NLP K-means algorithm incorporated into a proactive chatbot to assist failing students

Dr. Arlindo Almada
School of Computing and Digital Media, London Metropolitan University
Universidade Católica de Angola
London, UK
ara0223@my.londonmet.ac.uk

Dr. Qicheng Yu
School of Computing and Digital Media, London Metropolitan University
London, UK
q.yu@londonmet.ac.uk

Dr. Preeti Patel
School of Computing and Digital Media, London Metropolitan University
London, UK
p.patel@londonmet.ac.uk

***Abstract*—Predicting failure and individually assisting failing students is an ongoing challenge for most universities. This paper focuses on natural language processing and clustering the k-means algorithm applied to active chatbots. It aims to help students, and specifically to identify and predict failing students and proactively help them. Furthermore, it suggests an intervention to help students based on controllable academic factors that affect their academic performance. First, the authors outlined the research context for achieving this goal and created a predictive model of students' academic performance. The research results indicate a correlation between the variables with an accuracy of 0.935 and a precision of 0.76. Next, the k-means algorithm was used to cluster the students' problems or different factors that affect the students' academic performance. Finally, the k-means algorithm was integrated into an active chatbot to help students according to their problem groups.**

Keywords—Students' assistance, academic performance, proactive chatbot, group of problems, cluster

I. INTRODUCTION

The ideal solution for all failing students might be to give them individualised help during their studies. However, learning is not an easy journey, but is full of challenges and difficulties. Many students struggle to overcome these without help. Unfortunately, most institutions cannot quickly provide one-on-one support due to resource constraints, and this is an ongoing challenge for universities [1]. This situation often results in poor performance and students failing courses, an increased drop-out rate, and poor key performance indicators (KPIs) for institutions [2].

Two years ago, the COVID-19 pandemic meant that all students had to study online, which made the need to assist students more urgent not only in their academic subjects but also in the personal and social aspects of their lives. The new environment further showed how the personal and sociological factors under their control could affect their academic performance [3,4,5].

Providing a positive student experience is one of any University's priorities: ensuring engagement, academic success, professional development, and self-evaluation are a priority. What if they could identify and assist students with the controllable issues which affect their academic performance? This would involve helping students to deal with greater levels of uncertainty, fear, test anxiety and high-stress levels by finding a way to assist them with setting goals, time management, developing resilience, improving their communication skills and study techniques and overcoming sleep problems and generally improving their health and wellbeing. These questions were answered in the author's previous research [6], which aimed to identify factors under students' control as a basis for assisting them.

One of the author's previous research projects gaps [6] was to assist a large number of students with different groups of problems. This involved helping the more needy students in the same way as those with fewer problems, bearing in mind that some students need more attention than others. Therefore, this paper proposes to assist students by using a proactive chatbot which identifies the different problems students encounter and assists them accordingly.

This research builds on previous studies to develop a new approach to assisting students [7] through a more personalised and dynamic chatbot which provides individualised assistance to all students by taking into consideration the number and type of problems each student is grappling with at that point and drawing more attention to failing students.

II. ASSISTING STUDENTS WITH THE PROACTIVE CHATBOT BASED ON THEIR CLUSTERING

In December 2022, the popular Chat Generative Pre-trained Transformer (ChatGPT) was launched [8]. This is a conversational chatbot assistant with multi-purpose functions. However, it also highlights a new set of questions and problems - for instance, plagiarism at universities and source credibility. For a technology to be sustainable and survive, these are crucial issues that need to be addressed. In addition to this, the answers to be offered by the proactive chatbot are created by experts and need to be approved by the university.

Experience has shown that some students need more attention than others, and therefore a proactive chatbot with a certain level of extraversion was created. It has more interaction with those students who have a set of problems which requires more attention. This section first introduces the research background and our previous work related to the current study, and then shows how to categorise problems into different groups or clusters. It concludes by describing how the proactive chatbot is designed to address students' issues.

A. The Context of the Research

Seeking to enhance students' academic performance, the authors' previous research started by creating an educational model that combines the major controllable factors, i.e., the factors under students' control. *The model is a student-controllable learning factor model which combines the perspectives of Psychology, Self-responsibility, Sociology, Communication, Learning and Health & wellbeing (PS2CLH)* [9]. The PS2CLH model thus provides a basis for establishing students' learning profiles. Each perspective is manifested in a group of factors; for instance, the Psychology perspective is manifested in factors such as stress, depression, anxiety or fear and low self-esteem [9].

