs. Introverted (I); Sensing (S) vs. Intuition (N); Thinking (T) vs. Feeling (F). Dr. Isabel B. Myers, continued with Jung's theory, which later she added one more field as a fourth antagonism per; Judging (J) vs. Perceiving (P) (Briggs. M, 1980)

Please, Choose the personality type you would like to have as Assistant, selecting one on each pair!

Extrov. vs. Introv. Extroverted (E) Introverted (I)
Turn to others are an (E), Turn inward are an (I)

Sensing vs. Intuit. Sensing (S) Intuition (N)
Pragmatic way are an (S), Creative way are an (N)

Think vs. Feel Thinking (T) Feeling (F)
Seek objective truth are a (T), Seek harmony are an (F)

Judg. vs. Perceiv. Judging (J) Perceiving (P)
Get closure and act are a (J), Stay open and adapt are a (P)

Figure 58 - Personality types selection

This is a component which needs more research and work on the implementation because there are still many limitations in the technology. For instance, how can we represent the pairs *Sensing vs Intuition* and *Thinking and Feeling* on the chatbot? The chatbot has a persona chosen by students based on the four pairs (*Introversion vs Extraversion* and *Judging vs Perceiving*). At this stage, students can choose the chatbot's personality by choosing one item from each pair. Finally, we present the student's results.

Thank you, for filling the Self-evaluation Student Questionnaire! Below are your PS2CLH Coordinates.

☰ Thank you, for taking your time to fill the PS2CLH questionnaire!
Explain why, how and what is a student's Cluster
Please, click on the button below to see your cluster...

Psychology Coordinate

18

Self-Responsability Coordinate

26

PsychologySel-Self-Responsability Coordinate

73.33333333333333

Sociology Coordinate

26

Communication Coordinate

26

Sociology-Communication Coordinate

86.66666666666667

Learning Coordinate

26

HealthWellbeing Coordinate

26

Learning-HealthWellbeing Coordinate

86.66666666666667

Centroid

73.33333333333333,86.66666666666667,86.66666666666667

I am in the cluster number:

7

Figure 59 – Students' PS2CLH coordinates

After students finish the questionnaire, they are represented by their coordinates for each area Psychology, Self-responsibility, Sociology, Communication, Learning and Health & wellbeing. The figure above also shows the combined pair of coordinates, the student centroid, and the initial student cluster. Then the application lead to the Student's profile.

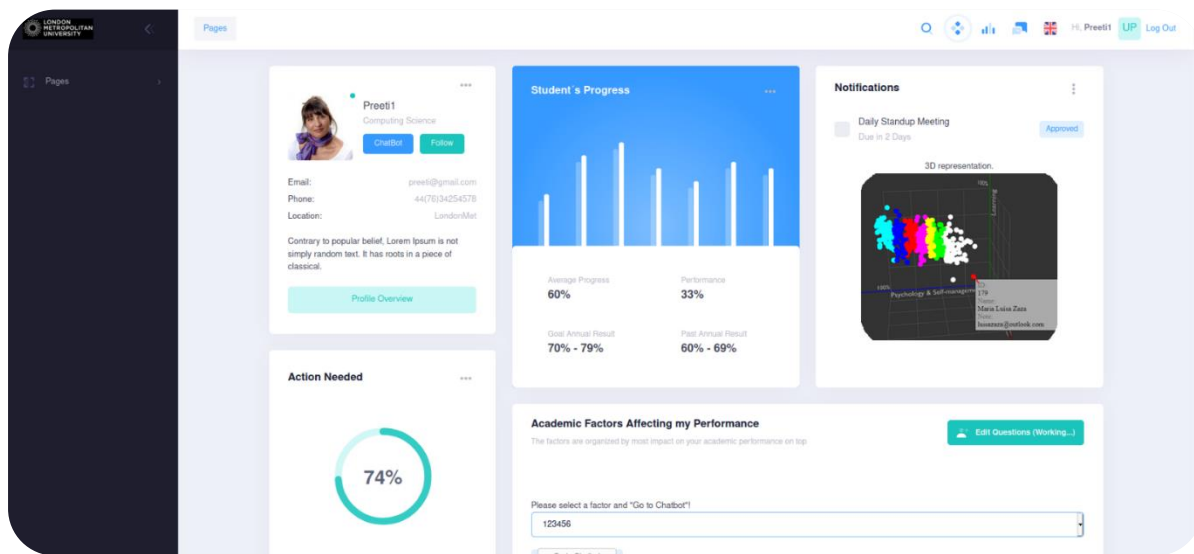


Figure 60 - Student's Profile

The student profile has the questions he or she selected from the PS2CLH questionnaire. It also has the student coordinates and the cluster, and presents statistics and a 3D representation of the student, allowing us to see how the students in our University are doing. Finally, the student has the proactive Chatbot that helps him with the factors and the material given to him by the lecturer and the assistant. In future, the assistant will have in his profile the questions not answered by the chatbot and the form that will allow him to add questions and answers, which will go in the knowledge base. The assistant will also be able to start a chat with the student, if it is necessary, especially for the field of psychology, where human interaction is fundamental to the solution of many problems.

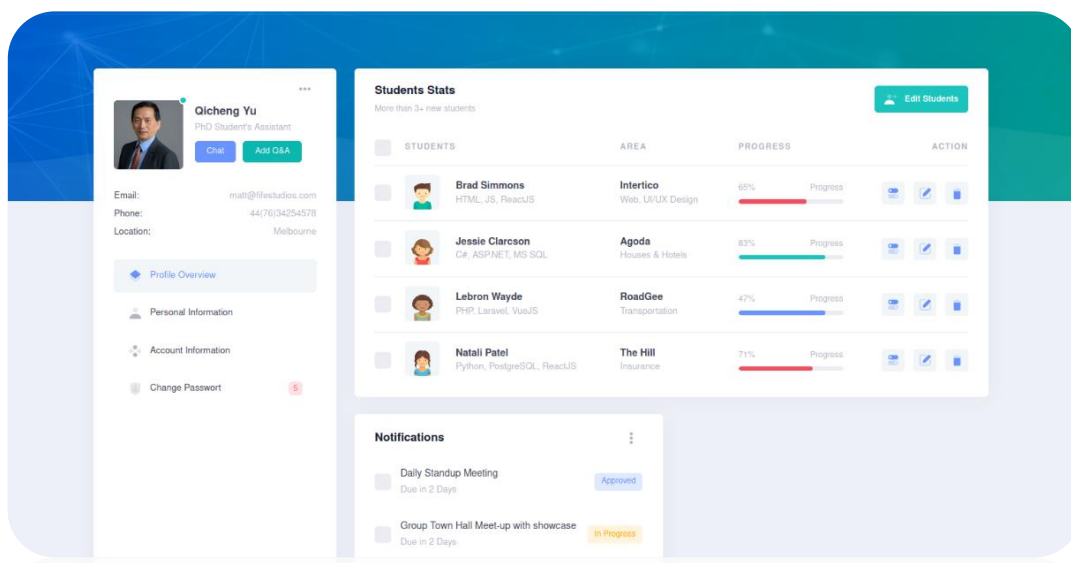


Figure 61 - Assistant's Profile

In the lecturer profile, we will also have a form which allows us to add questions asked by students, which are not found in the knowledge base, together with answers.

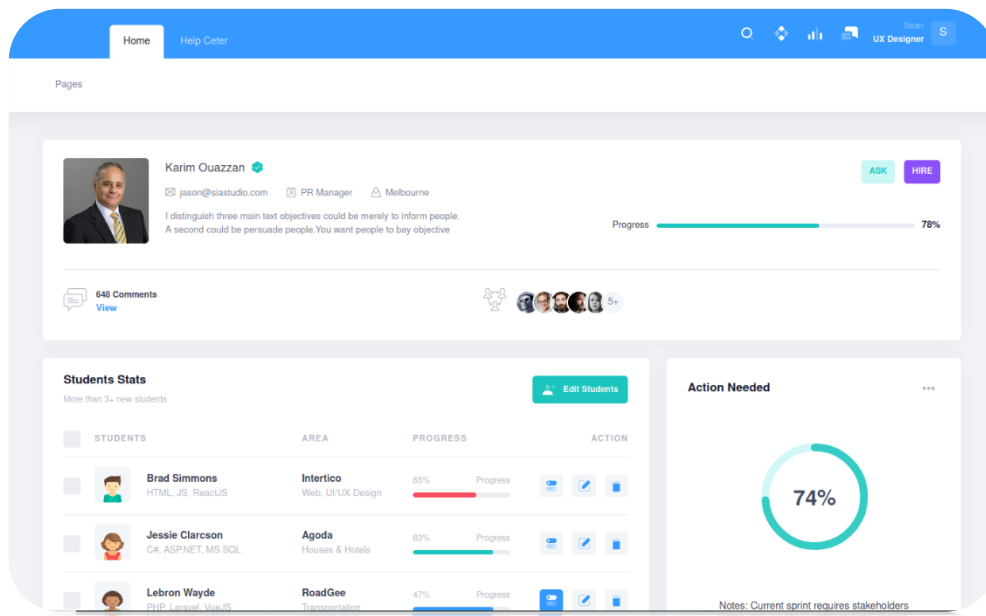


Figure 62 - Lecturer's Profile

5.6 Testing the proactive chatbot framework and extending the BERT chatbot: Results

We intend to test the proposed framework for a proactive chatbot for students based on the PS2CLH model.

The application tests of the efficiency of a model or architecture usually analyse the chatbot's performance or response time. In this case, we intend to take different approach. The framework is different universities in different countries, but these universities already use certain technologies that may be integrated into the proposed framework in the future. This implies that we will require different implementations to meet different needs, so the type of input language, word embeddings, text similarity measurements and question and answer models will vary accordingly.

We therefore do not aim to test which word embeddings (e.g. word2vec, TF-IDF or MV-LSTM) work better, or combinations of embeddings and text similarity measurements (such as

cosine similarity, Euclidean distance, Manhattan distance, Hamming distance, distribution distance or semantic distance) or whether different Q&A models should be used.

The aim of our tests is to test the accuracy of the chatbot’s answers and then compare the results of an extending chatbot with the proactive chatbot framework. Below we present the test setup.

Test Setup: We created two applications to perform the test. The first is the proactive chatbot framework, which we will call the “proactive chatbot”; the second application is an implementation of the Extending Bidirectional Encoder Representations from Transformers, and we will call this the “extending chatbot”. To ensure that the test would be fair and realistic, the two applications were programmed with the same word embeddings, the same text similarity measurements and the same Q&A model. The difference lies in the components added in the proactive chatbot, i.e., the methodology for presenting the answer to the student’s question. Below is a small sample of the questions used in the test.

	<i>Question (Knowledge D.)</i>	<i>Student’s questio: Simple</i>	<i>(Student’s question: Complex</i>
1.	What is a database?	Could you define database?	What is a data repository?
2.	How can I manage my time better?	How can I prioritise my time better?	How can I prioritise my time effectively?
3.	What are the objectives of SQL?	Please tell me the objectives of SQL.	Please explain the objectives of the Structured Query Language
4.	What are the symptoms of stress?	What are the side effects of stress?	What side effects does stress have?
5.	Why is it important to learn how to manage stress?	Why is it important know about stress?	Why is it important to understand stress and how to manage it?.
6.	What ways of trying to cope with stress are unhealthy?	Are there healthy and unhealthy ways of dealing with stress?	What ways of dealing with stress should I avoid?
7.	Why is it important to connect with others to deal with stress?	How can connecting with others help me deal with stress?	Why is it essential to work with others to solve stress problem?
8.	What things should I make time for to deal with stress?	How can I make time for fun and relaxation?	When I’m feeling stressed, how can I have fun and relax?
9.	How can I manage my time better?	What suggestions are there for managing time?	How can I improve my time management?
10.	How common is stress?	Is stress a common problem?	Is stress a frequent issue?
11.	Does stress affect your performance?	What are the symptoms of stress?	How can I deal with stress?.

12.	Why do people sometimes get frustrated?	What is a simple way to cope with frustration?	How can I deal with frustration?
13.	Why do I keep putting things off?	What can I do about procrastination?	Is there an effective way of avoiding procrastination?
14.	What can I do to stop anxiety?	How can I overcome anxiety?	Is there anything I can do to eliminate my anxiety?
15.	Why do I worry so much?	Why is it so difficult to stop worrying?	Is there an effective way of stopping worrying?
16.	Why am I sometimes anxious?	How can I avoid anxious thoughts?	What is the best way of overcoming anxiety?
17.	Is it possible to deal with uncertainty?	Are there effective ways of combatting uncertainty?	Is it possible to live with uncertainty?
18.	How can I tell if others affect me?	Should I try not to let others affect me?	Will it help if I can understand how others affect me?
19.	How can I practice mindfulness?	Is there a way of training mindfulness?	How do I learn to exercise mindfulness?
20.	Is it necessary to be creative?	How can I become more creative?	Are there effective ways to develop creativity?
...
100.	How much sleep do I really need?	What are the effects of insufficient sleep?	How can I overcome insomnia?

Table 5 - Small sample of the questions used in the test

The two applications used the same dataset of 100 questions and answers. The test consists of two types of question. The first is a simple question, i.e. one which uses some keywords from the question in the knowledge database. The second type is a complex question, i.e. one which asks the same thing as the first one, but uses different words.

Results: Please find below an example of the “*proactive chatbot vs extending chatbot*”.

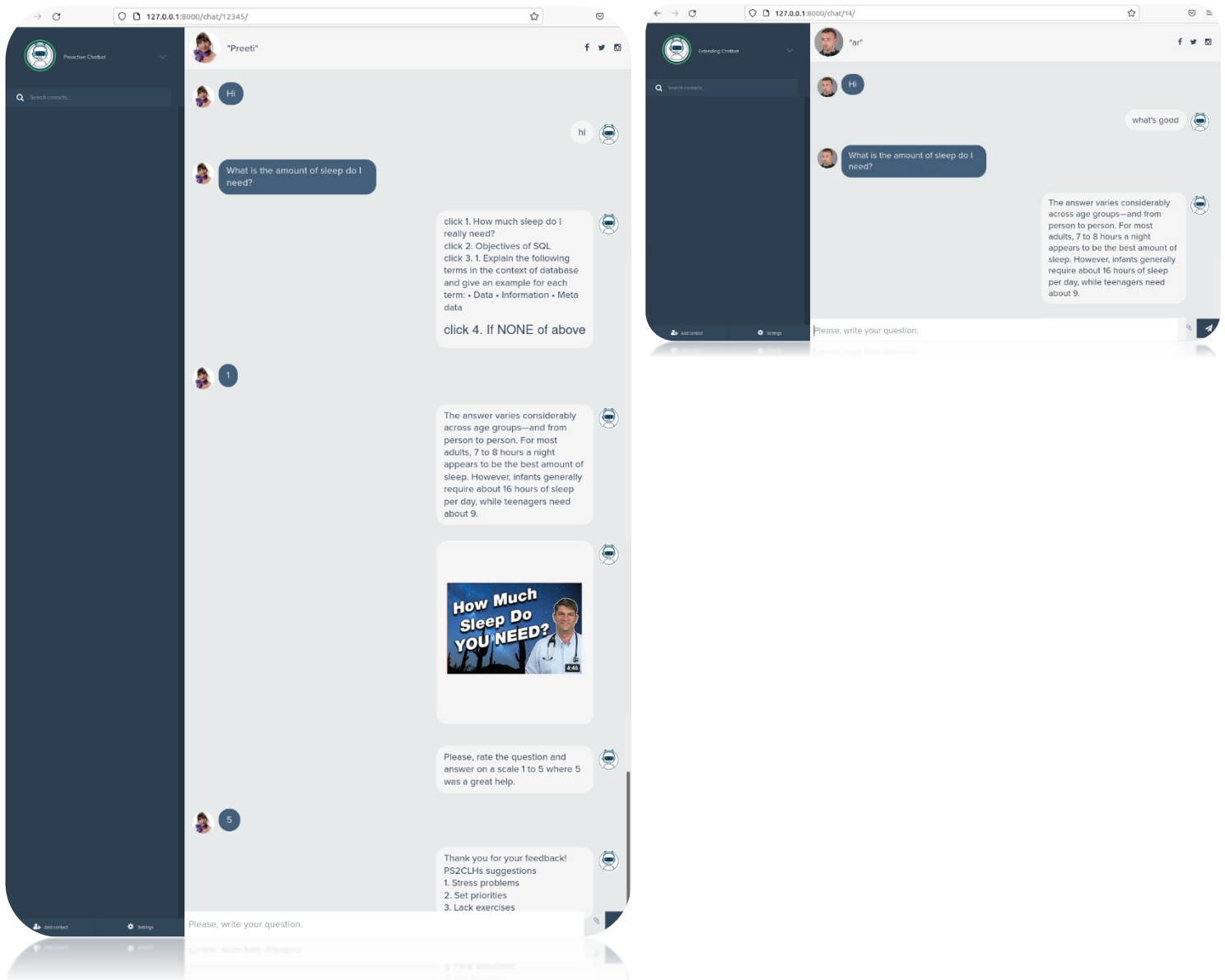


Figure 63 - Proactive chatbot vs Extending chatbot

In this example, the proactive chatbot takes longer to interact with the student, as it uses four interactions per step, in contrast the extending chatbot, which uses two interactions. However, the last two student interactions are minimal. In this example, the two chatbots give the correct answer to the student, and those students who wanted only a simple answer would be more satisfied. However, we believe that most students would have a better idea of the problem with the proactive chatbot, as they will have more information about it through the video and the PS2CLH suggestions.

