Exploring the Use of NLP Techniques for Building Learner Models: A Study on Text Mining for Personalized Learning

Jalal Lahiassi, Souhaib Aammou, Badr Touis and
Oussama EL Warraki

Abdelmalek Essaadi University, S2IPU, Morocco
lahiassi@gmail.com

Abstract. The importance of learner models has been growing in recent years due to the increasing focus on personalized learning. Learner models provide a way to understand learner behavior and preferences, which can be used to tailor instruction to the individual needs and preferences of each learner. This can lead to more effective and efficient learning, as learners receive instruction that is tailored to their unique needs and abilities.

Furthermore, the rise of online and distance learning has increased the need for learner models. In these environments, teachers and instructors often have limited visibility into learner behavior and progress. Learner models can provide valuable insights into learner behavior and preferences, which can be used to improve the effectiveness of online instruction.

In this article we explore the use of natural language processing (NLP) techniques for building learner models, from text data, based on Text Mining.

As well, text mining is the process of extracting useful information from unstructured text data. Extracting relevant information from text data using techniques such as tokenization, stemming, and stop word removal is one of the ways in which NLP can be used to build learner models.

Once the relevant information has been extracted, it can be used to create features that can be used to train machine learning models to predict learner behavior and preferences. These features can include things like the frequency of certain words or phrases, the sentiment of learner responses, and the topics discussed in learner writing.

Keywords: Learner models, Personalized learning, Natural language processing (NLP), Text mining.

1 Introduction

The fast-paced nature of modern life has led to an increasing reliance on e-learning, which is now being used in primary schools, universities, and many other educational institutions. E-learning is a form of teaching that involves the geographical separation of teachers and learners, with learners managing their learning activities and resources autonomously.

M. Khaldi et al. (eds.), *Proceedings of the International Conference on E-Learning and Smart Engineering Systems (ELSES 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 14,
https://doi.org/10.2991/978-94-6463-360-3_11

The volume of data generated by learners using Learning Management Systems can be immense. This data can include information about their interactions with the LMS, such as the number of courses they have enrolled in, the amount of time they spend on each course, the number of assignments they have completed, and their performance on quizzes and exams.

In addition to the traces of learner activity on LMS, there may also be learner features or attributes that are stored in their profiles. These learner features can include their preferred learning style, interests, past learning results, and other relevant information. There are currently multiple techniques for analyzing data such as Natural language processing (NLP)[1] and Text Mining[2].

The goal of this study is to create personalized learning pathways that are tailored adapted to the unique needs and preferences of each individual learner. By analyzing learners' written responses to quiz questions, discussion forums, and other course materials, we can classify the clusters that identify the areas of strength and weakness in their knowledge of a particular subject. This cluster can then be used to create personalized learning pathways that focus on the topics and concepts that each learner needs. This study emphasizes the importance of personalized learning pathways in promoting effective learning. By identifying the individual needs of each learner and creating pathways that cater to those needs.

2 Theoretical framework

2.1 Learner models

Learner models are a type of representation or model that provides insight into how learners approach and learn a particular topic or subject. They are developed using data mining techniques that analyze a learner's learning behavior to identify patterns and trends that reveal how they learn, what motivates them, and how they solve problems [3].

Learner models can be used to personalize the learning experience for each learner. By understanding how a learner learns, teachers can tailor their teaching strategies and materials to better suit the learner's needs, abilities, and interests. This can help to improve the learner's engagement, motivation, and learning outcomes[4].

2.2 Personalized learning

Personalized learning is an approach to education that tailors the learning experience to the unique needs, interests, and abilities of individual learners. Rather than a one-size-fits-all approach, personalized learning recognizes that every learner is different and has unique learning needs, and seeks to provide a learning experience that is customized to each individual learner[3].

The goal of personalized learning is to create a learning experience that is more engaging, relevant, and effective for each learner. By tailoring the learning experience to their unique needs and interests, personalized learning can help learners to stay motivated and engaged in their learning, and to achieve better learning outcomes[5].

Personalized learning is gaining in popularity as technology and data analytics become more advanced and accessible, and as educators recognize the limitations of traditional, one-size-fits-all approaches to education. With personalized learning, every learner has the opportunity to achieve their full potential and to become lifelong learners who are passionate about their education[6].

2.3 Natural Language Processing

Natural Language Processing (NLP) techniques can be used to extract information and insights from text data to build learner models. Text mining, a subset of NLP, involves analyzing and extracting information from unstructured text data, such as learner essays, online discussion forums, and other types of written text[1].

One approach to building learner models using NLP is to analyze the text data to identify patterns and themes related to a learner's learning behaviors and strategies [7].

These insights can then be used to create a learner model that captures the learner's learning preferences, interests, and motivations. This model can be used to personalize the learning experience by recommending content and activities that are aligned with the learner's learning goals and preferences[8]. Another approach to building learner models using NLP is to analyze the text data to identify patterns and trends related to a learner's performance and progress[2].

3 Methodology

3.1 Design process

By analyzing the text data generated by learners, NLP techniques can be used to gain insights into their thought processes, interests, and learning styles. These insights can then be used to personalize the learning experience for each individual learner by clustering method.