Universidade Católica de Angola.

The model makes it possible to further represent in three dimensions (3D) the students' factors by applying a self-evaluated questionnaire and from the responses to create coordinates according to the number of problems the students identify. See the paper [10] for more information.

Finally, a proactive chatbot framework [6] was developed based on the previous research, with the aim of supporting students in relation to the academic factors which affect their academic performances. This is presented in the following diagram.

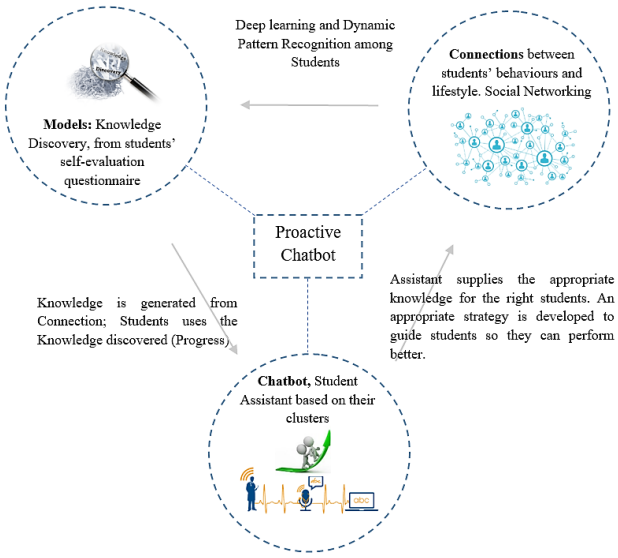

Fig. 1. The intervention context

Fig. 1 shows the context in which the proactive chatbot is applied. It starts by creating social networking; each student has a profile and fills out a self-evaluation questionnaire with questions related to their behaviour and lifestyle. The information this provides generates knowledge discovery, finds a pattern in data, and identifies the correlation between factors affecting students' academic performance as a basis for developing students awareness, making the unconscious conscious, concerning the group of problems which affects their results. The proactive chatbot thus builds on that knowledge to assist students individually.

Carl Jung, one of the fathers of psychoanalysis, stated: "*Until you make the unconscious conscious, it will direct your life, and you will call it fate*" [11]. Almost all human behaviour is known to be controlled by the unconscious or subconscious [12]. According to science, nearly all brain activity happens subconsciously [13]. Unfortunately, most of the population is not aware of this principle. Therefore, they are living on automatic pilot, doing the same actions and behaving in the same way in a Figure-of-eight loop. Albert Einstein stated: "*The definition of insanity is doing the same thing constantly and expecting a different result.*" Until people become aware and change the habituated reactions, responses and actions which they have conditioned into themselves unconsciously, they will not achieve different outcomes in their life.

Individuals cannot hope to improve their performance without understanding the underlying reasons for their actions. Similarly, without gaining a certain level of insight into the root causes of their emotions, they cannot consistently achieve better outcomes.

Therefore, people need to reflect on and study their behaviours and habits, and ask what feelings, behaviours, patterns and actions are on autopilot. What can human beings bring to the surface and into their consciousness? What can they do to develop daily routines and become more productive?

As human beings, people have daily habits, which can be good or bad in terms of their consequences. In brief, the human brain operates according to patterns or by following specific neural pathways; if people constantly do something, the path is set in the brain. They are therefore inclined to repeat that action, thus creating a new habit [14].

Bergson (1911) wrote of both active and passive habits. The first type of habit is passive, and develops from experiencing things which humans eventually get used to [15]. The second type is active, arising from repeated intention and effort, crystalising as skills which an individual performs with little or no thought [16]. In general, passive habits are seen as bad.

On the other hand, most active habits are reasonable considering the desired results. Consequently, homo sapiens tend to pursue active rather than passive habits, which are addictions - checking their Instagram feed every five minutes, eating a certain kind of unhealthy food, drinking, smoking and so on. Therefore, they are inclined to regret doing those addictive actions with undesirable effects.

Consequently, they are frittering away their brain power and draining themselves emotionally of the energy they need in life. Most people have engaged in active habits at some point in their lives, such as learning a foreign language, playing a musical instrument or learning how to do a sport. In the beginning, it is difficult and slow; they have to take each step in turn, until the activity is established in their brain and becomes automatic. When one stage becomes automated, they can go on to the next level and know they can become more and more proficient.