Below are the results of the tests performed with the proactive chatbot and the extending chatbot. The total number of questions was 100, and the number of correct question options given for the Chatbots was as below:

<i>Question</i>	<i>Proactive Chatbot</i>		<i>Extending Chatbot</i>	
	Simple	Complex	Simple	Complex
Type of Question:	Simple	Complex	Simple	Complex
What is a database?	Right	Wrong	Right	Wrong
How can I manage my time better?	Right	Right	Right	Right
What are the objectives of SQL?	Right	Wrong	Right	Wrong
What are the symptoms of stress?	Right	Right	Wrong	Wrong
Why is it important to learn how to manage stress?	Right	Right	Right	Right
What ways of trying to cope with stress are unhealthy?	Right	Right	Right	Wrong
Why is it important to connect with others to deal with stress?	Right	Wrong	Right	Wrong
What things should I make time for to deal with stress?	Right	Right	Right	Right
How can I manage my time better?	Right	Right	Right	Wrong
How common is stress?	Right	Right	Right	Right
Does stress affect your performance?	Right	Right	Right	Right
Why do people sometimes get frustrated?	Wrong	Wrong	Wrong	Wrong
Why do I keep putting things off?	Wrong	Wrong	Wrong	Wrong
What can I do to stop anxiety?	Right	Right	Wrong	Right
Why do I worry so much?	Right	Right	Right	Wrong
Why am I sometimes anxious?	Right	Right	Right	Right
Is it possible to deal with uncertainty?	Right	Wrong	Right	Wrong
How can I tell if others affect me?	Right	Right	Right	Right
How can I practice mindfulness?	Right	Right	Right	Right
Is it necessary to be creative?	Right	Wrong	Right	Wrong
How much sleep do I really need?	Right	Right	Right	Wrong

Table 6 - Results of a small sample of the questions used in the test

The formula for calculating Accuracy is:

$$Accuracy = \frac{\text{Correct \# of questions answered}}{\text{Total \# of questions made}} = 94/100 = 0.94 \approx 94\%$$

Equation 14 - Accuracy

Accuracy is the ratio of the number of answers correctly answered to the number of questions asked, and the chatbot test results show the percentage of accurate answers.

	<i>Proactive Chatbot</i>		<i>Extending Chatbot</i>	
Type:	<i>Simple</i>	<i>Complex</i>	<i>Simple</i>	<i>Complex</i>
Accurate Answers≈	94%	77%	91%	59%

Table 7 - Accuracy results

For both simple and complex questions, the proactive chatbot showed better results, i.e. a higher number of correct answers, than the extending chatbot. The difference was four more correct questions for simple answers and eighteen more for complex questions.

ANALYSIS: The experiment above was done in the following way: we asked the proactive chatbot the simple version of the question, then we asked the extending chatbot the same question; next we asked the proactive chatbot the complex question and then the extending chatbot the same complex question.

In terms of the big picture, we may say that we were influenced to a certain extent by the author's bias in relation to the difference between the simple and the complex versions of the question, as some words are shared between the two versions. However, the original questions and the complex questions are far apart, making it possible for the proactive chatbot to answer the complex version after saving the simple question correctly. Nevertheless, this attempts to replicate the students' situation, i.e. the fact that they share a typical language style and similar words.

5.7 Summary

This Chapter starts with an introduction and an outline of the fundamental relationship between the PS2CLH model and the proactive chatbot.

The proposed proactive chatbot framework for students uses a state-of-the-art Natural Language Processing - Deep learning question-and-answer technology. It is based on both the lecturer's subject and the author's student-controllable learning factor model which combines Psychology, Self-responsibility, Sociology, Communication, Learning and Health & Wellbeing (PS2CLH). This is where students, lecturers and university assistants have their profiles.

The framework is divided into two parts, the first of which contains the following components: Questions/Inputs, Data Preparation and Embedding, Word2Vec/TF-IDF/MV-LSTM and Vector/Matrix and text-similarity measurement. And the second part of the proposed framework, which is the Educational Chatbot Ecosystem.

This research reports on the outcomes of the employment of the innovative, proactive chatbot framework. The main novelty in the chatbot's framework is in the student-lecturer/assistant facilitator. After receiving the student question, the proactive chatbot suggests correlated controllable factors that influence the student's performance and improves the system's learning by saving the student's question and associating it with the answer it gives. Thus, there is no need to fine-tune or retrain the model to enhance the proactive chatbot's understanding of similar questions.

The next chapter concludes the thesis by presenting an evaluation of the research and a general conclusion, and makes suggestions for further research, .

Chapter 6 Evaluation, Conclusion and Future works

6.1 Evaluation

The research question was, “How can a specific model be developed to enhance academic performance which deals with the controllable academic factors? Furthermore, how can this model be used within a framework for implementing a readily available student learning assistant tool?”

Consequently, the primary purpose of this chapter is first to evaluate the methodology to establish the correlation among the PS2CLH model factors, and secondly to evaluate the design of the framework for a proactive chatbot for students.

6.1.1 Evaluation of the methodology to find the correlation among the PS2CLH model factors

The evaluation of this methodology will be done step by step, analysing each phase of the two experiments carried out in this research. We therefore start with the approval of the project by the university.

Project Approval by the University: One of the initial problems here was the lack of any official ethical approval document developed by the Universidade Católica de Angola, demonstrating that many developing countries do not have solid plans to develop scientific research in their universities. Given this situation, we had to adopt the ethical approval document from the London Metropolitan University UK, which the Universidade Católica de Angola then formally approved. More than an official document, this research felt a need for commitment from the University, which suggests that for future research a new commitment/ participation document for the University’s involvement is required. It is essential for the University to be part of the project because that will ensure that more attention will be paid to problems that arise during the research due to the nature of research, which will involve all the students and lecturers at the University. Despite the initial obstacles, the Universidade Católica de Angola provided all the necessary support for this research. With the approval and support of the University, we moved on to the second phase of the research, which is qualitative research.

Qualitative research to develop the survey: At this stage of the research, we were faced with another problem, which was the lack of scientific material developed in the area of Artificial Intelligence and Education in Angola. However, while there is a lack of investment by the governments of developing countries in scientific investigation, there is a lot of material on the subject in developed countries like England. Despite this state of affairs, interviews were conducted with students, university staff and lecturers, focusing on students' behaviour inside and outside the university. This qualitative research was fundamental for the first experiment. In the second experiment, we used the experience of the previous experiment. After this phase, we formulated the questions, looking at the results of the qualitative research.

Developing the questions for the PS2CLH model: In the first experiment, we had a great number of questions, and they were somewhat generic, resulting in incomplete student records. Students started to answer and then gave up because there were too many questions. This made us reduce the number of questions in the second experiment and make them more specific, paying more attention to the objective of each question and we were happy that more students were completing the questionnaire. After selecting the questions for the questionnaire, we started to build a web-based questionnaire.

Building the web-based questionnaire using Scrum Methodology: We developed the web-based questionnaire using Scrum methodology, which helped in the rapid development of the application. We used the python programming language in the Django 2.0 framework. The questions were Boolean with a 5-point Likert scale, and included a description explaining to the student what each factor meant and having some symptoms associated with them in ways that meant the students could still do the self-assessment if they had problems with that particular factor. . This approach was facilitative, as it gave students options relevant to what we intended to achieve as a result. Completing the development of the web-based questionnaire meant we had reached the stage where we could evaluate the collection of data from students.

Collection of Data from Students; Beginning the Quantitative Research: Carrying out two experiments in which data was collected from student massively helped this research because of the quality of the data collected. Uncertainty governed the first experiment because of the lack of scientific studies related to AI in Angola. In the first experiment, we collected 600 registers, with five areas and around 70 questions or factors. There were also many questions without a significant impact on the target variable, leading us to focus on the questions with

the most significant impact on the model. In the second experiment, we aimed to collect 500 registers. We started our second experiment having as a reference point the variables of the first experiment. Consequently, we reduced the variables with less impact on the prediction model and kept those with the highest scores for predictive importance. In addition, we added a new area, Health & Wellbeing and changed Self-Management to a 'Self-Responsibility' area. In the second experiment, we reduced the number of questions in each area, but we added a new area with a total of 53 questions. Again, we reduced the filling in time, and student interest increased with the reduced number of questions. After finishing the data collection, we moved on to data analysis and constructing the predictive model.

Start the CRISP-DM Methodology to Analyse the Data and Build the Model: We performed the data analysis using the CRISP-DM methodology on the database results from the student responses. We used the advanced statistics tool SPSS Modeler and R language. In 2018 and 2019, the SPSS Modeler software was the best there was for analysing statistics and applying data mining. The research goals were clear, which allowed it to turn into a data mining application. In both experiments, we were able to understand, extract, clean and process the data collected for use with data mining algorithms.

The data partition of the first experiment was: training 50%, test: 25% and validation: 25%. The best model was Random Trees 1, with 90.74% accuracy, and then XGBoost Tree1, with 85.6% accuracy. For the second experiment, the balanced partition split between results and percentage for the available data was: training 60%, testing 20% and validation 20%. We applied the SMOTE function to balance the lack of positive values. Then, the classification Logistics Regression function was used to build the model; the Python language result was around 94%. The best model from the SPSS Modeler results was the CRT model, which gave 94.12% accuracy. We had better results with the second experiment. We believe that one of the differentials was the significant factors chosen in relation to the target variable, probably because we had a better understanding of the factors affecting Angolan students' performance.

Predicting students' performance, Attributing importance, Correlation and Clusters: After creating the model, we added it to the framework application so that as soon as the student fills in the questionnaire, the system automatically generates a forecast of his academic result using the PS2CLH model as a reference. Both experiments presented an acceptable prediction model result, which showed the correlation among the factors. The most significant attributes were in

the areas of Learning and Self-Management, such as “Establish and achieve personal goals” and “Practice Tests”.

In the first and the second experiment, the correlation between the variables was determined. With the 3D visualisation in both experiments, we could see patterns showing different stages in the level of the students’ problems. The clusters were numbered from 1 to 7. We found that the best students were in the sixth and seventh clusters, which are the levels where the students selected have fewer problems with controllable factors. Finding the importance of attributes and establishing correlations among the variables allows us to add the proactivity function in the chatbot because it will suggest that student factors are related to what they are asking. It also shows on students’ profiles the variables they should prioritise. Below we evaluate the interventions.

Students’ Interventions: With the knowledge produced by data processing and data analysis, we see a clear direction for our interventions in the cluster. The interventions aim to assist students to reduce the number of factors affecting student performance overall, helping them overcome their limitations and develop new study habits to allow them to move to clusters 6 and 7. The experiment involved 50 students, of whom 25 selected students worked on some interventions with the researcher. This intervention showed that looking at the essential attributes and the correlation among the factors could assist many students because the model showed clear patterns which helped the researchers assist students.

We now ask why they were not doing what they said they could do. We realised that students had unrealistic expectations, and we had to redirect them to their primary goal. At this point, we can say that the high accurate level of the model prediction is not the main factor in this process: as long as there is a clear pattern and strong correlations among the factors selected, there will be a strong possibility of success. Having made the interventions, we move on to evaluating the results.

Evaluation of the Results: the evaluation of the results was done as a function of the students’ improvement concerning factors in the PS2CLH model. We saw how they improved their academic performance and how the factors controllable by the students contributed to this improvement.

Among the numerous machine learning models tested, the “Logistics Regression” model produced the best results. The prediction result presented by the model, based on the

perspectives of the proposed PS2CLH, was around 94% accurate. It showed that the selected variables from the proposed PS2CLH model correlate directly with the target variable or the student’s academic performance. The target variable was the students’ past performance, in terms of grade point average (GPA). (For more detail, see the predictor Importance variables in the Appendix section.) Here “studying and working” is one of the most significant predictors, followed by the variables “the average hours students play/distractions per day”, “Aim for excellence in everything you do”, “Establish and achieve personal goals”, “Practice Tests”, and more.

It explains the correlation between the time students spend on distractions such as social networking, spending hours with friends and more. The “Aim for excellence in everything you do” and “Establishing and achieving personal goals” variables were also found to be significant. Dealing well with these last variables proved to be game-changing for top students, which seems to be in harmony with the previous literature reviewed and ideas such as the marshmallow factor or delaying instant gratification and applying deliberate practice.

The Grade Point Average (GPA) is the score used to summarise students’ academic performance. According to (Singh, S. P., et al., 2016), a considerable number of researchers use the GPA to analyse students’ results. We applied the GPA to evaluate students’ performance in a semester (Tahir, S. & Naqvi, S. R., 2006).

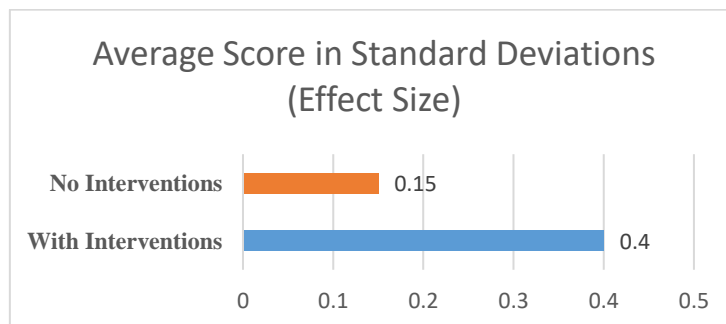


Figure 64 - Students’ average score in standard deviations

At the end of the semester, the raised awareness of PS2CLH perspective of these 25 students from the second experiment helped them set priorities and identify management factors affecting academic performance more effectively. Consequently, most of them have enhanced their academic performance by addressing these critical factors. Nevertheless, due to the limitations of the current sample data, the PS2CLH model will be further monitored for various applications.

As we have just seen, the experiment was carried out with a small number of subjects relative to the total number of students in a university. Therefore, we need an intervention that can be scalable and cost-effective for all students and to follow them throughout the academic year. It is proposed that the proactive chatbot framework will do this, and we will evaluate the method below.

6.1.2 Evaluation of the Method to Build the Framework Designed for a Proactive Chatbot for Students

The evaluation of this method will be done step by step, analysing each component of the framework. We start with the inputs.

Questions/Inputs, Data Preparation: The process starts by the chatbot receiving students' questions via writing or voice. This possibility of writing or speaking gives the student more possibilities to interact with the chatbot. However, voice recognition still has certain limitations, such as the difficulty of recognising some accents. The use of text and speech for the student's questions are strengths of the application. However, the Inputs writing and speech were designed only for regular students. The reality is that there are many universities worldwide with students with learning difficulties/disabilities who will not be able to use the application. Preparing the data helps in processing the question, but this process can often cause the loss of the certain words meaning, affecting question retrieval. However, in general we did not have too many problems dealing with data preparation. After this cleaning process, we used word embeddings.

Word Embedding and Vector/Matrix: Looking at the big picture of the research in question, we will see that there are six areas with several factors, and the proactive chatbot will also assist in different subjects. Considering the research context, we chose to use the TF-IDF, as it is the ideal choice for problems that include many words and large document files. It can still be applied to each training document at once, while word embedding using vectors must be applied to each word individually. Furthermore, the power of TF IDF lies in its simplicity. Therefore, it consumes less memory. One of the problems with TF IDF is that each word maps to a single value, making a sparse matrix which does not capture the meaning. However, we use the Transformer model in this framework that captures the word context.

Text Similarity Measurement: It is common to use TF IDF with cosine similarity, which gives better results. Again, referring to the context in which we will insert the proactive chatbot, it makes perfect sense to choose technologies that work more efficiently together. Moreover, it seems that cosine similarity is effective in multidimensional context and more straightforward implementations than other similarity measurements.

