Automatic Classification for Cognitive Engagement in Online Discussion Forums: Text Mining and Machine Learning Approach

Hind Hayati^(✉), Mohammed Khalidi Idrissi, and Samir Bennani

Ecole Mohammadia d'Ingénieurs, Université Mohammed V, Rabat, Morocco
hayati.hind@gmail.com

Abstract. For effective learning, students must set learning objectives and adopt the ad hoc cognitive behavior to achieve them. Our research work aims to ensure good scaffolding by offering tutors the opportunity to observe learners' cognitive behaviors, especially their cognitive engagement. In this respect, we propose in the present work an automatic system for classifying learners according to their levels of cognitive engagement. To this end, we focus on the analysis of social interactions within online discussion forums. Hence, the proposed system has two main steps: 1/Learners' vector construction and 2/SVM-based classifier. The results show the efficiency of the proposed system with an accuracy = 0.9 and a cohen's K = 0.89.

Keywords: Text Mining · Machine Learning · Online discussion forum · SVM · Cognitive engagement · Doc2vec · NLP

1 Introduction

The main objective of e-learning platforms is to change the traditional framework of education and make necessary improvements to teaching methods for better learning. However, e-learning remains a complex learning environment in which the learner feels autonomous, isolated, and responsible for his/her educational experience. Therefore, the learner must exhibit enough engagement to counterbalance any other factors resistant to his/her learning. There are three types of engagement: emotional, behavioral, and cognitive. In our work, we toggle the cognitive dimension of engagement since it reveals learners' reflexing and critical thinking. However, the latent nature of engagement and the lack of direct interaction between students and tutors make the prediction of engagement level difficult and challenging (Aleven 2010). Therefore, we focus on the online discussion forum as a tool of asynchronous communication which fosters social interaction.

An online discussion forum is a tool that allows free communication between different participants at any time by keeping track of the various exchanges. Given the degree of learner autonomy during online learning, Larkin-Hein (Larkin-Hein 2001) finds that discussions forum represent a promising way to both achieve emotional attachment and acquire an effective role in the program. Althaus (Althaus 1997) adds that learners learn better through their participation in online discussions because they

are placed in a socio-intellectual environment that encourages active participation, reflection, and equality among different learners.

In our work, we attempt to explore learners' transcripts in order to extract features revealing their cognitive behavior and more especially their level of cognitive engagement. To do that, we propose to automatically classify learners according to cognitive engagement levels based on their social interaction by combining both Text Mining and Machine Learning techniques. We can distinguish four levels of cognitive engagement: **Passive**, **active**, **constructive**, and **interactive**.

2 Detecting Learners' Level of Cognitive Engagement

We can't talk about effective and efficient learning without addressing learners' cognitive behavior, particularly their cognitive engagement. This latter reflects the quality and degree of mental effort that a learner can spend during the learning process. Therefore, our main objective is to determine the level of learners' cognitive engagement from their social interactions within discussion forums.

According to ICAP Framework (Chi and Wylie 2014), we can distinguish four levels related to cognitive engagement, namely: **Passive**: for a learner who simply receives the information without analyzing it, interpreting it, or even reacting to it. **Active**: for those who can understand the text, summarize it and focus on what they are learning. **Constructive**: the learner becomes productive and can generate and produce new ideas and construct knowledge. Finally, the **Interactive**, for whom can debate with peers and defend his/her ideas. To automatically classify learners to the four levels above we have two essential phases:

2.1 Learners' Vector Construction

This step is based on feature extraction to model learners and construct vectors. In our system, a learner can be detected by his/her messages categorized according to the cognitive presence phases (Hayati et al. 2019) as well as traces of his/her social interaction within the platform specifically the discussion forum. Therefore, we have two types of attributes:

The Cognitive Presence categorized messages:

For each learner we calculate

- **TE**: number of messages belonging to the Triggering Event phase.
- **EX**: number of messages belonging to the Exploration phase.
- **Int**: number of messages belonging to the Integration phase.
- **Res**: number of messages belonging to the Resolution phase.

The social interaction features:

- **Add_post**: number of added posts.
- **Discussion_view**: number of learner's consultation of the discussions.
- **Thread_count**: number of discussions initiated by the learner.
- **Nbr_peer_interaction**: number of peers interacting with the learner.

- **Nbr_vote**: number of votes collected by the learner.
- **Time_spent**: time spent by the learner in the online forum.