Consequently, people want active habits in their life. Humans need to develop these using baby steps. For instance, one thing they can do is try to eliminate bad habits so that the addictions slowly disappear. For example, they could disconnect from the internet for several hours a day or longer, so they stop wasting time. People have now got to fill up that time with something productive, something active, perhaps taking up something new that will force their mind to be involved, learning a new skill. By doing that, they are slowly developing this discipline inside themselves that will allow them to take on other challenges.

To keep developing active habits, they need continuous challenges. Thus, for instance, if people are working on a project, they could give themselves a deadline, perhaps aiming to complete it in a month. If they procrastinate and waste weeks, maybe a week before the deadline, they will start to feel energetic and try to cram everything in and get the thing done in the final few days. Perhaps this is not so good.

People can challenge themselves and give themselves only a week to finish an activity that typically takes a month. That challenge will force them to get into higher gear and work hard every day. Then again, when humans are in the passive, habituated mode, they believe that pleasure comes from activities that are just a waste of time, such as watching

internet porn, spending hours watching sports or checking their Instagram. All this might seem relaxing, fun or socially enjoyable. Thinking of discipline and working hard is not fun at all. People should switch from autopilot mode and realise that they cannot develop healthy habits unless they change their notion of pleasure, gratification or immediate reward.

In an interview, Kaku was asked which specific psychological test correlated with success in life, and he answered that the marshmallow experiment successfully predicted people's success. This test was invented by Walter Mischel, who researched delayed gratification in young children, correlating it with self-control in human growth. *The test consists of asking a child if they want 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 wished for shortcuts or those who did not want to put in hard work* [17]. Another scientist, Angela Duckworth, also studied the effects of self-control and delayed gratification; her important research, like "Grit", the power of passion and self-control, also shows evidence that self-control is correlated with people's success [18].

Long-lasting pleasure and satisfaction do not come from immediate rewards and immediate gratification. Instead, research shows that if individuals work hard and consistently until the end, it will give them a sense of fulfilment and happiness. It will be far greater than anything they could have obtained from the instant gratification their bad habits gave them [17].

The next section develops a predictive model of students' results using the factors that affect their performance, and a correlation is found among the variables. The goal is to find patterns in the data, which give insights related to students' behaviours, which in turn lead to a clear vision of how to automate the proactive chatbot to assist students.

B. Data Processing, Modelling and Correlation

As a result of limited accessibility for a significant number of students, the dataset based on the PS2CLH was gathered through an experiment conducted at Universidade Católica de Angola in 2018. The questionnaire consisted of 6 (Health & wellbeing), 7 Communication, 7 Psychology, 12 Sociology, 8 Self-responsibility, 8 Learning, and 5 multiple choice questions, with a 5-point Likert scale for responses. The research was conducted with 540 learners from various courses aged between 20 and 25; the data was collected between September 2nd and November 28th, 2018.

In the sequence, the data was processed by applying artificial intelligence algorithms to find the best data modelling.

The target variable (students' results) was imbalanced, meaning that the representation of outstanding students (1) is significantly below that of average students (0).

The distribution of the data after cleaning was as follows:

- Total number of students: 540
- Training data: 60%, 324 students
- Validation data: 20%, 108 students
- Test data: 20%, 108 students

The test data Confusion Matrix for the prediction of students' academic performance based on factors affecting their academic performance is presented in Fig. 2 as follows:

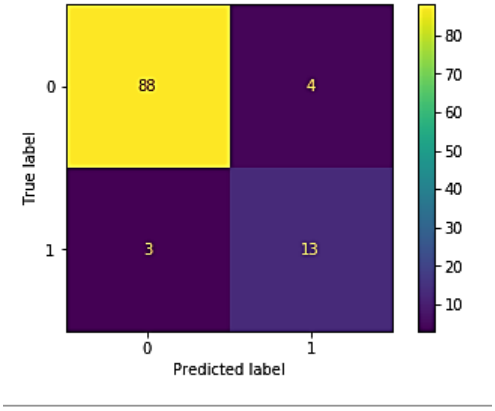

Fig. 2. Test Confusion Matrix

The Confusion Matrix is a table that shows the performance of a classification model. In this case, the rows represent the true labels (0 and 1) and the columns represent the predicted labels (0 and 1). The values in the cells of the Confusion Matrix represent the number of instances for which the model predicted a certain label (either 0 or 1).

In the given Confusion Matrix:

- True label 0 was predicted as 0 in 88 instances.
- True label 0 was predicted as 1 in 3 instances.
- True label 1 was predicted as 0 in 4 instances.
- True label 1 was predicted as 1 in 13 instances.