The Q&A and Transformer models: The choice of the transformer model for the framework was based on its self-attention of the stacked layers functionality that allows recognition of long-range dependencies in sequential data. This approach helps contextualise and find the correct answer to a specific question more quickly. It also deals with the vanish and explosion problems encountered in previous models during the training phase. However, the question-and-answer model of the transformer reveals the limitation in the size of the dataset, i.e. its limited number of paragraphs.

Assembly Parts: Even though the proactive chatbot presents options for possible questions, the student will have the option to confirm the desired question, thus breaking the natural rhythm of a conversation. However, students are used to search engines such as Google, Microsoft Bing and others. Therefore, we believe that this form of communication with the proactive chatbot will not be completely unfamiliar to students.

Interaction Facilitator: This component makes the proactive chatbot more usable and reliable in terms of answering questions. In addition, there are students (perhaps introverts) who feel more comfortable asking questions outside of the classroom environment. Therefore, this component gives greater prominence to the proactive chatbot.

Profile Customiser: We developed the pair “Extroverted (E) vs Introverted (I)”, but did not complete the development of the sixteen personalities types due to its complexity. However, the chatbot still acts according to students’ profiles; each profile is unique as it is based on the questionnaire the student has filled in. This is an exciting component in relation to the individualisation of student assistance. However, it could be developed in the future.

Multimodality: One of our programming problems was viewing videos in the Mozilla Firefox web browser. The other difficulty was finding images or videos for the different questions, which will be an additional job for lecturers or assistants. However, this component helps in the student’s learning experience.

Rating: This component is an effective way to get chatbot performance feedback from the students, but the reality is that few students will rate the chatbot's responses. Nevertheless, having the students' ratings helps get a sense of whether the student found the chatbot assistance helpful and thus allow us, lecturers and assistants, to improve the questions and answers in the knowledge base in the future.

Suggest Factor: We usually pay more attention to the problems that appear to be the biggest. However, the reality is that these are often just the result of other problems which appear to be harmless, for example the domino effect. Consequently, we end up dropping the biggest domino by moving what we can, i.e. the smallest one. This can be seen when students try to solve problems such as stress and fail, but if they solve another problem, such as the lack of a study plan that is causing the stress, the stress can disappear. Therefore, we see this component as being an innovation in the field of student assistance.

Knowledge Database: Even though we limit the number of questions associated with an answer, the knowledge base will inevitably grow rapidly, negatively affecting performance. However, the benefit of keeping the best ratings from students' questions will outweigh the potential problems.

Profiles: Generally, universities have applications with web profiles of students and lecturers. This component could be a connection point between existing university applications and the proactive chatbot framework, with the flexibility to integrate new components in different universities.

Comparison between the proactive chatbot and the extending chatbot: The table below presents a functional overview of what one chatbot does better than another.

	<i>Proactive Chatbot</i>	<i>Extending Chatbot</i>
Accuracy of answers	<input checked="" type="checkbox"/>	
Execution time of chatbot's response		<input checked="" type="checkbox"/>
Instant learning from previous questions	<input checked="" type="checkbox"/>	
Human-like conversation		<input checked="" type="checkbox"/>
Dealing with questions that are not in knowledge database	<input checked="" type="checkbox"/>	
Multiple ways of displaying content	<input checked="" type="checkbox"/>	
Proactivity	<input checked="" type="checkbox"/>	
Accepts student feedback	<input checked="" type="checkbox"/>	
Q&A in relatively large data set	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Assists students	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Table 8 - Comparison between the proactive chatbot and the extending chatbot

As we can see, the two chatbots assist the student, both using a decent number of questions in their data set. However, the proactive chatbot has greater accuracy. On the other hand, the extending chatbot gives a faster response to the student's questions. However, one of the reasons for this are the procedures that make the proactive chatbot learn from past questions and deal with unexpected situations where the questions asked by the student are not found in the knowledge database. It uses multimodality in its responses yet receives feedback from students and proactively acts by suggesting other factors correlated with the question asked by the student. All of this makes the conversation less natural between the proactive chatbot and the student. The key observation was that, despite the limited real-world simulation with students, the proactive chatbot learned and used more information in giving a response for each interaction. For example, creating a cluster of similar questions related to a particular chatbot answer led us to suggest that it improved accuracy as the students interacted with the chatbot.

Despite encouraging results, the proposed proactive chatbot needs to be further monitored and tested in different universities and countries.

6.2 Limitations

The first limitation was the researcher's lack of background knowledge of the multidisciplinary field, specifically of psychology, sociology, communication, learning and health and well-being. In addition, during the collection phase, a problem was the fact that the Universidade Católica de Angola did not have an ethical approval document, demonstrating that many developing countries do not have solid procedures for developing scientific research in their universities. There was also a lack of scientific material produced in the area of Artificial Intelligence and education in Angola.

The other limitation was the lack of an efficient way of evaluating students' controllable factors on the web questionnaire. To process the data, we used a data mining tool which presented some limitations in explaining the procedures and how we got a particular result.

There are still limitations in understanding the clusters of students' behaviour using the 3D visual representation. The other problem is that we used interventions for only 25 students. Lastly, we did not test the proactive chatbot framework in a real-life university environment with students; therefore, we were not able to investigate the real-world effects of the proactive chatbot.

6.3 Implications

Creating the PS2CLH model develops an awareness of the current controllable factors that affect students' performance, both for students and for university managers, giving them a basis for taking proactive actions and intervening accordingly. The PS2CLH model can potentially become a reference for future studies related to students' controllable factors. In addition, the 3D representation of the students' factors using the PS2CLH model is a viable tool for resolving a concern for university managers, subject tutors and academic mentors. This representation helps to show, measure, analyse and monitor student performance against controllable factors affecting their academic achievement and thus to enhance the student experience. Finding the current connection of particular university student behaviours and lifestyles, and particularly the controllable factors, opens a new possibility for effective interventions for different clusters of students. Finally, the proactive chatbot framework has the potential to facilitate and enhance

the way lecturers or mentors interact with students, improving the student's learning experience, and for learning from this interaction and increasing the usability of the chatbot.

6.4 Conclusion

The aim was to develop a specific model to enhance academic performance that incorporates all the controllable academic factors and further integrate it into a framework for the implementation of a readily available and student-usable tool, in which we used Natural Language Process - Deep Learning (NLP-DL).

We started by investigating the literature to build the PS2CLH model that combines the perspectives of Psychology, Self-responsibility, Sociology, Communication, Learning and Health & Wellbeing to facilitate a student-controllable learning factor model. Then, we built a methodology to find the correlation among the PS2CLH model factors. Next, we developed a method to build the framework designed for the proactive chatbot for students. Finally, we built an application that incorporates a proactive chatbot that could potentially assist students.

During the first phase of the research, two experiments were conducted to test the practicability of the PS2CLH model. According to the results, students improved their awareness of PS2CLH perspectives on factors influencing their achievement and which are under their control. The PS2CLH model helped them to control how much influence the factors have on them. It empowered the students.

Although we still face the challenge of assisting all the students in their academic work for the duration of their programmes, new solutions to tackle this issue are emerging. This research proposed an AI deep learning chatbot using the state-of-the-art model Transformer combined with text similarity measurement and other components, resulting in a proactive chatbot framework to assist students on the PS2CLH factors and their academic subjects individually. This approach improved the accuracy of the proactive chatbot. It could instantly learn in interaction with the student.

However, the proposed framework still has limitations. For instance, there is no image or video for every question-answer on our knowledge database to fully exploit a multimodal environment. In our case, we just represented the pair Introverted – Extroverted on the chatbot persona due to the complexity of doing anything more sophisticated. In addition, future

framework testing is required. The best way of doing this would be to test how it works over a semester at different universities. Despite advances in Artificial Intelligence, we still have a long way to go to achieve the initial claim that AI could achieve the level of human rationality level or become a technology that perfectly replaces human interaction.

To conclude this Thesis, the metaphor of planting a garden is presented below, referencing James C. Hunter's book "The Servant". It sees a parallel between assisting a student and creating a healthy environment for a plant. In our case, the proactive chatbot framework intends to create a healthy environment for students to improve their academic performance.

From Hunter's perspective, to shape a healthy garden, it is first necessary to find a piece of land exposed to the sun and turn over the soil to get it ready for planting. Then the seeds can be planted, they are regularly watered, and the soil is fertilised, taking care of pests, controlling the temperature and wild planting the garden every so often. He declares that we will see some growth in due time, and soon the fruit will come, (Hunter, 2012). This is how we created the PS2CLH model to identify each student's controllable factors which affect much of their performance. We then built students' profiles, using the proactive chatbot to assist them periodically in relation to the PS2CLH factors through which they can control their behaviour and lifestyle, and to assist them in their academic subjects. The framework will be used as a facilitator to help lecturers and specialists engage with, involve and assist students. We can expect to see students' growth after some time. However, assistance is any communication that influences choice. While this research proposes that the proactive chatbot framework can provide the necessary 'friction' or influence, students should make their own choices to improve. Bearing in mind the garden rules, growth does not occur because of us. We can deliver a healthy environment and provide the knowledge students need to question their behaviours, and then they can choose to change and improve their academic performance.

6.5 Future works

For future work in this area, we recommend that data collection be appropriately developed with student tests and scientific diagnostics, finding a scientific way of evaluating students' academic problems instead of a self-evaluation questionnaire. The other recommendation for future work would be to individualise and personalise the proactive chatbot based on the 16 personality types in ways that can involve students and captivate them in their interaction with the chatbot. In addition, work need to be done to find a more efficient integration of text similarity measurements and the questions and answers model and to improve the structure of the knowledge database.

In terms of further research, we suggest incorporating parallelism in the text similarity measurement algorithm for each question with multiply answers. This could identify the highest score for each question and would thus help to improve machine learning on-time processing. Furthermore, the chatbot could learn the student's languages and ways to ask more efficiently. This work could speed up the chatbot's process as the knowledge database is filled with more and more information. We also suggest future research on classification in the PS2CLH model and lecturers' subjects. The idea could be to use classification to identify the area of students' questions more precisely when they ask a question.

References

- Metiri Group, 2008. *Multimodal Learning Through Media*., s.l.: Charles Fadel, Global Lead, Education; Cisco Systems, Inc.
- Abdullah, A., 2005. Some determinants of student performance in Financial Management Introductory course: an empirical investigation. *Journal of King Saudi University Administrative Sciences*, 5(1), pp. 1-26..
- Acholonu, I. & Njie, S., 2020. *African students' access to higher education is a priority for the continent's development*. Firoz Lalji, Africa and Higher Education.
- Adam, M., Wessel, M. & Benlian, A., 2020. *AI-based chatbots in customer service and their effects on user compliance*.. s.l., Electron Markets.
- Adamopoulou, E. & Moussiades, L., 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2(100006).
- Akama, E., 2017. Student's Performance Prediction Using FP-Tree Data Mining Techniques. *International Journal of Science and Research (IJSR)*.
- al., D. M. e., 2002. Information Visualisation for Site. *Planning Technical Report of Data Mining, INVISIP IST-2000-29640, WP No2: Technology Analysis*, 28 February.
- Almada, A., Yu, Q. & Patel, P., 2019. *PS2CLH: A Learning Factor Model for Enhancing Students' Ability to Control Their Achievement*. Tokyo, ACE2019.
- Almeida, F. & Xexéo, G., 2019. *Word Embeddings: A Survey*. s.l., Cornell University.
- Ameisen, E., 2018. *How to solve 90% of NLP problems: a step-by-step guide*. [Online] Available at: <https://blog.insightdatascience.com/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e> [Accessed 20 12 2020].
- Andoni, A., Indyk, P. & Krauthgamer, R., 2008. Earth mover distance over high-dimensional spaces. *In Proceedings of the Symposium on Discrete Algorithms*, 20–22 January (San Francisco, CA, USA), p. pp. 343–352.
- Applegate, C. & Anne, D., 2006. The Impact of Paid Work on the Academic Performance of Students: A Case Study from the University of Canberra. *Australian Journal of Education*, 50(2).
- Bandler, A., Roberti, A. & Fitzpatrick, O., 2014. *How to take charge of your life – The user's guide to NLP*. ISBN -978-0-00-755593-2 ed. London: HarperCollins Publishers.
- Bankston, C. L. & Caldas, S. J., 1998. Family structure, schoolmates, and racial inequalities in school achievement.. *Journal of Marriage and the Family*, 60(3), pp. 715-723.

- Becker, W. E. & Watts, M. C., 1996. A national survey on teaching undergraduate economics. *American Economic Review*, 86(2), p. 448–453.
- Bengio, Y., Courville, A. & Vincent, P., 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), p. 1798–1828.
- Benotti, L., Martínez, M. & Schapachnik, F., 2014. Engaging high school students using chatbots. *Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education, ITiCSE*, p. 63–68.
- Bertolini, K., Stremmel, A. & Thorngren, J., 2012. *STUDENT ACHIEVEMENT FACTORS*. [Online]
Available at: <https://files.eric.ed.gov/fulltext/ED568687.pdf>,
[Accessed December 2019].
- Bird, J., Ekárt, A. & Faria, D., 2020. *Chatbot Interaction with Artificial Intelligence: Human Data Augmentation with T5 and Language Transformer Ensemble for Text Classification*, Birmingham: Aston University, United Kingdom.
- Blei, D.M., Ng, A.Y. & Jordan, M.I., 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 2003, Volume 3, p. 993–1022.
- Blog, R., 2016. *AlphaGo: Mastering the ancient game of Go with Machine Learning*. s.l., Google Research Blog.
- Bouissac, P., 2007. *Oxford Reference*. Print ISBN-13: 9780195120905 ed. Oxford: Oxford University Press.
- Bowles, S. & Gintis, H., 1976. *Schooling in capitalist America*. New York, Basic Books.
- Brandtzaeg, P. B. & Følstad, A., 2017. Why people use chatbots. *International Conference on Internet Science*, p. 377 – 392.
- Briggs, M., 1980. *Gifts Differing: Understanding Personality Type*, s.l.: s.n.
- British Library, 2015. *Qualitative and Quantitative Research*. [Online]
Available at: <http://www.bl.uk/bipc/resmark/qualquantresearch/qualquantresearch.html>
[Accessed 27 August 2015].
- Bush, F., 1945. *As We May Think*, s.l.: Atlantic July.
- Chauhan, A., 2017. *What Is Variable Importance and How Is It Calculated?*. [Online]
Available at: <https://dzone.com/articles/variable-importance-and-how-it-is-calculated>
[Accessed 17 August 2022].

- Chemers, M., Hu, L.-t. & Garcia, B., 2001. Academic Self-Efficacy and First-Year College Student Performance and Adjustment. *Journal of Educational Psychology*, Vol. 93(No. 1), pp. 55-64.
- Chen, X., Jia, S. & Xiang, Y., 2020. A review: Knowledge reasoning over knowledge graph. *Expert Syst. Appl.* 2020, 141(112948).
- Clarizia, F., et al., 2018. Chatbot: An education support system for student. *International symposium on cyberspace safety and security, Springer (2018)*, pp. 291-302.
- Clark, D., 2018. *The Fallacy of 'Robot' Teachers*. [Online]
Available at: <http://donaldclarkplanb.blogspot.com/2018/04/the-fallacy-of-robot-teachers.html>
[Accessed 10 April 2018].
- Clutterbuck, D. & Megginson, D., 1999. *Mentoring Executives and Directors*. p. 3.
- Codecademyteam, 2021. *What Is a Framework? - CODECADEMY TEAM*. [Online]
Available at: <https://www.codecademy.com/resources/blog/what-is-a-framework/>
[Accessed 29 9 2021].
- Committee on the Science of Children Birth to Age 8: Deepening and Broadening the Foundation for Success, 2015. *Transforming the Workforce for Children Birth Through Age 8: A Unifying Foundation*. [Online]
Available at: <https://www.ncbi.nlm.nih.gov/books/NBK310550/>
[Accessed 14 June 2019].
- Connolly, T., Begg, C. & Holowczak, R., 2010. *Business Database Systems*. 1 ed. s.l.:s.n.
- Connor, M. & Pokora, J., 2007. *Coaching at work*. Maidenhead: Open University Press.
- Connor, M. P. & Pokora, J. B., 2007. *Coaching at School*. Maidenhead: Open University Press.
- Csikszentmihalyi, M., 1997. *Finding Flow – State Diagram*. [Art].
- Csikszentmihalyi, M., 2008. *Flow: The Psychology of Optimal Experience (Harper Perennial Modern Classics)*. ISBN 0061339202; ed. s.l.:HarperCollins - Paperback..
- Cuevas, J., 2015. *Is learning styles-based instruction effective? A comprehensive analysis of recent research on learning styles..* s.l., Theory Res. Educ. 13:308–333. doi: 10.1177/1477878515606621.
- Cummings, A. M., 2014. *The Impact of Student Support Services on Academic Success at a Select Historically Black College and University*. [Online]
Available at: <https://digitalcommons.unf.edu/etd/532>
[Accessed 10 8 2021].