In the research design for this study, we have outlined the following phases (Fig. 1)**:**

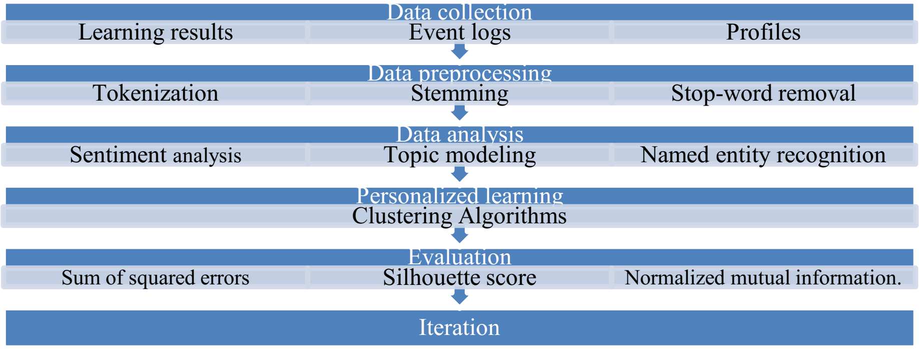


Fig.1: NLP design process for enabling personalized learning

Design process outlines a comprehensive approach for using text data techniques for personalized learning based on text mining. By following these phases, we aim to provide a more customized personalized and engaging learning experience for each individual learner.

3.2 Clustering

Clustering is a machine learning technique that involves grouping a set of objects in such a way that objects in the same group (called cluster) are more similar to each other than to those in other groups (clusters). Clustering is an unsupervised learning method, meaning that it is performed without prior knowledge of the labels or categories of the data.

In our case, we will be using partitional clustering, where the data is divided into k distinct and non-overlapping clusters based on a user-defined parameter. The most popular partitional clustering algorithm is k-means, which assigns data points to the nearest centroid iteratively until the centroids no longer change or a maximum number of iterations is reached.

3.3 k-means algorithm

The k-means algorithm is a commonly used clustering algorithm that partitions a given dataset into k clusters. The algorithm works as follows: Let X be a matrix of n rows and p columns, where each row xi represents the event data for case i, and each column j represents a feature of the events. Thus, X can be represented as X = [x1, x2, ..., xn]^T, where each xi = [xi1, xi2, ..., xip] represents a row vector of event features.

Let Y be a vector of n elements, where each yi represents the pass/fail outcome for case i. Thus, Y can be represented as Y = [y1, y2, ..., yn]^T.

Let k be the number of clusters we want to create.

The k-means algorithm can be represented as:

1. Initialize k cluster centroids randomly, denoted by μ1, μ2, ..., μk.
2. Assign each event xi to the nearest centroid based on Euclidean distance:
   - For j = 1, 2, ..., k, let Cj be the set of events assigned to centroid μj.
   - For i = 1, 2, ..., n, let j = argmin(||xi - μj||), and assign xi to Cj.
3. Recalculate the centroids based on the mean of the events assigned to them:
   - For j = 1, 2, ..., k, let μj' = (1/|Cj|)Σxi in Cj.
4. Repeat steps 2 and 3 until convergence, when the assignments no longer change.

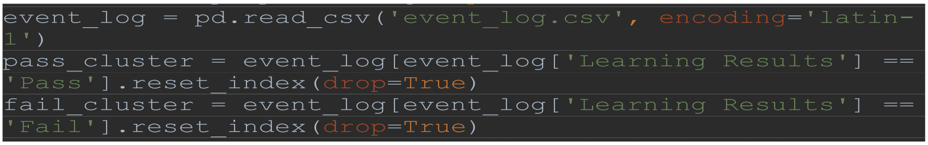
Once the clustering is completed, we can use it to predict the pass/fail outcomes for new cases using the following steps: Given a new case with event data xi, find the cluster centroid μj that is closest to xi based on Euclidean distance.

Predict the pass/fail outcome for xi to be the majority outcome of the cases in cluster Cj.

4 Implementation

The program for clustering is a two-step process that involves applying k-means clustering to first split the event log into Pass and Fail clusters, and then further dividing each cluster into sub-clusters based on learner features.

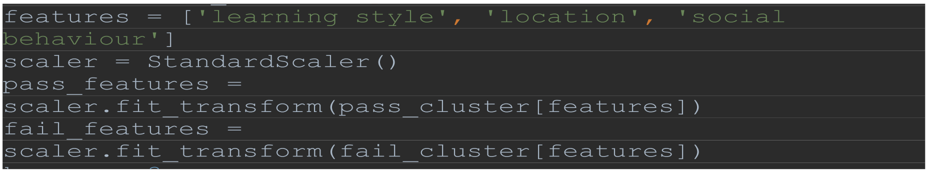
The program takes as input the event log, the number of clusters "X" the number of sub-clusters "Y" and the clustering parameter, which is initialized by Learning Results. The first step of the algorithm applies k-means clustering to the event log to generate "X" clusters according to the parameter Learning Results. This step separates the event log into two distinct clusters based on whether the learners passed or failed the learning resource.



```
event_log = pd.read_csv('event_log.csv', encoding='latin-
1')
pass_cluster = event_log[event_log['Learning Results'] ==
'Pass'].reset_index(drop=True)
fail_cluster = event_log[event_log['Learning Results'] ==
'Fail'].reset_index(drop=True)
```