It should be noted that the Confusion Matrix helps evaluate the performance of a classification model by providing insights into the number of instances that were predicted correctly (diagonal values) and instances that were misclassified (off-diagonal values). The Confusion Matrix is a valuable tool for assessing the accuracy and effectiveness of a classification model in predicting different classes.

In the presence of an imbalanced dataset, a Standard scalar was applied to standardise the features. The dataset was later separated into training, testing and validation. To further improve the model, the SMOTE function was used. SMOTE stands for Synthetic Minority Oversampling Technique, and generates virtual training records for the minority class using linear interpolation [19].

The final step in creating the model was to apply logistics regression to the training dataset. The application of logistic regression model to a dataset is a statistical method of analysing the relationship between predictor variables and a binary outcome variable. It was faster than other models. This technique makes predictions and classifies data points into discrete categories based on their characteristics. Using logistic regression on the training dataset made it possible to identify effectively patterns and trends that may impact students' academic performance. Furthermore, this model makes it possible to build a predictive model that can provide insights and facilitates informed decisions about future outcomes.

Next, the model's efficiency was evaluated by using performance metrics. First, the Receiver Operating Characteristic curve (ROC) and the Area under the ROC Curve (AUC) score were used, representing the Area under the (AUROC or AUC/ROC). Then, to evaluate the classification models used, precision, recall and the F1 scores (the F1 scores vary from 0 to 1) were all used to measure the efficiency of the model.

```
ROC_AUC_score 0.8845108695652174
Accuracy 0.9351851851851852
Precision 0.7647058823529411
Recall 0.8125
F1 score 0.787878787878788
```

Fig. 3. The Results Generated by the Model

As illustrated in Fig. 3, the analysis gave promising results, with a notably high Receiver Operating Characteristic Area Under the Curve (ROC AUC) score of 0.88, a precision rate of 0.76 and an overall accuracy level of 0.935. Additionally, the F1 score and the recall metrics were also commendable at approximately 0.79. However, it is essential to note that improvements could be achieved by implementing feature selection methods and considering results for the major class rather than the minor class only.

These outcomes indicate that the selected variables pertaining to students' factors are strongly associated with their academic performance, demonstrating distinct patterns that differentiate top-performing students from others. Moreover, the findings suggest a correlation between student factors and their academic performance, highlighting the significance of these variables in predicting educational outcomes. Overall, the results underscore the potential of the model developed as a valuable tool for understanding and predicting students' performance.

Subsequently, the correlation algorithm was built. 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* [20].

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

where:

r: Pearson correlation coefficient

X_i, Y_i: Individual data points of the two variables being correlated.

Σ: Summation symbol, which indicates that the values inside the parentheses are to be summed over all data points.

sqrt: Square root function, used to calculate the square root of the values inside the parentheses.

Results of the top seven positive Pearson correlation coefficients between the student result and the PS2CLH variables.

- Establish and achieve personal goals 0.71
- Set priorities 0.63
- Practice tests 0.57
- Hours of play/distraction per day 0.52
- Sleep problems 0.49
- Stress 0.47
- Anxiety or fear 0.45

The Pearson Correlation coefficient 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 [21].

C. Group of Problems or Clustering of Factors That Affect Students' Performance

There are supervised and unsupervised learning algorithms in the Natural Language Processing field. The clustering algorithm relates to unsupervised learning and focuses on assigning data into groups. *These groups (or clusters) are created by uncovering hidden patterns in the data to the end of grouping data points with similar patterns in the cluster* [22].

The integration of PS2CLH domains, encompassing psychology and self-responsibility, is shown in coordinate X, sociology and communication are shown in coordinate Y, and learning, health and wellbeing in coordinate Z. Building a 3D student representation of the factors that affect their performance make it possible to represent students' problems, effectively monitor the issues which are affecting them and understand the different patterns of the PS2CLH model, leading to a better understanding of the student's behaviours in various clusters.