- Cunningham-Nelson, S., Boles, W., Trouton, L. & Margerison, E., 2019. A review of chatbots in education: Practical steps forward. *30th annual conference for the australasian association for engineering education (AAEE 2019): Educators becoming agents of change: Innovate, integrate*, p. 299–306.
- Damerau, F.J., 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM 1964*, Volume 7, p. pp. 171–176.
- De Angeli, A., Lynch, P. & Johnson, G., 2001. Personifying the e-market: a framework for social agents. In: M. Hirose, ed. *IFIP TC.13 International Conference on Human-Computer Interaction*. Amsterdam: IOS Press, p. 198–205.
- De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D.L., 2000. The mahalanobis distance. *Chemom. Intell. Lab. Syst 2000*, Issue [https://doi.org/10.1016/S0169-7439\(99\)00047-7](https://doi.org/10.1016/S0169-7439(99)00047-7), p. 1–18.
- De Martino et al., 2002. *Information Visualisation for Site*. INVISIP IST-2000-29640 WP No2: Technology Analysis ed. New York: Planning Technical Report of Data Mining.
- Deerwester, S., et al., 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci. 1990*, Volume 41, p. 391–407.
- Department of Health, 2001. *Research Governance Framework for Health and Social Care*, London: Department of Health.
- Department of Health, 2005. *Research Governance Framework for Health and Social Care*. 2nd ed. London,: Department of Health.
- Devlin, J., Chang, M. & Toutanova, K., 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, s.l.: Google AI Language.
- Devlin, J., Chang, M.W., Lee, K. & Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv 2018*, Volume arXiv:1810.04805.
- Deza, M. & Deza, E., 2009. Encyclopedia of distances. In *Encyclopedia of Distances*. Springer: Berlin/Heidelberg, Issue Germany, p. pp. 1–583.
- Dice, L.R., 1945. *Measures of the amount of ecologic association between species*. [Online] Available at: <https://doi.org/10.2307/1932409> [Accessed 21 9 2021].
- D'Silva, G., et al., 2020. Career counselling chatbot using cognitive science and artificial intelligence. *Advanced computing technologies and applications, Springer (2020)*, pp. 1-9.

- Dsouza, R., Sahu, S., Patil, R. & Kalbande, D. R., 2019. Chat with bots intelligently: A critical review & analysis. *IEEE - In 2019 international conference on advances in computing*, p. 1–6.
- Duckworth, A., 2017. *Grit: Why passion and resilience are the secrets to success*. ISBN 978-1-5011-1110-5 ed. New York: NY 10020.
- Duhigg, C. & Ruben, G., 2014. *The Power of Habit: Why We Do What We Do in Life and Business*. ISBN - 10: 1847946240; ed. s.l.:Paperback.
- Dunlosky, J., Rawson, K. A., Marsh, E. J. & Nat, M. J., 2013. Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, 14(1), p. 4 –58.
- Elken, M., Hovdhaugen, E. & Wiers-Jenssen, J., 2015. *Higher Education in the Nordic Countries*. Denmark, The Nordic Council of Ministers.
- Ellison, G., 2013. *Ethics Approval for Projects based overseas*. [Online]
Available at: <http://student.londonmet.ac.uk/media/london-metropolitan-university/london-met-documents/professional-service-departments/r>
[Accessed 10 12 2017].
- Ellison, G., 2013. *Informed Consent*. [Online]
Available at: <http://student.londonmet.ac.uk/media/london-metropolitan-university/london-met-documents/professional-service-departments/research-office/course-doc>
[Accessed 2017 11 2017].
- Ericsson, A. & Pool, R., 2016. *Peak: Secrets from the New Science of Expertise*. ISBN 978-0-5-45623-5 ed. New York: New York 10016.
- Fayyad, U., Shapiro, G. P. & Smyth, P., 1996.. From data mining to knowledge discovery in databases.. *AI Magazine*, 17(3), pp. 37-54.
- Fenta, A. & Kelkay, B., 2018. The impact of entertainment related factors that affect the academic performance of graduating class students. *International Journal of Applied Research*, 4(10)(Abaya campus, Arba Minch University,), pp. 258-265.
- Ferriss, T., 2017. *Tools of Titans – The tactics, routines, and habits of billionaires, icons, and world class performers*. ISBN – 978-1-328-68378-6 ed. New York: Library of Congress Cataloging,.
- Flexibility, Behavioural, 2012. *Do Something Different - What is behavioural flexibility and why does it matter?.* [Online]
Available at: <https://dtd.me/business/2012/06/15/what-is-behavioural-flexibility-and-why-does-it-matter/>
[Accessed 2 May 2017].

- Følstad, A. & Brandtzæg, P. B., 2017. Chatbots and the new world of HCI. *Interactions*, 24(4), p. 38–42.
- Gardner, H., 1983. In: *Frames of Mind: Theory of Multiple Intelligences*. New York: Basic Books, pp. 10-12.
- Gardner, H., 1991. *The Unschooled Mind: How Children Think and How Schools Should Teach*. New York: Basic Books.
- George, T., & Merkus, J., 2022. *Explanatory Research | Definition, Guide & Examples..* [Online]
Available at: <https://www.scribbr.co.uk/research-methods/explanatory-research-design/>
[Accessed 20 October 2022].
- George, T., 2022. *Exploratory Research | Definition, Guide, & Examples..* [Online]
Available at: <https://www.scribbr.co.uk/research-methods/exploratory-research-design/>
[Accessed 20 October 2022].
- Ghose, S. & Barua, J., 2013. *Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor*. In: *International Conference on Informatics, Electronics & Vision (ICIEV)*. s.l., IIIE, p. 1–5.
- Gilmer, J., et al., 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, 70(Sydney, Australia, 6–11 August 2017), p. 1263–1272.
- Goffe, W. L. & Sosin, K., 2005. Teaching with technology: May you live in interesting times. *Journal of Economic Education*, 36(3), p. 278–291.
- Goodfellow, I., Bengio, Y. & Courville, A., 2017. *Deep Learning*. [Online]
Available at: <http://neuralnetworksanddeeplearning.com/about.html>
[Accessed 14 March 2017].
- Google Cloud, 2021. *Google Cloud - Speech-to-Text*. [Online]
Available at: <https://cloud.google.com/speech-to-text>
[Accessed 29 9 2021].
- Grossman, D.A. & Frieder, O., 2012. *Information Retrieval: Algorithms and Heuristics*. Springer Science & Business, 15(Berlin/Heidelberg, Germany, 2012).
- Guess, A., 2011. *Sentiment Analysis v. Semantic Analysis*. [Online]
Available at: <http://www.dataversity.net/sentiment-analysis-v-semantic-analysis/>
[Accessed 27 November 2016].
- Gunawan, D., Sembirin, C. & Budiman, M., 2018. *The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents*. s.l., Journal of Physics Conference Series 978(1):012120.

- Hattie, J., 2009. *Visible Learning, A synthesis of over 800 meta-analyses relating achievement*. New York: Routledge, 270 Madison Avenue.
- Hattie, J., 2018. “*Visible Learning plus,*” *250+ Influences on Student Achievement*. [Online] Available at: https://www.visiblelearningplus.com/content/research-john-hattie_250_influences_10.1.2018.pdf [Accessed 7 December 2018].
- Hedjazi, Y. & Omid, M., 2008. *Factors Affecting the Academic Success of Agricultural Students at University of Tehran, Iran*. Tehran, J. Agric. Sci. Technol.
- Hicks, M. & Foster, S., 2010. *Adapting Scrum to Managing a Research Group*. [Online] Available at: <http://www.cs.umd.edu/~mwh/papers/score.pdf> [Accessed 20 February 2017].
- Hill, B., 2020. *Internal Locus of Control (and Why It’s Important to Success)*. [Online] Available at: <https://businessprofessionals.com/internal-locus-of-control/> [Accessed 17 08 2022].
- Hinton, G. et al., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *Signal Processing Magazine*.
- Ho, C. C., Lee H. L., Lo, W. K. & Lui, H. F., 2018. Developing a chatbot for college student programme advisement. *2018 international symposium on educational technology (ISET) - IEEE*, p. 52–56.
- Hobert, S., 2019. Say hello to ‘coding tutor’! design and evaluation of a chatbot-based learning system supporting students to learn to program. *Fortieth International Conference on Information Systems At: Munich*.
- Holmes, W., Bialik, M. & Fadel, C., 2019. *Artificial Intelligence in Education*. Boston: Center for Curriculum Redesign.
- Hoose, N., 2021. *Information Processing Theory*. [Online] Available at: <https://courses.lumenlearning.com/edpsy/chapter/information-processing-theories/> [Accessed 24 9 2021].
- Hoy, B., 2018. *Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants*. Rochester, Medical Reference Services Quarterly.
- Hu, B., Lu, Z., Li, H. & Chen, Q., 2014. Convolutional neural network architectures for matching natural language sentences. *In Proceedings of the Advances in Neural Information Processing Systems*, Issue Montreal, QC, Canada, 8–13 December 2014, p. 2042–2050.
- Huang, F., 2021. Japanese doctoral students’ stress: Main findings from a national survey in 2017. *International Journal of Chinese Education*, 10(1).

- Huang, P.S., et al., 2013. Learning deep structured semantic models for web search using clickthrough data. *In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, Issue Burlingame, CA, USA, 27 October–1 November 2013, p. 2333–2338.
- Hunter, C. J., 2012. THE SERVANT – A simple story about the true essence of Leadership. *Crown Business*, p. 130–132.
- IBM Software Business Analytics, 2010. *CRISP-DM 1.0: Step-by-step data mining guide*, s.l.: IBM Corporation.
- ICX Association blog, 2016. *ICX Association blog*. [Online]
Available at: <https://icxa.org/2016/05/chatbots-friendly-or-frightening/>
[Accessed 16 February 2018].
- Informatik, F., Hochreiter, S. & Schmidhuber, J., 1997. Long Short-term Memory. *Neural Computation*, 9(8), pp. 1735-1780.
- Irving, R.W. & Fraser, C.B., 1992. Two algorithms for the longest common subsequence of three (or more) strings. *In Proceedings of the Annual Symposium on Combinatorial Pattern Matching*, Issue Tucson, AZ, USA, 29 April–1 May 1992, p. pp. 214–229.
- Isaksen, V. J., 2013. *Popular social science - The Psychology behind Defence Mechanisms*. [Online]
Available at: <http://www.popularsocialscience.com/2013/02/07/the-psychology-behind-defense-mechanisms/>
[Accessed 14 April 2017].
- Ismail, M. & Ade-Ibijola, A., 2019. Lecturer's apprentice: A chatbot for assisting novice programmers. *2019 international multidisciplinary information technology and engineering conference (IMITEC) - IEEE*, pp. 1-8.
- Jaccard, P., 1912. The distribution of the flora in the alpine zone. [https://doi.org/10.1111/j.1469-8137.1912.tb05611.x\(1](https://doi.org/10.1111/j.1469-8137.1912.tb05611.x(1). *New Phytol* 1912), p. pp. 37–50.
- Jennifer, G., 2013. Millennials Enter Growing, Controversial Field of Life Coaching. *USA Today*.
- Jung, C. G., Adler, Gerhard & Hull, R. F.C, 2014. *Collected Works of C.G. Jung, Volume 6: Psychological Types*, Princeton: Princeton University Press.
- Jung, C., 1971. Psychological Types. In: R. & K. Paul, ed. *Collected Works of C.G. Jung, Vol. 6*. London: s.n.
- Jurafsky, D. & Martin, H., 2020. *Speech and Language Processing - N-gram Language Models*. [Online]

Available at: <https://web.stanford.edu/~jurafsky/slp3/3.pdf>
[Accessed 25 January 2021].

Kaku, M., 2018. *Why You Should Be Optimistic About the Future | Michio Kaku on Impact Theory*. [Online]
Available at: <https://www.youtube.com/watch?v=tGulK44YaOM>
[Accessed 10 December 2018].

Khan, A., 2020. *Why is Education Industry opting for AI Chatbots? How Are They Benefiting It? - Botsify*. [Online]
Available at: <https://botsify.com/blog/education-industry-chatbot/>
[Accessed 28 9 2021].

Khazan, O., 2018. *The Atlantic - The Myth of 'Learning Styles'*. [Online]
Available at: <https://www.theatlantic.com/science/archive/2018/04/the-myth-of-learning-styles/557687/>
[Accessed 28 9 2021].

Kingma, P. & Ba, L., 2017. *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION*. s.l., ICLR arXiv:1412.6980v9.

Kitaev, N., Kaiser, Ł. & Levskaya, A., 2020. Reformer: The Efficient Transformer. *ICLR 2020, Machine Learning (cs.LG)*, 18 February.

Kontostathis, A. & Pottenger, W.M., 2006. A framework for understanding Latent Semantic Indexing (LSI) performance.. *Inf. Process. Manag.* 2006, Volume 42, p. 56–73.

Krizhevsky, A., Sutskever, I. & Hinton, G., 2017. *ImageNet Classification with Deep Convolutional Neural Networks*. Nevada, Lake Tahoe.

Kuligowska, K., 2015. Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research*, 2(2), pp. 1-16.

Kullback, S. & Leibler, R.A., 1951. *On information and sufficiency*. [Online]
Available at: <https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Sufficiency/10.1214/aoms/1177729694.full>
[Accessed 21 9 2021].

Kusner, M., Sun, Y, Kolkin, N. & Weinberger, K., 2015. From word embeddings to document distances. *In Proceedings of the International Conference on Machine Learning*, Issue Lille, France 6–11 July 2015, p. pp. 957–966.

Laiqa, R., Shah, R. U. & Khan, S. M., 2011. Impact of quality space on students' academic achievement. *International Journal of Academic Research*, pp. 706-711.

- Lakhiani, V., 2016. *The code of the extraordinary mind – Ten Unconventional laws to redefine your life & succeed on your own terms*. ISBN- 13: 978-1-62336-708-4 ed. New York: Rodale Inc..
- Landauer, T.K. & Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 1997, pp. 104, 211.
- Landauer, T.K., Foltz, P.W. & Laham, D., 1998. An introduction to latent semantic analysis. *Discourse Process* 1998, Volume 25, p. 259–284.
- Landry, S. H., 2014. *The role of parents in early childhood learning*, Texas: Children's Learning Institute, University of Texas Health Science Center.
- Lawson, B. R., 2001. *The Language of Space*. Boston, The Architectural Press.
- Le, Q. & Mikolov, T., 2014. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning. Issue Beijing, China, 22–24 June 2014, p. pp. 1188–1196.
- Lesli, G. M. & Ingrid, S., 2013. *The impact of non-cognitive skills on outcomes for young people*. London, Institute of Education, University of London.
- Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl*, Volume 10, p. pp. 707–710.
- Li, Q., Wang, B. & Melucci, M., 2019. CNM: An Interpretable Complex-valued Network for Matching. *arXiv 2019*, Volume arXiv:1904.05298.
- Lima, I. R., Freire, T. C. & Costa, H. A. X., 2012. *Adapting and Using Scrum in a Software Research and Development Laboratory*. [Online]
Available at: http://www.fsma.edu.br/si/edicao9/FSMA_SI_2012_1_Princip
- Liu, Z., Xiong, C., Sun, M. & Liu, Z., 2018. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. *arXiv 2018*, Volume arXiv:1805.07591.
- LLC, O. P., 2019. *developgoodhabits*. [Online]
Available at: <https://www.developgoodhabits.com/learning-new-things/>
[Accessed 20 Junho 2020].
- Mamlin, N., Harris, . K. R. & Case, L. P., 2001. A Methodological Analysis of Research on Locus of Control and Learning Disabilities - Rethinking a Common Assumption. *Journal of Special Education*, pp. 214-225.