Fig.2 Step 1: K-means Clustering for Learning Resource Success/Failure Classification

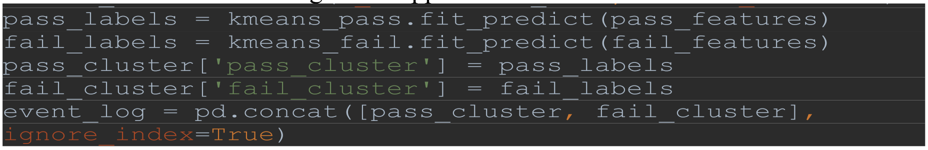
In the second step of the program, k-means clustering is applied again to each of the "X" clusters generated in the first step, to create "Y" sub-clusters based on learner features such as learning style, location, and social behavior. This step allows for a more detailed analysis of the behavior of learners within each of the initial Pass and Fail clusters.



```
features = ['learning style', 'location', 'social
behaviour']
scaler = StandardScaler()
pass_features =
scaler.fit_transform(pass_cluster[features])
fail_features =
scaler.fit_transform(fail_cluster[features])
```

Fig.3 : K-means Clustering on Learner Features: Creating Sub-Clusters for Detailed Analysis

The output of the program is a set of "X * Y" sub-clusters, each containing a group of learners with similar characteristics and learning behaviors. This information can be used to gain insights into the factors that contribute to successful learning outcomes and inform instructional design and support interventions.



```
pass_labels = kmeans_pass.fit_predict(pass_features)
fail_labels = kmeans_fail.fit_predict(fail_features)
pass_cluster['pass_cluster'] = pass_labels
fail_cluster['fail_cluster'] = fail_labels
event_log = pd.concat([pass_cluster, fail_cluster],
ignore_index=True)
```

Fig.4:Clustering Program Output: Grouped Data Visualization

Each sub-cluster generated by the program contains a set of events performed by learners who are very close in their location, learning style, and social behavior.

“Location" in learning refers to the mode of learning, such as face-to-face or online, and learners in the same sub-cluster based on sentiment analysis may have similar experiences. "Learning style" describes a learner's preferred way of learning, and learners in the same sub-cluster based on learning style may share similar activity preferences. "Social behavior" refers to how learners interact with others during learning, and learners in the same sub-cluster based on social behavior may have similar approaches to working with others, which can affect their learning outcomes.

5 Conclusion

In conclusion, the clustering program used in this study involves a two-step process that utilizes k-means clustering to divide the event log into Pass and Fail clusters, and further divide each cluster into sub-clusters based on learner features. The program requires inputs such as the event log, the number of clusters, the number of sub-clusters, and the clustering parameter initialized by Learning Results.

The output of the program is a set of sub-clusters containing learners with similar characteristics and learning behaviors, which can be used to gain insights into the factors contributing to successful learning outcomes and inform instructional design and support interventions.

It should be noted that this study is currently in the proposal stage, and further evaluations will need to be carried out to determine the effectiveness of the program. While the proposed clustering program has the potential to provide valuable insights into learner behavior and inform instructional design, its efficacy in improving learning outcomes remains to be determined through rigorous evaluation.

References

[1] A. Kao and S. R. Poteet, *Natural Language Processing and Text Mining*. Springer Science & Business Media, 2007.

[2] “Text Mining: Techniques, Applications and Issues - ProQuest.” https://www.proquest.com/openview/86b5831a74364ad4b36255cc0f697c52/1?pq-origsite=gscholar&cbl=5444811 (accessed Mar. 12, 2023).

[3] H. A. Alamri, S. Watson, and W. Watson, “Learning Technology Models that Support Personalization within Blended Learning Environments in Higher Education,” *TechTrends*, vol. 65, no. 1, pp. 62–78, Jan. 2021, doi: 10.1007/s11528-020-00530-3.

[4] R. Burton *et al.*, “Vers une typologie des dispositifs hybrides de formation en enseignement supérieur,” *Distances et savoirs*, vol. 9, no. 1, pp. 69–96, 2011.

[5] “Preservice teachers’ Web 2.0 experiences and perceptions on Web 2.0 as a personal learning environment | SpringerLink.” https://link.springer.com/article/10.1007/s12528-019-09227-w (accessed Mar. 12, 2023).

[6] A. Partington, “Personalised Learning for the Student-Consumer,” *Frontiers in Education*, vol. 5, 2020, Accessed: Mar. 12, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/feduc.2020.529628

[7] D. W. Otter, J. R. Medina, and J. K. Kalita, “A Survey of the Usages of Deep Learning for Natural Language Processing,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604–624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.

[8] "Uncovering themes in personalized learning: Using natural language processing to analyze school interviews: Journal of Research on Technology in Education: Vol 52, No 3." https://www.tandfonline.com/doi/abs/10.1080/15391523.2020.1752337 (accessed Mar. 12, 2023).