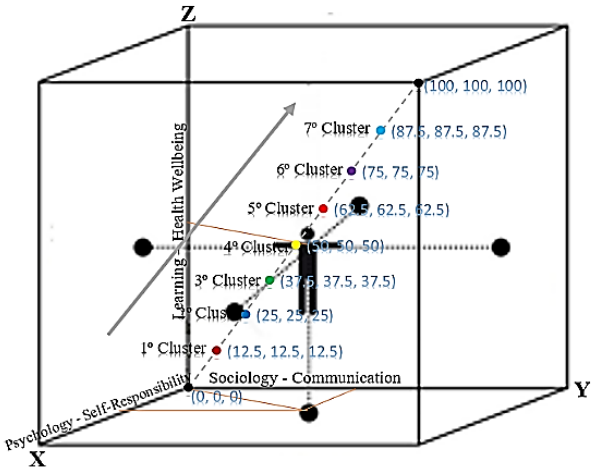

Fig. 4. 3D representation of the students' clusters centroids [9]

The image in Fig. 4, depicting 3D centroid clusters, has a value of k set at seven. The PS2CLH was employed in this study to identify controllable factors that impact students' performance, with variations observed in accordance with their respective k clusters.

The K-means algorithm first finds the k points or centroids, and then minimises the sum of squared errors and

aims to identify the total error. This procedure is performed iteratively until it no longer yields further reductions. At that point, the sum of squared errors has reached a minimum, and the K centroids will be classified into groups or clusters [22].

Thus, the K-means algorithm goes through the following steps:

- *Firstly, initialising k centroids randomly.*
- *Calculating the sum of squared deviations.*
- *Assigning a centroid to each of the observations.*
- *Calculating the sum of total errors and comparing it with the sum in the previous iteration.*
- *If the error decreases, recalculating centroids and repeating the process* [23].

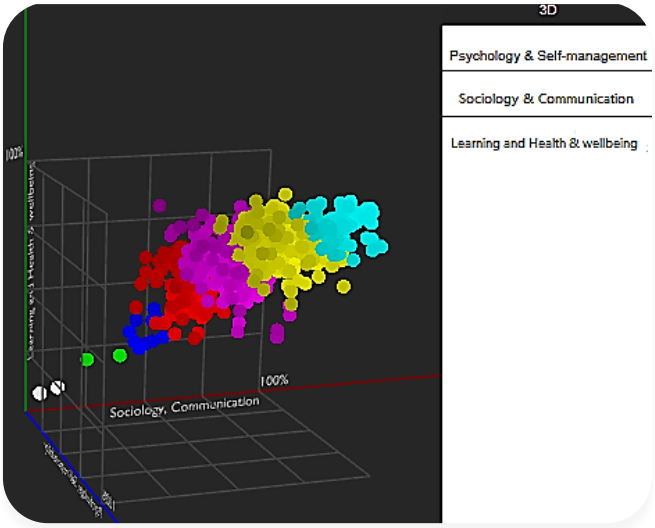

Fig. 5. Representation of the K-means by colour

Fig. 5 uses different colours to represent students in the same cluster; each point represents a student's position. The k-means cluster representation, based on these students' controllable factors, will be inserted into the proactive chatbot and will perform the following functions: the cluster will identify different groups of student problems, thus finding failing students. In addition, it will show the patterns which emerge, and identify the top students, giving valuable insights and highlighting the factors which make the difference between different groups of learners, and ultimately enable university decision-makers to take suitable action.

III. Proposed Intervention Using the Proactive Chatbot to Assist Students Based on Problem Groups

Typically, universities have a student assistance department. However, it is impractical to identify those who need more attention and assist all students individually due to the number of learners. The following intervention to deal with this issue is therefore proposed.

The aim is to effectively assist all university students from the beginning of the academic year, remembering that each country and each university presents its own challenges. The intervention, therefore, starts by identifying and understanding the major educational factors that affect students' results. To do this, qualitative research is undertaken by interviewing students, lecturers and local experts in psychology, sociology, learning, communication and health and wellbeing. The final objective is to create a questionnaire based on the most relevant controllable problems affecting those particular students.

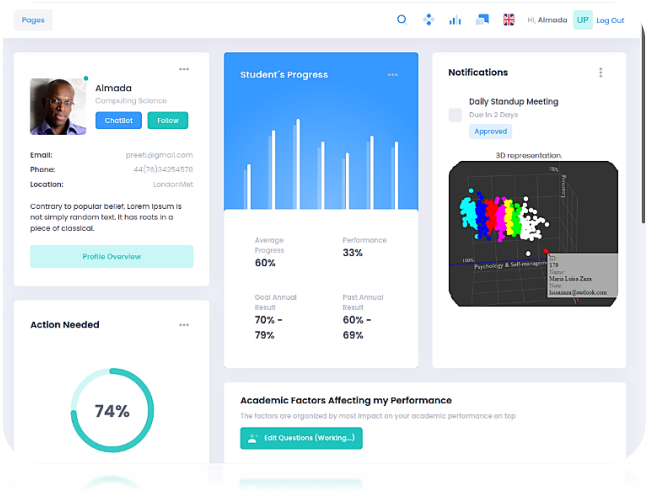

Fig. 6. Example of a student profile [6]

Fig. 6 shows the student profile, highlighting the buttons for interaction with the chatbot and the self-evaluated questionnaire featuring PS2CLH questions.