- Mamlin, N., Harris, K. R. & Case, L. P., 2001. A Methodological Analysis of Research on Locus of Control and Learning Disabilities - Rethinking a Common Assumption.. *Journal of Special Education*, pp. 214-225.
- Manning, C.D. & Schütze, H., 1999. Foundations of Statistical Natural Language Processing. *MIT Press*, Issue Cambridge, MA, USA,.
- May, C., 2018. *The Problem with "Learning Styles" (Scientific American)*. [Online] Available at: <https://www.scientificamerican.com/article/the-problem-with-learning-styles/> [Accessed 27 9 2021].
- McCarthy, J., 1959. *Programs with Common Sense*. London, In Proceedings of the Teddington Conference on the Mechanization of Thought Processes, 756-91.
- McCorduck, P., 2004. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. 2nd ed. s.l.:A. K. Peters, Ltd.
- Meisel, W., 2016. *Specialized Digital Assistants and Bots:Vendor Guide and Market Study*. Tarzana CA: TMA Associates.
- Memon, A. R., et al., 2015. An electronic information desk system for information dissemination in educational institutions. In 2nd International Conference on Computing for Sustainable Global Develop. *IEEE*, p. 1275–1280.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J., n.d. *arXiv 2013*, Volume arXiv:1301.3781.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. *Cornell University and Google Scholar*, 2(arXiv:1301.3781).
- Miller, Suzanne M. & McVee, Mary B, 2013. *Multimodal Composing in Classrooms*. doi:10.4324/9780203804032 ed. s.l.:ISBN 9780203804032.
- Mischel, W., 2015. *The Marshmallow Test: Understanding Self-control and How To Master It*. ISBN 978055216886 ed. London: Penguin Random House UK.
- Mitra, S. & Dangwal, R., 2010. Limits to self-organising systems of learning: the Kalikuppam experiment. *British Journal of Educational Technology*, 41(5).
- Morgan, S., 2012. Should a Life Coach Have a Life First?. *The New York Times*.
- Morgan, S., 2012. Should a Life Coach Have a Life First?. *The New York Times*.
- Morshed, J., 2016. *How chatbots will change the face of campus technology*. [Online] Available at: <https://www.ecampusnews.com/campus-administration/chatbots-campus-students/> [Accessed 10 June 2019].

- Mulkey, L. M., Crain, . R. L. & Harrington, A. J., 1992. One-parent households and achievement: economic and behavioural explanations of a small effect. *Sociology of Education*, Volume 65, pp. 48-65.
- Müller, V. C. & Bostrom, N., 2016. Future progress in artificial intelligence: A survey of expert opinion. In: V. C. Müller, ed. *Fundamental issues of artificial intelligence*. Berlin: Springer, p. 555–572.
- Murphy, M., 2015. 'Everything you need to know about deep learning and neural networks'. [Online]
Available at: <http://www.techworld.com/big-data/why-does-google-need-deep-neural-network->
[Accessed July 2015].
- Murtarelli, G., Gregory, A. & Romenti, S., 2021. A conversation-based perspective for shaping ethical human–machine interactions: The challenge of chatbots. *Journal of Business Research*, Issue 129, p. 927–935.
- Neto, J., 2021. *Analytics Vidhya - Best NLP Algorithms to get Document Similarity*. [Online]
Available at: <https://medium.com/analytics-vidhya/best-nlp-algorithms-to-get-document-similarity-a5559244b23b>
[Accessed 29 9 2021].
- Nielsen, F., 2010. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv*, Volume arXiv:1009.4004.
- Nietzsche, F. W., 2014. *Schopenhauer As Educator: Friedrich Nietzsche's Third Untimely Meditation*. Chicago, CreateSpace Independent Publishing Platform.
- Noble, J. P., Roberts, W. L. & Sawyer, R. L., 2006. *Student Achievement, Behavior, Perceptions, and Other Factors Affecting ACT Scores*, s.l.: ACT Research Report Series, 2006-1.
- Norouzi, M., Fleet, D.J. & Salakhutdinov, R.R., 2012. Hamming distance metric learning. In Proceedings of. *Lake Tahoe, NV, USA*, Volume 3–6, p. pp. 1061–1069.
- O'Brien, E., 2014. *10 Things Life Coaches Won't Tell You*, s.l.: MarketWatch.
- Okonkwo, C. W. & Ade-Ibijola, A., 2021. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2(100033).
- Ondas, S., Pleva, M. & Hladek, D., 2019. How chatbots can be involved in the education process. *IEEE - In In 2019 17th international conference on emerging eLearning technologies and applications (ICETA)*, p. 575–580.
- Pang, L., et al., 2016. Text matching as image recognition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Issue Phoenix, AZ, USA, 12–17 February 2016.

- Parsloe, E., 1999. *The Manager as Coach and Mentor*. p. 8.
- Pascanu, Gulcehre, C., Cho, K. & Bengio, Y., 2014. *How to construct deep recurrent neural networks*. s.l., Proceedings of the Second International Conference on Learning Representations (ICLR 2014).
- Paschoal, L. N., de Oliveira, M. M. & Chicon, P. M. M., 2018. A chatterbot sensitive to student's context to help on software engineering education.. *2018 XLIV Latin American computer conference (CLEI) - IEEE*, pp. 839-848.
- Pennington, J., Socher, R. & Manning, C.D., 2014. Glove: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Issue Doha, Qatar, 25–29 October 2014, p. pp. 1532–1543.
- Percy, A., 2014. *Student mental health: the situation is more nuanced than it seems*, Oxford: The Guardian.
- Piatetski, G. & Frawley, W., 1991. *Knowledge Discovery in Databases*. ISBN 0262660709 ed. MA: MIT Press Cambridge.
- Prebble, T., et al., 2004. *Impact of Student Support Services and Academic Development Programmes on Student Outcomes in Undergraduate Tertiary Study: A Synthesis of the Research Report to the Ministry of Education*, s.l.: Massey University College of Education.
- Proctor Gallagher Institute, 2015. *You were born rich*. ISBN - 9781599303673 ed. Scottsdale: MacCrary Publishing.
- Prof Sims, J., 2019. *The Teaching and Learning International Survey (TALIS)*, London: UK, Department of Education; UCL, Institute of Education.
- Profillidis, V.A. & Botzoris, G.N., 2018. *Modeling of Transport Demand, Analyzing, Calculating, and Forecasting Transport Demand*. 1st Edition ed. s.l.:Paperback ISBN: 9780128115138.
- Ranoliya, B. R., Raghuwanshi, N. & Singh, S., 2017. *Chatbot for University Related FAQs*. *In: 2017 International Conference on Advances in Computing*. Udupi, Communications and Informatics (ICACCI).
- Rapp, A., Curti, L. & Boldi, A., 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text- based chatbots. *International Journal of Human-Computer Studies*, Issue 102630.
- Research & Enterprise Development Centre, 2014. *An introduction to ethics issues and principles in research involving human and animal participants*, s.l.: Canterbury Christ Church University.

Research Blog, 2016. *AlphaGo: Mastering the ancient game of Go with Machine Learning*. s.l., Google Research Blog.

Richardson, J., 2008. *A Life of Picasso: The Triumphant Years 1917-1932*. s.l.:Knopf Doubleday Publishing Group.

Robbins Seminar, 2016. *Total Success 2016 Tony Robbins Highlights*. [Online] Available at: <https://www.youtube.com/watch?v=DW0H8sN6iQM> [Accessed 14 December 2016].

Robbins, A., 2000. *Unlimited Power - The new science of personal achievement*. ISBN-13 9780671316457 ed. United States: Simon & Schuster Ltd.

Robbins, M., 2017. *The 5 Second Rule – Transform your life, work, and confidence with everyday courage*. ISBN – 978-1-68261-238-5 ed. s.l.:SAVIO Repvblc.

Robertson, S.E. & Walker, S., 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the International ACM Sigir Conference on Research and Development in Information Retrieval SIGIR '94*, Issue Dublin, Ireland, 3–6 July 1994, p. pp. 232–241.

Rohn, J., 1999. *The Why Behind Personal Development!*. [Online] Available at: <https://www.youtube.com/watch?v=5Pj5lekpVEE> [Accessed 20 December 2016].

Roman, M., Shahid, A., Khan, S & Koubaa, A., 2021. Citation Intent Classification Using Word Embedding. *IEEE*, 9(10.1109/ACCESS.2021.3050547), pp. 9982-9995.

Rong, X., 2014. Word2vec parameter learning explained. arXiv:1411.2738(arXiv 2014).

Rotter, J., 1966. Generalized expectancies for internal versus external control of reinforcements'. *Psychological Monographs*, Volume 80, p. 609.

Rowley, J., 2007. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, 33(2), p. 163–180.

Rubin, G., 2015. *Better Than Before: Mastering the Habits of Our Everyday Lives*. s.l.:Hardcover.

Russell, J. & Norvig, P., 2009. *Artificial Intelligence: A Modern Approach*. 3rd ed. New Jersey: Prentice Hall.

Sahami, M. & Heilman, T.D., 2006. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on WorldWideWeb*, Issue Edinburgh, Scotland, UK, 23–26 May 2006, p. 377–386.

Salton, G. & Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* 1988, Volume 24, p. pp. 513–523.

Sandu, N. & Gide, E., 2019. Adoption of ai-chatbots to enhance student learning experience in higher education in India. *2019 18th international conference on information technology based higher education and training (ITHET) - IEEE*, pp. 1-5.

Sandu, N. & Gide, E., 2020. *Adoption of AI-Chatbots to Enhance Student Learning Experience in Higher Education in India*. Sydney, Australia, IOP Conference Series Materials Science and Engineering.

Sansonnet, J. P., Leray, D. & Martin, J., 2006. Architecture of a Framework for Generic Assisting Conversational Agents. In: J. Gratch, et al. eds. *Intelligent Virtual Agents. IVA 2006. Lecture Notes in Computer Science*. Heidelberg: Springer, pp. 145-156.

Sansonnet, J, Leray, D. & Martin, JC, 2006. *Architecture of a Framework for Generic Assisting Conversational Agents*. Marina Del Rey, CA, USA, Intelligent Virtual Agents, 6th International Conference.

Schuchmann, S., 2019. *Analyzing the Prospect of an Approaching AI Winter*. s.l., s.n.

Schumaker, R. P., Ginsburg, M., Chen, H. & Liu, Y., 2007. An evaluation of the chat and knowledge delivery components of a low-level dialog system: The AZ-ALICE experiment. *Decision Support Systems*, 42(4), p. 2236–2246.

Schwaber, K. & Beedle, M., 2001. *Agile Software Development with SCRUM*. 1st ed. s.l.:Paperback.

Schwaber, K. & Beedle, M., 2001. *Agile Software Development with SCRUM*. 1st ed. s.l.:Paperback.

Scrum Methodology, 2013. *Scrum Methodology & Agile Scrum Methodologies*. [Online] Available at: <http://scrummethodology.com/> [Accessed 27 August 2015].

Seldon, A., 2018. *The Fourth Education Revolution: Will Artificial Intelligence Liberate or Infantilise Humanity*. Buckingham: University of Buckingham Press.

Shang, L., Lu, Z. & Li, H., 2015. *Neural Responding Machine for Short-Text Conversation "*, *Sha Tin, Hong Kong. Paper.*, Hong Kong: Noah's Ark Lab..

Sharkey, C., 2016. Should we welcome robot teachers?. *Ethics and Information Technology*, 18(4), p. 283–297.

Shawar, B. A. & Atwell, E., 2007. Chatbots: Are they Really Useful?. *LDV Forum*, 22(1), p. 29–49.

Shen, Y., et al., 2014. A latent semantic model with convolutional-pooling structure for information retrieval. *In Proceedings of the 23rd ACM International Conference on*

Information and Knowledge Management, Issue Shanghai, China, 3–7 November 2014, p. 101–110.

Shoppe, R., 2019. *Effect of the Breakthrough Student Assistance Program on Grades, Behavior, and Attendance*, s.l.: Walden University.

Singh, A., 2012. *Does Physical Activity Lead to Higher Grades? Education Report*. Amsterdam, .: VU University Medical Center.

Singh, S. P., Singh, P. & Malik, S., 2016. *Research Paper Factors Affecting Academic Performance of Students*. India, Gurukul Kangri Uni-versity, Haridwar, Uttarakhand .

Sivic, J., 2009. Efficient visual search of videos cast as text retrieval. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 31(4), p. 591–605.

SQuAD, 2020. *SQuAD2.0 - The Stanford Question Answering Dataset*. [Online] Available at: <https://rajpurkar.github.io/SQuAD-explorer/> [Accessed 17 2 2020].

Srivastava, B., 2021. Did chatbots miss their “Apollo Moment”? Potential, gaps, and lessons from using collaboration assistants during COVID-19. *Patterns - AI Institute, University of South Carolina, 1112 Greene St., Columbia, SC 29208, USA*, 2(8).

Srivastava, N, et al., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, pp. 1929-1958.

Statistics Solutions, 2020. *Pearson’s Correlation Coefficient*. [Online] Available at: <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/> [Accessed 1 November 2022].

Tahir, S. & Naqvi, S. R., 2006. FACTORS AFFECTING STUDENTS’ PERFORMANCE. *Bangladesh e-Journal of Sociology*, 3(1).

Tesla, 2017. *WHAT IS GPU-ACCELERATED COMPUTING?*. [Online] Available at: <http://www.nvidia.com/object/what-is-gpu-computing.html> [Accessed 14 March 2017].

The University of Kansas, 2021. *Different Learning Styles—What Teachers Need To Know*. [Online] Available at: <https://educationonline.ku.edu/community/learning-styles-what-teachers-need-to-know> [Accessed 25 9 2021].

Three Initiates, 2010. *The Kybalion – Mind of the all Hermetic classics*. Toronto: Prohyptikon Publishing Inc.

- TimesMojo, 2022. *How Does LSTM Solve Exploding Gradient?*. [Online]
Available at: <https://www.timesmojo.com/how-does-lstm-solve-exploding-gradient/>
[Accessed 20 08 2022].
- Tinto, V., 1975. Dropout from Higher Education: A Theoretical Synthesis of Recent Research.. *Review of Educational Research*, vol. 45(JSTOR), p. 89–125.
- Tugend, A., 2015. Before Starting as a Coach, It Helps to Go Into Training. *New York Times*.
- Turing, A., 1950. Computing Machinery and Intelligence. *Mind*, LIX(236), p. 433–460.
- UNESCO, 1997. Chapter VI, Section VII. Institutional Rights, Duties and Responsibilities. In: *Recommendation concerning the Status of Higher-Education Teaching Personnel*. Paris: UNESCO.
- United Nations Children’s Fund, UNICEF, 2011. Progress Evaluation of UNICEF’s Education in Emergencies and Post-Crisis Transition,. *Programme: Angola Case Study, Evaluation Report*, p. 22.
- University of Kansas, 2021. *Different Learning Styles—What Teachers Need To Know*. [Online]
Available at: <https://educationonline.ku.edu/community/learning-styles-what-teachers-need-to-know>
[Accessed 10 8 2021].
- Vashishth, S., Yadati, N. & Talukdar, P., 2020. Graph-based Deep Learning in Natural Language Processing. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, Issue Hyderabad, India, 5–7 January 2020, p. 371–372.
- Vaswani, A., et al., 2017. *Attention Is All You Need*. s.l., Cornell University.
- Vaswani, A., et al., 2017. *Attention Is All You Need*,. s.l., Cornell University.
- Wallace, R. S., 2009. The anatomy of ALICE. In: R. Epstein, G. Roberts & G. Beber, eds. *Parsing the Turing Test*. Dordrecht: Springer, p. 181–210.
- Wan, S., et al., 2016. A deep architecture for semantic matching with multiple positional sentence representations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Issue Phoenix, AZ, USA, 12–17 February 2016.
- Wang, S. & Manning, C.D., 2012. Baselines and bigrams: Simple, good sentiment and topic classification.. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*,. 2(Short Papers, Jeju Island, Korea, 8–14 July 2012), p. pp. 90–94.
- Wang, D. et al., 2014. High-Dimensional Data Stream Classification via Sparse Online Learning. *IEEE International Conference on Data Mining*, pp. 1007-1012.

- Weizenbaum, J., 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), p. 36–45.
- Weng, L., 2019. From GAN to WGAN. arXiv:1904.08994 (arXiv 2019).
- Williams, M. & Burden, R. L., 1997. *Psychology for language teachers: a social constructivist view*. New York: Cambridge.
- Wilson, D. T., 2015. Redirect: Changing the Stories We Live. In: s.l.:Paperback, pp. 30 - 35.
- Wilson, T., 2013. *Redirect: Changing the Stories We Live By*. ISBN 978-0-141-04224-4, ed. London: Penguin Group.
- Winkler, W.E., 1990. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. [Online]
Available at: <https://files.eric.ed.gov/fulltext/ED325505.pdf>
[Accessed 20 9 2021].
- Wu, L., et al., 2018. Word mover's embedding: From word2vec to document embedding. arXiv:1811.01713(arXiv 2018).
- Wu, Z., et al., 2020. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, Issue <https://ieeexplore.ieee.org/document/9046288>, pp. 4 - 24.
- Xu, A., et al., 2017. A new chatbot for customer service on social media.. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, May, pp. 3506-3510.
- Yang, Z. & Hu, Z., 2017. *On Unifying Deep Generative Models*, s.l.: Carnegie Mellon University.
- Yousry, M., 2011. Discover Your Hidden Memory & Find the Real You. In: *TJ International*. Cornwall: Hay house UK Ltd, pp. 62 - 65.
- Zhou, J., et al., 2018. Graph neural networks: A review of methods and applications. *arXiv 2018*, Issue arXiv:1812.08434.
- Zhuohao, W., Dong, W. & Qing, L., 2021. Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF. *Chinese Journal of Electronics*, 30(4), pp. 652-657.

Appendix

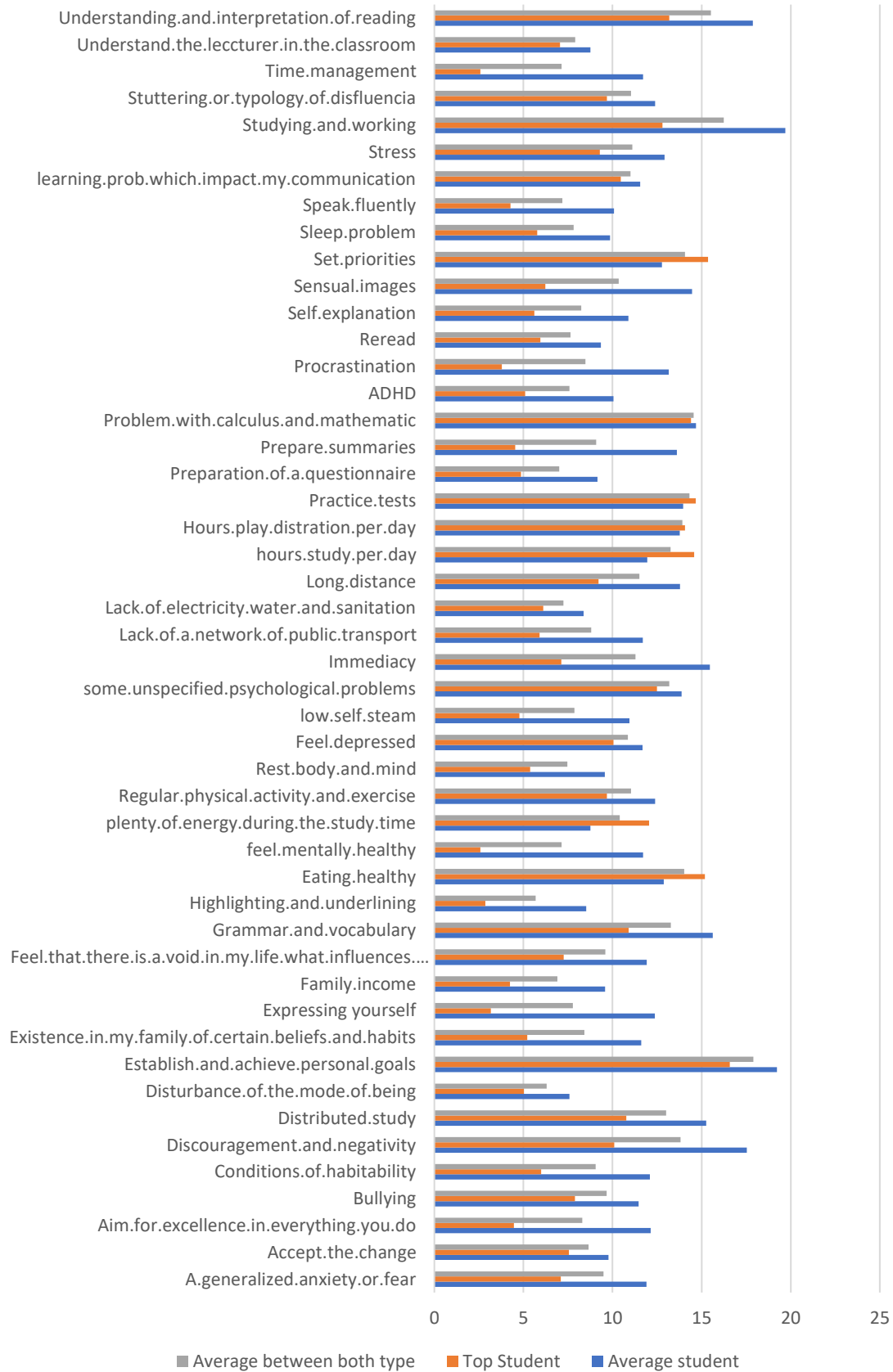
Travel risk assessment

TRAVEL RISK ASSESSMENT FORM		LONDON METROPOLITAN UNIVERSITY	
TEMPLATE ONLY – PLEASE EDIT FOR EACH SPECIFIC TRIP.			
Risk Assessment For		Assessment Undertaken By	
Service / School: Universidade Católica de Angola		Name: Arlindo Djassi Diogo de Almada	
Destination: Angola-Luanda		Date: 02/05/2018	
Purpose of trip: Data Collection		Signed by Dean of School / equivalent or nominee:	
		Date:	
		Assessment Reviewed	
		Name: <i>This section to be used if this risk assessment is to be used for further identical trips</i>	
		Date:	
		This risk assessment must be reviewed against latest FCO travel advice prior to travel	

Note: Not all of the hazards or controls listed below will be relevant to your intended travel - delete as appropriate

List significant hazards here:	List existing controls, or refer to safety procedures etc.	For risks, which are not adequately controlled, list actions needed.	Remaining level of risk: high, med or low
Air travel to (IRAQ) <i>Long haul flight - DVT / Dehydration</i>	Traveller advised to follow all DVT / dehydration precautions advised by aircraft cabin crew. Procedural Guidance on Travel Related Deep Vein Thrombosis (DVT) Specific safety advice from FCO to be included where relevant		
Accommodation <i>Fire, personal security</i>			

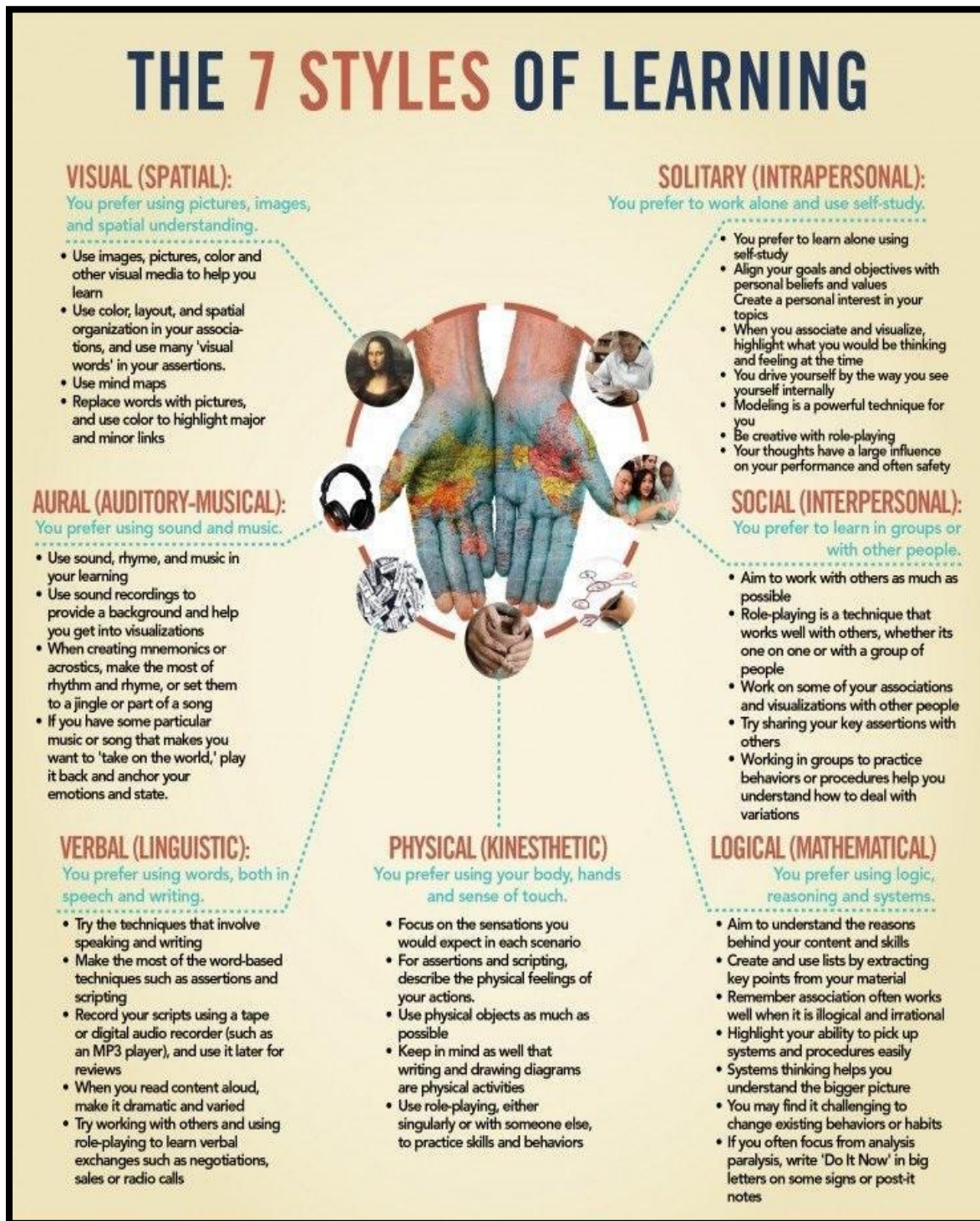
Feature importance



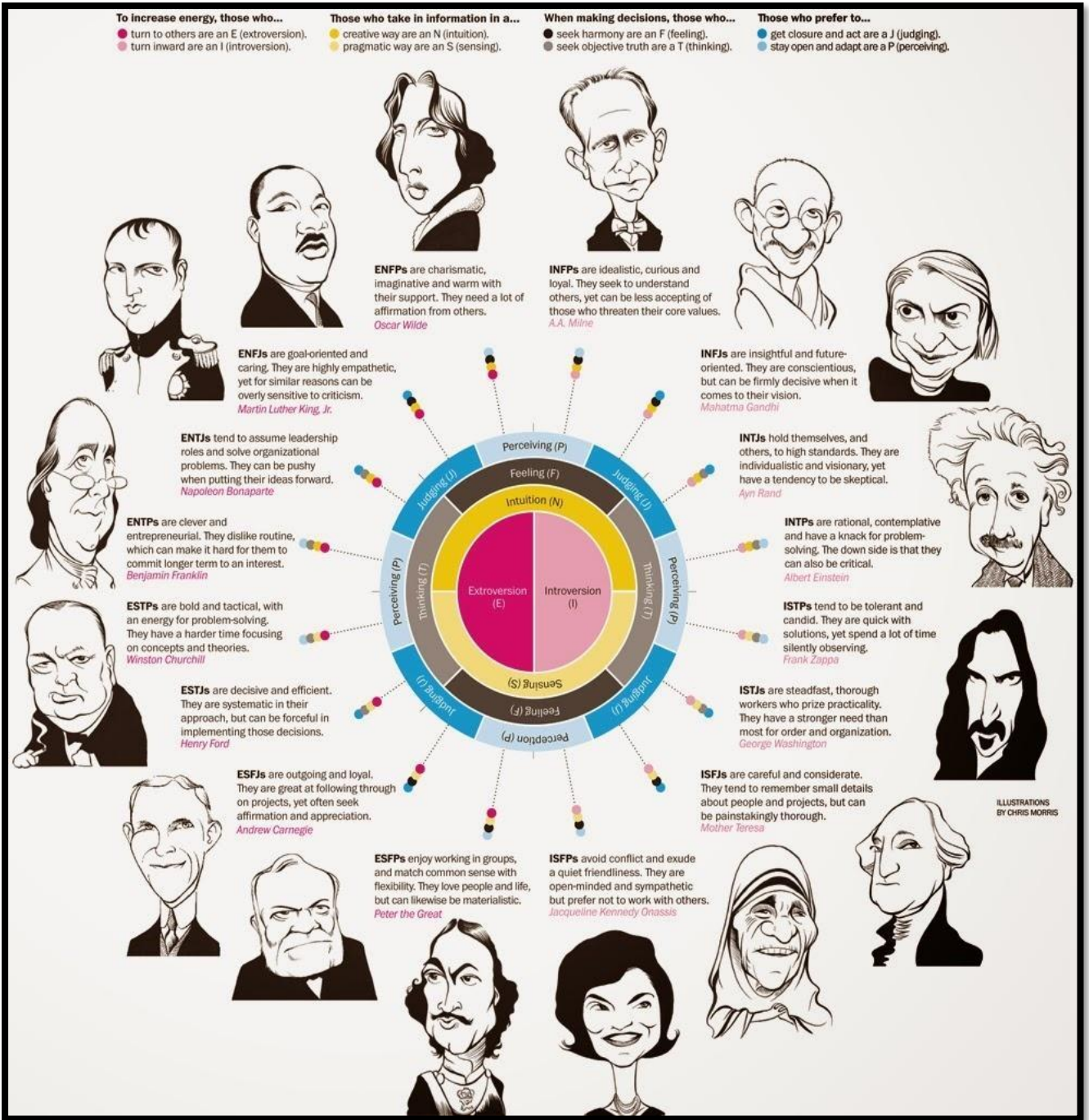
Feature	Regular Student	Top Student	Average between both type
A.generalized.anxiety.or.fear	11.900842	7.093103	9.4969725
Accept.the.change	9.771442	7.545632	8.658537
Aim.for.excellence.in.everything.you.do	12.133342	4.470418	8.30188
Bullying	11.461367	7.889721	9.675544
Conditions.of.habitability	12.09755	5.994017	9.0457835
Discouragement.and.negativity	17.536234	10.100659	13.8184465
Distributed.study	15.254397	10.774459	13.014428
Disturbance.of.the.mode.of.being	7.587777	5.027322	6.3075495
Establish.and.achieve.personal.goals	19.230261	16.579838	17.9050495
Existence.in.my.family.of.certain.beliefs.and.habits	11.612639	5.217129	8.414884
Expressing.yourself	12.373958	3.164849	7.7694035
Family.income	9.577273	4.242995	6.910134
Feel.that.there.is.a.void.in.my.life.what.influences.my.hours.of.study	11.919715	7.257566	9.5886405
Grammar.and.vocabulary	15.630944	10.901848	13.266396
Highlighting.and.underlining	8.518984	2.862623	5.6908035
Eating.healthy	14.87112	13.184391	14.0277555
feel.mentally.healthy	11.711269	2.581557	7.146413
plenty.of.energy.during.the.study.time	13.753262	12.056725	12.9049935
Regular.physical.activity.and.exercise	12.396418	9.68412	11.040269
Rest.body.and.mind	9.560582	5.377362	7.468972
Feel.depressed	11.682132	10.044437	10.8632845
low.self.steam	10.942398	4.775179	7.8587885
some.unspecified.psychological.problems	13.881609	12.499861	13.190735
Immediacy	15.461153	7.125493	11.293323
Lack.of.a.network.of.public.transport	11.697352	5.905367	8.8013595
Lack.of.electricity.water.and.sanitation	8.374959	6.109184	7.2420715
Long.distance	13.792366	9.212987	11.5026765
hours.study.per.day	15.942935	12.576001	14.259468
Hours.play.distratation.per.day	15.771161	14.066528	14.9188445
Practice.tests	14.96721	13.665361	14.3162855
Preparation.of.a.questionnaire	9.153478	4.86045	7.006964
Prepare.summaries	13.613933	4.54459	9.0792615
Problem.with.calculus.and.mathematic	14.684981	14.409741	14.547361
Procrastination	13.161259	3.788622	8.4749405
Reread	9.345112	5.942175	7.6436435
Self.explanation	10.884025	5.61431	8.2491675
Sensual.images	14.462227	6.230684	10.3464555
Set.priorities	15.765833	12.362556	14.0641945
Sleep.problem	9.853957	5.778763	7.81636
Speak.fluently	10.087846	4.27585	7.181848
learning.prob.which.impact.my.communication	11.546819	10.459346	11.0030825
Stress	12.912676	9.29386	11.103268

Studying and working	18.698665	12.798235	15.74845
Stuttering or typology of disfluencia	12.396418	9.68412	11.040269
Time management	11.711269	2.581557	7.146413
Understand the lecturer in the classroom	8.753262	7.056725	7.9049935
Understanding and interpretation of reading	17.87112	13.184391	15.5277555

Variables' prediction importance



The seven styles of learning (Gardner) (LLC, 2019)



16th Psychological Types (Jung, C.G.)