After building the web-based self-evaluated questionnaire, the next step is data collection. Although collecting a significant number of registers is essential for optimal results in the data analysis phase, it is crucial in this phase to make students understand that it is not a judgmental act or test so that they will cooperate fully.

Once the data preparation and understanding are completed, the cleaning process is undertaken. Then build the model, and find the correlations between the variables and their impact on the students' results, as demonstrated in the previous section. What is created is a group of problems or clustering of factors as presented in the last section.

At this stage, the raw data has been organised through information processing techniques such as creating a model, identifying correlations, and clustering. This transformed data will now become knowledge. In the subsequent phase of the intervention, investigators will utilise this knowledge to construct individual student profiles and develop a proactive chatbot to assist them.

The proactive chatbot framework comprises two distinct parts: the wide-ranging extended chatbot and the Educational Chatbot Ecosystem. The first part draws inspiration from the BERT (Bidirectional Encoder Representations from Transformers) model, a machine learning technique for natural language processing. It serves as the initial interaction point for students' input or questions. The second part, the Educational Chatbot Ecosystem, functions as an educational support system that enhances the chatbot's AI capabilities, facilitating and improving interactions between students and lecturers/assistants. This ecosystem encompasses several components: Knowledge Database, Suggest Factor, Rating System, Interactor Facilitator, Profile Customizer and Multimodality features [6].

The Interaction Facilitator component enhances the usability and reliability of the proactive chatbot by facilitating its interaction with lecturers or experts to address questions effectively. In addition, it acknowledges that certain students, especially introverts, may prefer to ask questions outside of the traditional classroom environment. As a result, this component empowers the proactive chatbot to play a more prominent role in assisting students, providing them with a comfortable avenue to seek answers to their queries. Furthermore, using multimodality in the chatbot's responses enhances the student's learning experience. The chatbot leverages multimodality to provide responses that incorporate various modes of communication, and it proactively suggests related factors correlated with the student's question based on student feedback.

The Knowledge Database component ensures that the proactive chatbot learns from past questions and can handle unexpected situations where the student's queries are not found in the knowledge database. In addition, by utilising the K-means algorithm to create student clusters, the proactive chatbot can adjust the level of student interaction, automating and differentiating its interventions to provide more attention to students who require it the most. This idea enables the chatbot to adapt and provide personalised support to meet the unique needs of individual students.

Once the questionnaires have been completed, those factors that students mention they are struggling with appear on their profile. According to student clustering, the proactive chatbot will increase or decreases the level of interaction with students. For example, for students in critical clusters 1 and 2, as shown in Fig. 4, where the number of problems is higher, the proactive chatbot will provide daily assistance by sending orientations on how to deal with that specific issue, and it will pay close attention to students in cluster 1. Furthermore, the proactive chatbot will send guidance daily for students in clusters 3, 4 and 5. Finally, for students in clusters 6 and 7, the proactive chatbot will send orientations once or twice weekly. In addition to this proactive interaction, students can interact with the chatbot at any time, asking questions about any factors they may be concerned about.

A. Discussion of Results

The discussion of the results of incorporating the NLP K-means algorithm into the proactive chatbot intervention can be summarised as below.

Effective identification of students' controllable educational factors: the qualitative research undertaken through interviews with students, lecturers, and local experts in various fields to identify and understand the significant educational factors that affect students' results provided a basis for creating a questionnaire which focused on the most relevant controllable problems affecting the students. This will ensure that the intervention is tailored to address the specific challenges faced by the students, bearing in mind the unique context of the country and the university.

On the other hand, the data used for training the K-means algorithm may contain biases in relation to demography, culture or sampling. Such biases could impact the clustering results and lead to the unequal treatment of certain groups of students or unequal access to support. For example, if the data used were biased towards a specific demographic group, the proactive chatbot might not cater effectively to the needs of other demographic groups, leading to a potential disparity in the assistance provided. In addition, using the NLP K-means algorithm in the proactive chatbot intervention raises ethical considerations related to fairness, transparency and accountability. Ensuring that the clustering process is fair and transparent and that the results are interpretable and clearly explainable is essential. Ethical factors should also be considered in relation to the potential impact of the intervention on students' autonomy, agency and decision-making, and measures should be in place to mitigate any potential harm or unintended consequences.