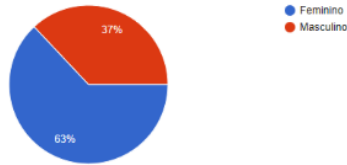
Universidade – País

229 responses

- Angola
- Universidade Católica de Angola
- Universidade Católica de Angola
- Universidade Católica de Angola - Angola
- ISPTEC - Angola
- Angola
- Universidade católica de Angola
- ISPTEC - Angola
- Instituto superior politécnico de tecnologias e ciências (ISPTEC) - Angola - Luanda

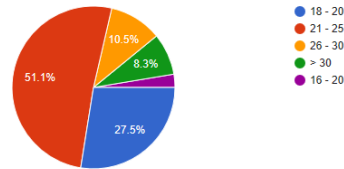
Género

227 responses



Idade

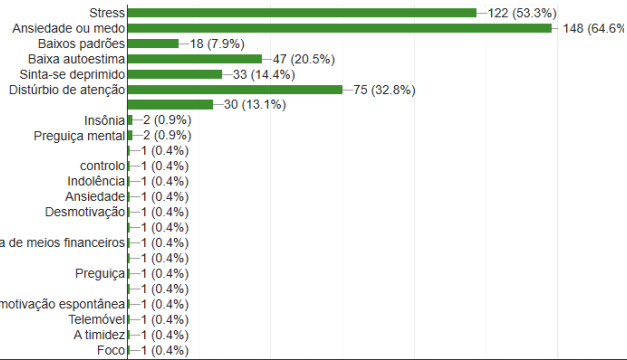
229 responses



Os fatores internos e externos afetam negativamente o desempenho académico do aluno. Por favor, responda às seguintes questões da forma mais completa e honesta possível.

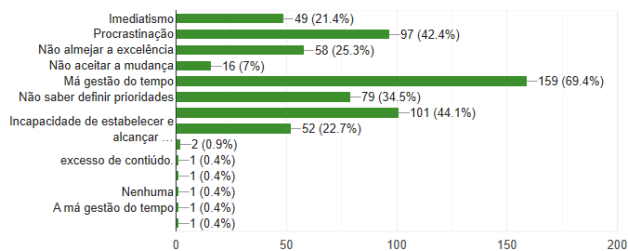
1 - Quais são os fatores psicológicos que acredita que afetam o seu desempenho académico?

229 responses



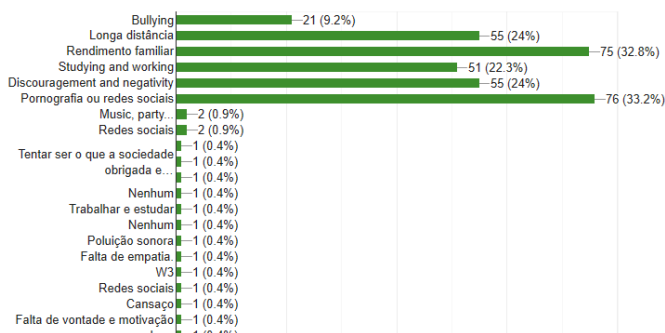
2 - Quais são os fatores de autorresponsabilidade que acredita que afetam o seu desempenho acadêmico?

229 responses



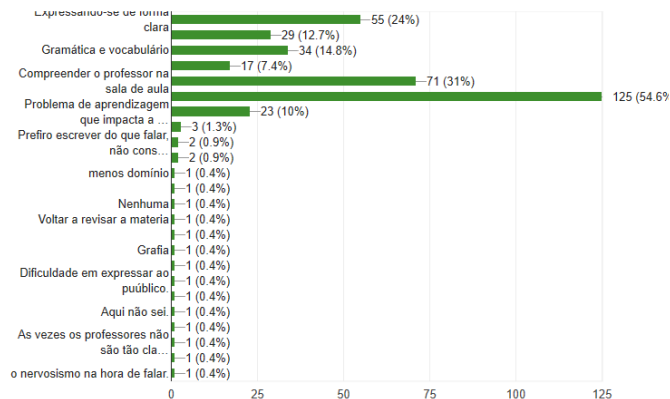
3 - Quais são os fatores sociológicos que acredita que afetam o seu desempenho acadêmico?

229 responses



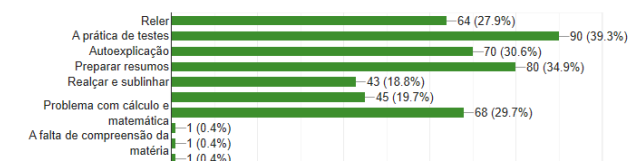
4 - Quais são os fatores de Comunicação que acredita que afetam o seu desempenho acadêmico?

229 responses



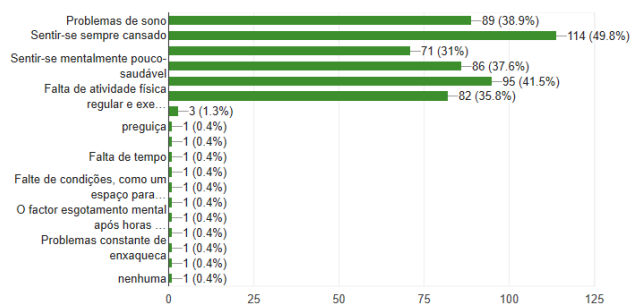
5 - Quais são os fatores de Aprendizagem que acredita que afetam o seu desempenho acadêmico?

229 responses



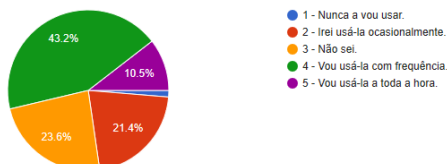
6 - Quais são os fatores de Saúde & bem-estar que acredita que afetam o seu desempenho acadêmico?

229 responses



7 - Quão útil encontraria uma ferramenta automatizada que poderia ajudá-lo a gerir os fatores controláveis que afetam o seu desempenho? Além disso, ajuda-o a lidar com material académico recomendado pelo seu docente.

229 responses



49 responses

Accepting responses

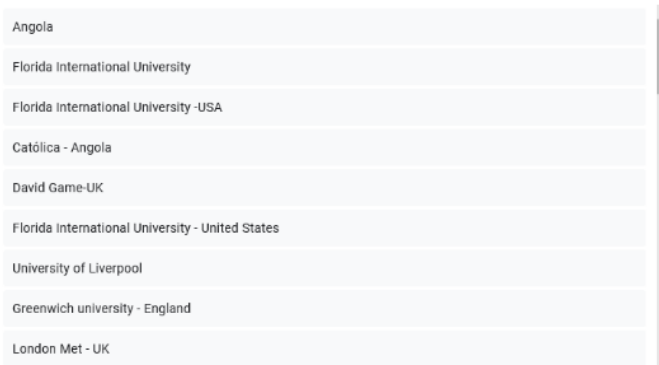
Summary

Question

Individual

University - Country

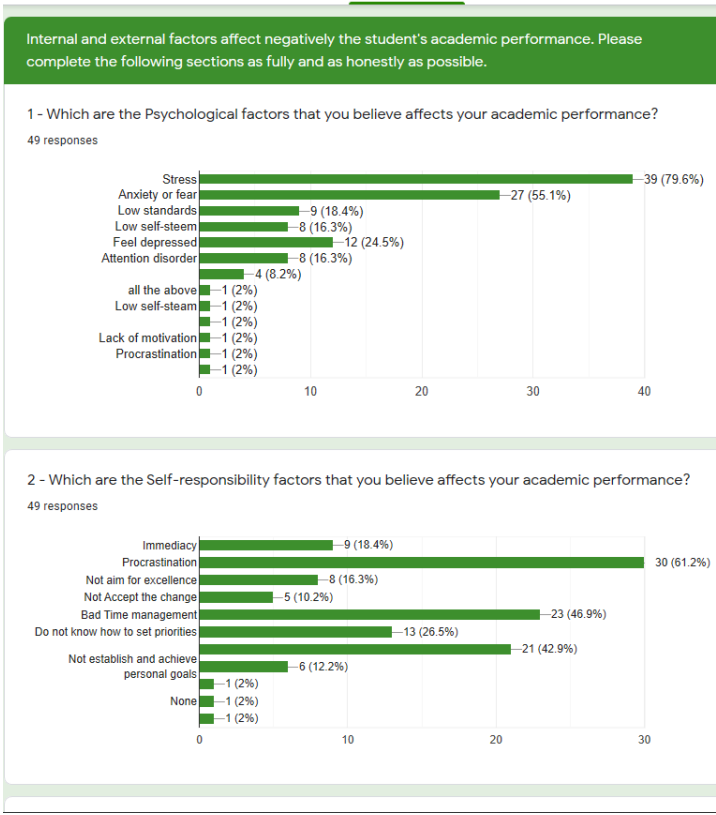
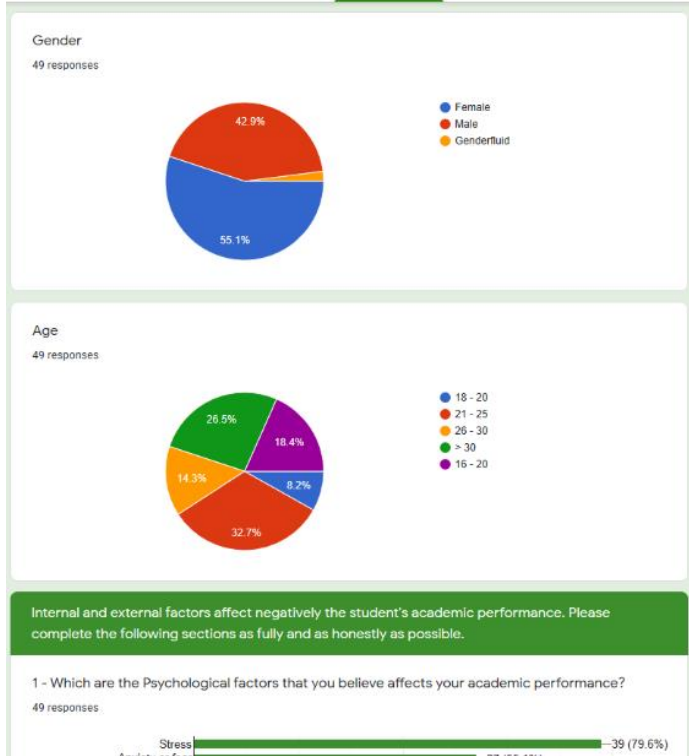
49 responses



Gender

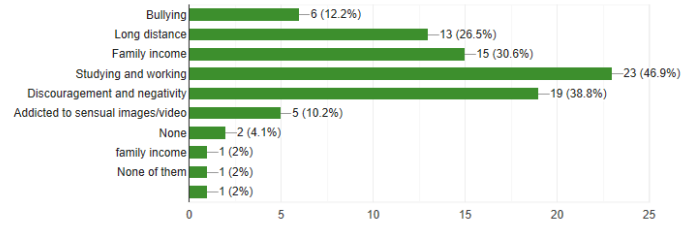
49 responses





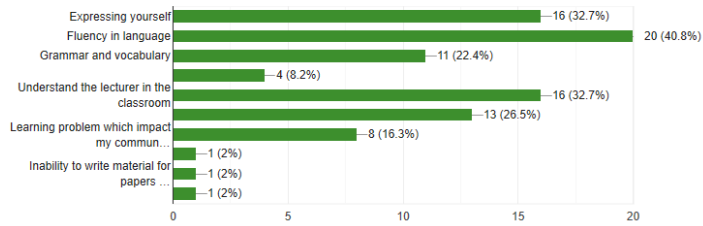
3 - Which are the Sociological factors that you believe affects your academic performance?

49 responses



4 - Which are the Communication factors that you believe affects your academic performance?

49 responses

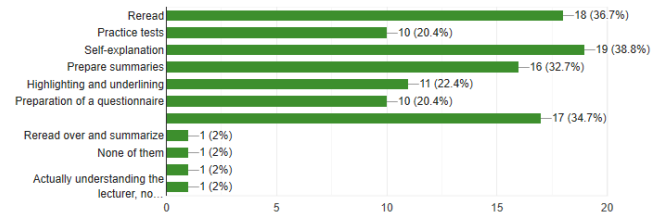


5 - Which are the Learning factors that you believe affects your academic performance?

49 responses

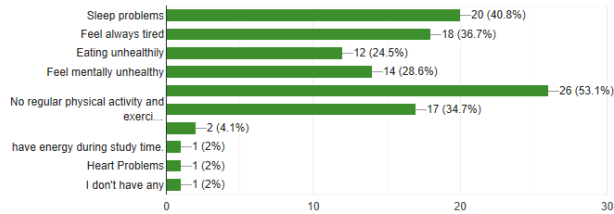
5 - Which are the Learning factors that you believe affects your academic performance?

49 responses

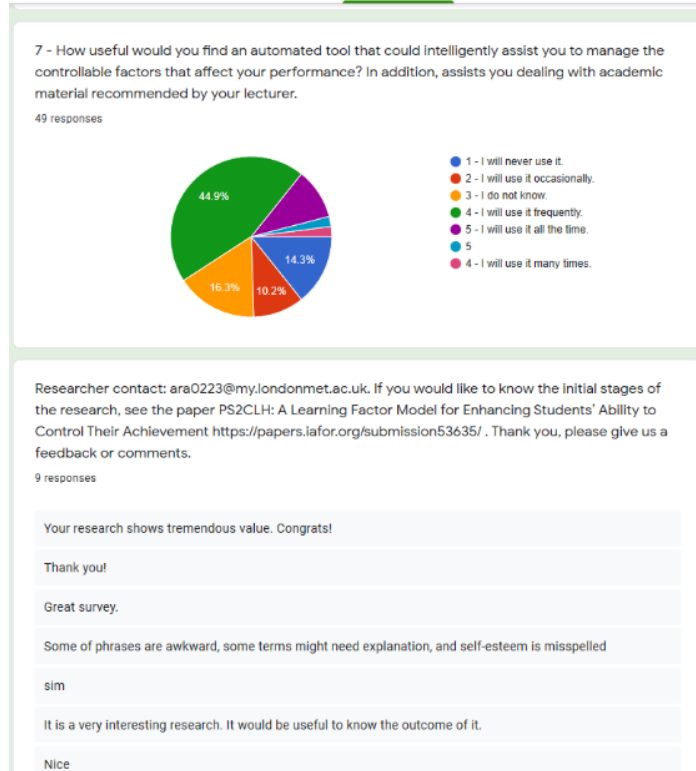


6 - Which are the Health & wellbeing factors that you believe affects your academic performance?

49 responses



7 - How useful would you find an automated tool that could intelligently assist you to manage the controllable factors that affect your performance? In addition, assists you dealing with academic material recommended by your lecturer.