Data collection and preparation: the data was collected through the web-based self-evaluated questionnaire, and care was taken to ensure that students understood that this was not a test or in any way judgmental, which resulted in cooperative participation. The data collected was cleaned and prepared for further analysis. The information processing techniques included creating a model, identifying correlations and clustering, and transforming the raw data into knowledge. The clustering of factors, as presented in the previous section, helped in organising the data and understanding the relationships between variables, leading to the identification of problem groups or clusters. However, collecting and using student data for clustering and profiling in the proactive chatbot intervention may raise privacy and security concerns. It is, therefore, crucial to ensure that any data collected is stored securely and used only for the intended purpose of providing support to students. It is also essential to obtain informed consent from students for data collection, ensure compliance with relevant data protection regulations, and protect their privacy and rights.

Proactive chatbot intervention: the proactive chatbot was developed based on the transformed data and individual student profiles. The chatbot uses cluster information to increase or decrease student interaction with students. The chatbot assists students in critical clusters who have more problems daily and pays close attention to their needs. For students in other groups, the chatbot sends daily or weekly guidance, depending on the severity of the issues they are facing. Additionally, students can interact with the chatbot anytime, seeking answers to their questions and concerns.

However, the accuracy of the K-means algorithm in clustering students into different problem groups or clusters depends on the quality and representativeness of the data used for training. Suppose the data used is not comprehensive or does not fully capture the diversity of students' profiles and the challenges they face. In that case, the clustering results may not accurately reflect the actual needs of the students. This can lead to inaccurate profiling and may result in ineffective assistance being provided by the proactive chatbot. Furthermore, implementing and following through proactive chatbot interventions using the NLP K-means algorithm may require significant resources in terms of technology infrastructure, personnel, and ongoing maintenance. Ensuring the scalability of the intervention to cater to a large number of students and addressing technical issues such as updates, maintenance and integration with other systems is vital to ensure the sustainability and effectiveness of the intervention.

Overall, the incorporation of the NLP K-means algorithm into the proactive chatbot intervention has enabled the effective identification of educational factors, data collection and preparation, the transformation of data into knowledge

and the provision of proactive assistance to students based on the type of problem they are facing. This approach ensures that students receive tailored support and guidance which addresses their specific needs and improves their academic outcomes.

In conclusion, although incorporating the NLP K-means algorithm into a proactive chatbot intervention has the potential to improve the assistance provided to students, there are potential issues related to the accuracy of clustering, bias in the data, privacy and security concerns, ethical considerations, questions of user acceptance and engagement, as well as scalability and maintenance. Therefore, it is essential that these issues should be considered during the design, implementation and evaluation of the proactive chatbot intervention to ensure its effectiveness and ethical use.

IV. Conclusion

This research proposes an innovative intervention to assist struggling students by using a proactive chatbot that addresses their specific problem areas. Through meticulous data processing, it was observed that the factors selected in the PS2CLH model are correlated to students' academic performance. A positive trend was identified in the data, which enabled the authors to create seven distinct clusters. This knowledge thus facilitated a focused approach, allowing more attention to be given to clusters where students exhibited a higher number of factors.

This means that the proactive chatbot is designed to adapt its interactions based on the individual needs of each student; the proactive chatbot can change the level of extroversion by interacting more often with failing students. This approach has the potential to address different situations faced by students effectively. Notably, in the context of limited university staff resources, this intervention maximises the effectiveness of university assistance. However, validating the authors' assumptions through real-life interventions is imperative to collect further evidence of the proactive chatbot's efficacy in supporting struggling students.

Future work in this area could involve further fine-tuning of the K-means algorithm to improve its accuracy and effectiveness in identifying students at risk of failing. Additionally, integrating other NLP techniques, such as sentiment analysis or topic modelling, could enhance the chatbot's ability to provide personalised and context-aware assistance to students. Further research and development could also explore the use of machine learning models such as deep learning as a form of generative pre-trained transformer or reinforcement learning, which combined with the K-means algorithm would create a more sophisticated and intelligent chatbot that can adapt and improve its assistance based on students' feedback and outcomes.