Top Decision Rules for 'TargetVariable'

Decision Rule	Most Frequent Category	Rule Accuracy	Ensemble Accuracy	Interestness Index
(AimForExcellenceInEverythingYouDo > 3.0) & (OnAverageHowManyHoursYouPlaydistratationPerDay <= 3.0) & (OnAverageHowManyHoursYouPlaydistratationPerDay <= 4.0) & (PracticeTests > 1.0) & (DistributedStudy <= 3.0)	0.0	1	1	1
(EstablishAndAchievePersonalGoals > 1.0) & (DistributedStudy > 2.0) & (Expression > 2.0) & (AimForExcellenceInEverythingYouDo > 3.0) & (EstablishAndAchievePersonalGoals <= 3.0)	0.0	1	1	1
(OnAverageHowManyHoursYouPlaydistratationPerDay > 2.0) & (SetPriorities <= 1.0) & (AimForExcellenceInEverythingYouDo <= 3.0) & (EstablishAndAchievePersonalGoals <= 3.0)	0.0	1	1	1

(SetPriorities <= 1.0) & (EstablishAndAchievePersonalGoals > 3.0) (OnAverageHowManyHoursYouPlaydistratationPerDay > 1.0) & (Expression > 3.0) & (SetPriorities > 1.0) & (IDoNotFeelThatIGiveToMyselfThe ValueThatIDeserveToHaveAsAPerson <= 4.0) & (DistributedStudy > 2.0)

0.0 1 1 1

0.0 1 1 1

Table - Random trees 1 model Decision rule table

A	B	C	D	E	F	G	H	I	J	K	L
1	Time User	University - Country	Gender	Age	1 - Which are the Psych	2 - Which are the Self-res	3 - Which are the Sociolo	4 - Which are the Communi	5 - Which are the Learning	6 - Which are the Health &	7 - How useful would y
2	20200514	11 London Met - UK	Male	> 30	all the above	lack of priorities to and time ma	family income	interpretation	Reread over and summarize	have energy during study time.	5
3	20200516	1: University of Houston	Male	21- 25	Anxiety or fear	Procrastination	Discouragement and negativit	Expressing yourself	Highlighting and underlining	Sleep problems	4 - I will use it many times.
4	20200516	1: Florida International University	Female	16- 20	Stress	Procrastination	None	Expressing yourself	Self-explanation	None	4 - I will use it frequently.
5	20200516	1: Netherland	Female	16- 20	Stress	Do not know how to set priorities	Studying and working	Grammar and vocabulary	Problem with calculus and math	Feel always tired	4 - I will use it frequently.
6	20200516	1: Lamar University- USA	Female	16- 20	Stress;Anxiety or fear;Low	Procrastination;Not Accept the c	Long distance;Family income;	Expressing yourself	Reread;Practice tests;Prepare st	Feel always tired;Eating unhealt	5 - I will use it all the time.
7	20200516	2: Florida International University - USA	Genderfl	16- 20	Stress;Anxiety or fear	Immediacy;Procrastination;Bad	Addicted to sensual images;vi	Inability to write material for pape	Self-explanation	Lack of energy during the stud;	1 - I will never use it.
8	20200516	2: Florida International University	Male	16- 20	Stress;Anxiety or fear;Low	Immediacy;Procrastination;Do n	Long distance	Expressing yourself;Fluency in l	Reread;Practice tests;Problem w	Feel always tired	1 - I will never use it.
9	20200516	3: Florida International University-US	Female	16- 20	Stress;Anxiety or fear	Procrastination;Lack of self-con	Long distance	Understanding and interpreter	Reread	Sleep problems;No regular phys	4 - I will use it frequently.
10	20200516	4: Florida International University -USA	Female	16- 20	Stress;Feel depressed	Procrastination;Lack of self-con	Long distance	Expressing yourself;Grammar ar	Reread;Self-explanation;Highlig	Feel always tired;Lack of energy	4 - I will use it frequently.
11	20200516	7: FIU	Male	16- 20	Stress;Attention disorder	None	None	None	Reread;Practice tests;Self-expla	None	1 - I will never use it.
12	20200516	9: UK	Male	18- 20	Stress;Anxiety or fear	Procrastination;Bad Time mana	Family income;Studying and	Fluency in language;Grammar ai	Reread;Self-explanation;Highlig	Sleep problems;Feel always tire	3 - I do not know.
13	20200516	9: Portugal	Female	21- 25	Stress;Anxiety or fear	Procrastination;Not Accept the c	Discouragement and negativit	Sluttering or typology of disfluen	Prepare summaries;Preparation	Sleep problems;Feel always tire	4 - I will use it frequently.
14	20200516	10: Universidade Agostinho Neto - Angola	Female	> 30	Anxiety or fear;Low standar	Immediacy;Procrastination;Bad	Long distance;Family income;	Fluency in language;Grammar ai	Practice tests;Prepare summariz	Sleep problems;Eating unhealt	3 - I do not know.
15	20200516	12: Florida International University, USA	Male	21- 25	Stress;Anxiety or fear	Procrastination	Bullying;Family income;Study;	Sluttering or typology of disfluen	Problem with calculus and math	Feel always tired;Heart Problem;	2 - I will use it occasionally
16	20200516	12: Angola	Male	21- 25	Stress	Bad Time management	Studying and working	Learning problem which impact i	Prepare summaries	Lack of energy during the stud;	3 - I do not know.
17	20200516	1: University of Houston- United States of	Male	21- 25	Stress	Bad Time management	Studying and working	Understand the lecturer in the cl	Self-explanation	Lack of energy during the stud;	3 - I do not know.
18	20200516	1: Angola	Male	21- 25	Anxiety or fear;Low standar	Immediacy;Procrastination;Do n	None of them	Expressing yourself	None of them	Lack of energy during the stud;	1 - I will use it occasionally
19	20200516	2: Royal Central school of Speech and Dr	Male	26- 30	Stress;Anxiety or fear	Not Accept the change	Family income;Discouragemen	Expressing yourself;Understand	Self-explanation	Feel always tired;Eating unhealt	4 - I will use it frequently.
20	20200516	2: Varsity College- South Africa	Female	21- 25	Stress;Anxiety or fear;Low	Immediacy;Procrastination	Family income;Discouragemen	Understand the lecturer in the cl	Practice tests;Prepare summariz	Sleep problems;Feel always tire	3 - I do not know.
21	20200516	2: Caçálica - Angola	Male	26- 30	Stress;Anxiety or fear;Low	Immediacy;Not aim for excellen	Bullying;Studying and workin	Expressing yourself;Sluttering o	Reread;Prepare summaries	Sleep problems;Eating unhealt	4 - I will use it frequently.
22	20200516	2: Caçálica - Angola	Male	26- 30	Stress;Anxiety or fear;Low	Immediacy;Not aim for excellen	Bullying;Studying and workin	Expressing yourself;Sluttering o	Reread;Prepare summaries	Sleep problems;Eating unhealt	4 - I will use it frequently.
23	20200516	2: Caçálica - Angola	Male	26- 30	Stress;Anxiety or fear;Low	Immediacy;Not aim for excellen	Bullying;Studying and workin	Expressing yourself;Sluttering o	Reread;Prepare summaries	Sleep problems;Eating unhealt	4 - I will use it frequently.
24	20200516	2: David Game-UK	Female	> 30	Stress	Procrastination	Studying and working	Fluency in language	Self-explanation	Lack of energy during the stud;	4 - I will use it frequently.
25	20200516	3: Florida International University - Unitec	Male	21- 25	Stress;Anxiety or fear;Low	Procrastination;Not aim for exce	Long distance;Discouragemen	Fluency in language;Understanc	Reread;Self-explanation;Prepar	Sleep problems;Feel mentally u	4 - I will use it frequently.
26	20200516	3: universidade caçálica-portugal	Female	21- 25	Stress;Anxiety or fear;Alter	Bad Time management;Lack of	Discouragement and negativit	Expressing yourself;Fluency in l	Highlighting and underlining;P	Sleep problems;Feel mentally u	4 - I will use it frequently.
27	20200516	3: FIU- Miami Dade	Male	26- 30	Stress;Anxiety or fear;Not k	Procrastination;Not aim for exce	Family income;Studying and	Grammar and vocabulary;Under	Reread;Prepare summaries;Pro	Sleep problems;Feel always tire	3 - I do not know.
28	20200516	4: David Game-UK	Female	> 30	Stress	Procrastination	Studying and working	Fluency in language	Self-explanation	Lack of energy during the stud;	4 - I will use it frequently.
29	20200516	5: Babeç™™ Bolyai University Cluj-Napoc	Female	> 30	Anxiety or fear;Low self-ste	Lack of self-control to avoid dist	Studying and working	Expressing yourself;Fluency in l	Reread;Self-explanation;Prepan	Feel always tired;Lack of energy	5 - I will use it all the time.
30	20200516	5: Portugal	Male	21- 25	Stress	Immediacy	Long distance	Learning problem which impact i	Prepare summaries	No regular physical activity and	2 - I will use it occasionally
31	20200516	6: Babeç™™ Bolyai University Cluj Napoc	Male	> 30	Lack of motivation	Procrastination;Not aim for exce	Studying and working	Fluency in language	Self-explanation	I don't have any	1 - I will never use it.
32	20200516	6: Kalangonjo - Angola	Male	26- 30	Attention disorder	Bad Time management	Long distance;Family income	Understand the lecturer in the cl	Problem with calculus and math	Sleep problems	5 - I will use it all the time.
33	20200516	6: DGHE, UK	Female	26- 30	Attention disorder	Bad Time management;Do not l	Studying and working	Grammar and vocabulary	Problem with calculus and math	Eating unhealthily;Lack of ener	4 - I will use it frequently.
34	20200516	10: Liverpool	Female	> 30	Stress;Anxiety or fear	Do not know how to set priorities	Studying and working	Expressing yourself;Fluency in l	Self-explanation;Prepare summ	Feel always tired;No regular phy	4 - I will use it frequently.
35	20200517	12: LoneStar College	Female	21- 25	Stress;Anxiety or fear;Feel	Procrastination;Bad Time mana	Bullying;Family income;Disco	Fluency in language;Grammar ai	Problem with calculus and math	Sleep problems;Feel always tire	4 - I will use it frequently.
36	20200517	12: University of Liverpool - UK	Female	> 30	Stress;Anxiety or fear;Low	Bad Time management;Lack of	Bullying;Discouragement and	Expressing yourself;Fluency in l	Reread;Prepare summaries	Feel mentally unhealthy;No reg	4 - I will use it frequently.
37	20200517	12: Florida International University- USA	Male	21- 25	Stress;Low self-steem;Feel	Procrastination;Not aim for exce	Family income;Studying and	Fluency in language;Understanc	Practice tests;Prepare summariz	Sleep problems;Lack of energy	3 - I do not know.
38	20200517	12: Florida International University - Unitec	Female	21- 25	Anxiety or fear;Low standar	Procrastination;Not aim for exce	Family income;Studying and	Grammar and vocabulary;Under	Preparation of a questionn	Feel always tired;Eating unhealt	4 - I will use it frequently.
39	20200517	2: University of Liverpool	Female	> 30	Stress	Bad Time management	Discouragement and negativit	Fluency in language	Preparation of a questionn	No regular physical activity and	4 - I will use it frequently.
40	20200517	3: Florida international university	Female	18- 20	Stress;Anxiety or fear	Procrastination;Bad Time mana	Studying and working;Discou	Expressing yourself;Fluency in l	Reread;Self-explanation;Highlig	Sleep problems;Feel always tire	5 - I will use it all the time.
41	20200517	11: Liverpool	Female	> 30	Stress;Anxiety or fear	Do not know how to set priorities	Long distance;Family income	Fluency in language;Understanc	Practice tests	Feel mentally unhealthy;Lack of	1 - I will never use it.
42	20200517	11: University of Chester, england	Female	21- 25	Stress;Anxiety or fear;Low	Procrastination;Bad Time mana	Studying and working;Discou	Understand the lecturer in the cl	Reread;Practice tests;Self-expla	Sleep problems;Feel mentally u	3 - I do not know.
43	20200517	6: University of Liverpool	Female	> 30	Stress	Bad Time management	Discouragement and negativit	Fluency in language	Preparation of a questionn	No regular physical activity and	4 - I will use it frequently.
44	20200517	7: What that is necessary	Female	> 30	Stress;Feel depressed;Son	Procrastination;Bad Time mana	Bullying;Family income;Stud;	Fluency in language;Learning pi	Preparation of a questionn	F Sleep problems	4 - I will use it frequently.
45	20200518	8: Lone star college- the United States	Female	18- 20	Stress;Procrastination	Procrastination;Bad Time mana	Discouragement and negativit	Fluency in language	Problem with calculus and math	Lack of energy during the stud;	1 - I will never use it.
46	20200518	12: IZ*	Male	21- 25	Stress	Procrastination	Discouragement and negativit	Learning problem which impact i	Preparation of a questionn	Eating unhealthily;No regular pf	2 - I will use it occasionally
47	20200518	9: Feup-Portugal	Male	18- 20	Stress	Procrastination;Lack of self-con	Long distance;Family income;	Expressing yourself	Reread;Practice tests;Self-expla	Sleep problems	2 - I will use it occasionally
48	20200521	1: Angola	Female	26- 30	Stress;Anxiety or fear;Low	Immediacy;Procrastination;Not i	Long distance;Family income;	Grammar and vocabulary;Under	Self-explanation;Prepare summ	Sleep problems;Feel always tire	5 - I will use it all the time.
49	20200521	2: Florida International University -USA	Female	16- 20	Stress;Feel depressed	Procrastination;Lack of self-con	Long distance	Expressing yourself;Grammar ar	Reread;Self-explanation;Highlig	Feel always tired;Lack of energy	4 - I will use it frequently.
50	20200528	1: Universitat Viadrina - Germany	Female	21- 25	Anxiety or fear;Feel depres	Procrastination	Long distance;Discouragemen	Understand the lecturer in the cl	Reread;Self-explanation;Highlig	Sleep problems;Feel always tire	4 - I will use it frequently.

Please, provide Psychological concerns

I feel stressed frequently.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: Headaches and Stomachaches; Sleep Issues; Sweating; Changes in socialization; dry mouth; changes in appetite; increased irritability and impatience; difficulty concentrating; excessive worry and negative thoughts; and more similar symptoms.

Regularly, I am anxious or fearful.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: Feeling nervous; restless or tense; having a sense of impending danger; panic or doom; Having an increased heart rate; Breathing rapidly; sweating; Trembling; Feeling weak or tired; Trouble concentrating or thinking about anything other than the present worry; and more similar symptoms.

I have low standards in my academic results.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: Accepts mediocrity; Do not care about their low academic performance; Below standard; and more similar symptoms.

I might have low self-esteem.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: Apologizing too much; Afraid to express diferent opinions or ideas; Fear of making mistakes believing others are more capable; and more similar symptoms.

I regularly feel depressed.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: Constant depressed mood or irritable; you feel down or irritated most of the day, nearly every day, decreased interest or pleasure; You lose interest in doing things you used to enjoy, such as sports, hobbies, movies, or hanging out with friends; and more similar symptoms.

Usually, I feel Loneliness.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: Spend a lot of time alone; Constantly checking social media; Are unproductive; Get stuck on the negatives; Seem to be sick or ill frequently; Seem overly attached to your possessions or hobbies; and more similar symptoms.

NEXT

Please, provide Self-Responsibility concerns

I am not a patient person, I want it now.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I do not aim for excellence.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I might have bad time management.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I do not know how to set priorities.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I have a lack of self-control to avoid distractions.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

Normally, I do not stablish and achieve personal goals.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

PREVIOUS

NEXT

Please, provide Social concerns

Sometimes, I do feel discriminated.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I live far away from the University.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Long distance: [know more](#).

I feel that my family income negatively affects my studies.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I am a student with a par time/full time job.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

I might be a University dropout.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I might be addicted to pornograph.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

PREVIOUS

NEXT

Please, provide Communication concerns

It is challenge for me to express myself.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

I am not fluent in English Lnaguage.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

I do have grammar and vocabulary problems.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

I have problems understanding my lecturer in classroom.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

I have difficulties understanding or interpretation of what I read.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

I might have a communication problem which impact my learning.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

PREVIOUS

NEXT

Please, provide Learning concerns

I face problems on rereading.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I have complications in practicing tests.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I have difficulties in self-explanation study type.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I do have problems in making a summarization of my studies.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

I might have problem highlighting and underline what I study.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

It is problematic for me to prepare a questionnaire.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree

Frequent symptoms: [know more](#).

PREVIOUS

NEXT

Please, provide Health & wellbeing concerns

I have constantly sleep problems.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

Frequently, I feel tired.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

I might be eating unhealthily.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

I feel mentally unhealthy.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

Frequently, I feel a lack of energy during study time .

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

I do not practice regular physical exercises.

Strongly_disagree Tend_to_disagree Do_not_know Tend_to_agree Strongly_agree
Frequent symptoms: [know more](#).

PREVIOUS

NEXT

Carl Jung presented the theory of psychological types, which has the same purpose, individualise people by their different functions and attitudes of consciousness. Categorised by their preference of wide-ranging behaviours, three areas of preferences and dichotomies: Extroverted (E) vs. Introverted (I); Sensing (S) vs. Intuition (N); Thinking (T) vs. Feeling (F). Dr. Isabel B. Myers, continued with Jung's theory, which later she added one more field as a fourth antagonism per; Judging (J) vs. Perceiving (P) (Briggs. M, 1980)

Please, Choose the personality type you would like to have as Assistant, selecting one on each pair!

Extrov. vs. Introv. Extroverted (E) Introverted (I)
Turn to others are an (E), Turn inward are an (I)

Sensing vs. Intuit. Sensing (S) Intuition (N)
Pragmatic way are an (S), Creative way are an (N)

Think vs. Feel Thinking (T) Feeling (F)
Seek objective truth are a (T), Seek harmony are an (F)

Judg. vs. Perceiv. Judging (J) Perceiving (P)
Get closure and act are a (J), Stay open and adapt are a (P)

PREVIOUS

SUBMIT