References

[1] F. M. Guangul, A. H. Suhail, M. I. Khalit, and B. A. Khidhir, "Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College," Educ Asse Eval Acc, vol. 32, pp. 519–535, 2020.

[2] M. Dresel et al., "Competencies for successful self-regulated learning in higher education: structural model and indications drawn from expert interviews," Stud. High. Educ., vol. 40, no. 3, pp. 454–470, 2015.

[3] S. C. Gewalt, S. Berger, R. Krisam, and M. Breuer, "Effects of the COVID-19 pandemic on university students' physical health, mental health and learning, a cross-sectional study including 917 students from eight universities in Germany," Plos one, vol. 17, no. 8, p. e0273928, 2022.

[4] W. Leal Filho et al., "Impacts of COVID-19 and social isolation on academic staff and students at universities: a cross-sectional study," BMC public health, vol. 21, no. 1, pp. 1–19, 2021.

[5] L. Ihm, H. Zhang, A. Van Vijfeijken, and M. G. Waugh, "Impacts of the Covid-19 pandemic on the health of university students," The International Journal of Health Planning and Management, vol. 36, no. 3, pp. 618–627, 2021.

[6] A. Almada, Q. Yu, and P. Patel, "Proactive chatbot framework based on the PS2CLH model: an AI-Deep Learning chatbot assistant for students," in Intelligent Systems and Applications: Proceedings of the 2022 Intelligent Systems Conference (IntelliSys) Volume 1, 2022: Springer, pp. 751–770.

[7] M. Carter et al., "Decision-making regarding adjustments for students with special educational needs in mainstream classrooms," Research Papers in Education, vol. 37, no. 5, pp. 729–755, 2022.

[8] C. Leiter et al., "ChatGPT: A Meta-Analysis after 2.5 Months," arXiv preprint arXiv:2302.13795, 2023.

[9] A. Almada, Q. Yu, and P. Patel, "PS2CLH: A Learning Factor Model for Enhancing Students' Ability to Control Their Achievement," Tokyo, ACE2019, 2019.

[10] A. Almada, Q. Yu, and P. Patel, "Representation of the Student's Controllable Performance Features Based on PS2CLH Model," Barcelona Conference on Education, Barcelona, Spain, 2022.

[11] C. Jung, "Psychological Types." In: R. &. K. Paul, ed. Collected Works of C.G. Jung, Vol. 6. London: s.n, 1971.

[12] C. Jung, G. Adler, R. Gerhard and F. C. Hull, "Psychological Types" Collected Works of C.G. Jung. Volume 6:, Princeton: Princeton University Press, 2014.

[13] C. S. Soon, M. Brass, H.-J. Heinze, and J.-D. Haynes, "Unconscious determinants of free decisions in the human brain," Nature neuroscience, vol. 11, no. 5, pp. 543–545, 2008.

[14] J. Ren et al., "Anatomically defined and functionally distinct dorsal raphe serotonin sub-systems," Cell, vol. 175, no. 2, pp. 472–487. e20, 2018.

[15] H. Bergson, "Matter and memory" N. M. Paul & W. S. Palmer, Trans. George Allen & Co, 1911.

[16] J. Brewer, "Mindfulness training for addictions: Has neuroscience revealed a brain hack by which awareness subverts the addictive process?," Current Opinion in Psychology, vol. 28, pp.198–203, 2019.

[17] W. Mischel, "The Marshmallow Test: Understanding Self-control and How To Master It," ISBN 978055216886 ed. London: Penguin Random House UK, 2015.

[18] A. Duckworth, "Grit: Why passion and resilience are the secrets to success," ISBN 978-1-5011-1110-5 ed. New York: NY 10020, 2017.

[19] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model," IEEE Access, vol. 9, pp. 78621–78634, 2021.

[20] A. Rebekić, Z. Lončarić, S. Petrović, and S. Marić, "Pearson's or Spearman's correlation coefficient-which one to use?," Poljoprivreda, vol. 21, no. 2, pp. 47–54, 2015.

[21] V. A. Profillidis and G. N. Botzoris, "Modeling of Transport Demand, Analyzing, Calculating, and Forecasting Transport Demand," Elsevier, 2018.

[22] X. Jin and J. Han, "K-Means Clustering," in Encyclopedia of Machine Learning, C. Sammut and G. I. Webb Eds. Boston, MA: Springer US, 2010, pp. 563–564.

[23] S. Z. Selim and M. A. Ismail, "K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, no. 1, pp. 81–87, 1984.