Thus for a learner i we have the vector

$$\vec{A}_i(\text{TE}, \text{Ex}, \text{Int}, \text{Res}, \text{Add}_{\text{post}}, \text{Discussion}_{\text{view}}, \text{Thread}_{\text{count}}, \text{Nbr}_{\text{peer}_{\text{interaction}}}, \text{Nbr}_{\text{vote}}, \text{Time}_{\text{spent}})$$

2.2 SVM-Based Classifier

This phase relies on the use of SVM as a Machine Learning algorithm for classifying learners as per the four levels of cognitive engagement.

SVM was originally designed for binary classification. Yet, several studies have studied the case of multi-class classification, either by combining binary classifications or by considering all classes at once (Mayoraz and Alpaydm 1999) (Hsu and Lin, n.d.). Indeed, there are two essential approaches, namely “one-vs-one” (OVO) and “one-vs-all” (OVA). OVO consists of definitions for each pair of classes a specific classifier, so, if we have k classes OVO method constructs $k(k - 1)/2$ classifiers. OVA hinges on constructing for each class a classifier that separates its points from all the others. In fact, if we have k classes OVA approach constructs k classifier.

3 Test and Results

3.1 Data Set Description

To test our system we used data from discussion forum samples of different courses in software engineering. Therefore, to classify learners’, we construct a database with all the calculated features whereupon two experts coded according to the four levels of cognitive engagement. The inter-rater agreement was good: percent agreement = 87%. Our data is balanced.

3.2 Training and Testing Phases

After constructing learners-vectors and coding them to the four levels we start the training phase of our SVM classifier after what we will test it and compare in Table 1 accuracy results (classification accuracy, cohen’s K, recall precision and f1 score) for the two approaches OVO and OVA.

From the obtained results we can see that the best choice in our context is the OVA approach. To better observe the obtained results, we have detailed the normalized confusion matrix (Fig. 1).

Table 1. Accuracy results for OVA and OVO approaches

	OVA	OVO
F1 Score	0.90	0.84
Cohen's K	0.86	0.79
Classification Accuracy	0.90	0.85
Precision Recall	0.95	0.93

Fig. 1. Normalized Confusion Matrix; 1: passive, 2: active, 3: constructive, 4: interactive

4 General Conclusion

This work sought to explore learners' cognitive engagement within online discussion forums. This later represents a socio-constructivist environment that encourages higher-order thinking behaviors. In fact, this type of asynchronous communication foster conducts like being socially interactive and adding new knowledge constructively. Regarding the literature review, there are four levels of cognitive engagement: Passive, active, constructive, and interactive.

Our research presents a new automated system to predict learners' cognitive engagement while examining their social interactions in online discussion forums and using Text Mining & Machine Learning techniques. These two approaches propose interesting methods for preprocessing text, analyzing data, and discovering knowledge.

Based on a corpus of posts extracted from learners' participation within courses in software engineering offered through an online learning platform, we explore whether the learner is a passive, active, constructive, or interactive participant. The achieved results have demonstrated interesting precision as classification accuracy = 0.9 and Cohen Kappa = 0.89, which shows that the proposed system is very effective with an almost perfect agreement.

Nevertheless, like any other research, there are limitations to this work too. Our approach, focus only on learners' posts in online discussion forums to predict their

level of cognitive engagement. Yet, it can be learners who are highly engaged with the course materials even if they never display a good level of cognitive engagement in the discussion forum.

As perspective, we can use our system as an input for the recommended systems. In fact, the reported results can be used to recommend new resources for learners according to their level of engagement.

References

- Aleven, V.: Rule-Based Cognitive Modeling for Intelligent Tutoring Systems, pp. 33–62. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14363-2_3
- Althaus, S.L.: Computer-mediated communication in the university classroom: an experiment with on-line discussions. *Commun. Educ.* **46**(3), 158–174 (1997). <https://doi.org/10.1080/03634529709379088>
- Chi, M.T.H., Wylie, R.: The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* **49**(4), 219–243 (2014). <https://doi.org/10.1080/00461520.2014.965823>
- Hayati, H., Chanaa, A., Khalidi Idrissi, M., Bennani, S.: Doc2Vec & Naïve Bayes: learners' cognitive presence assessment through asynchronous online discussion TQ transcripts. *Int. J. Emerg. Technol. Learn. (IJET)* **14**(08), 70 (2019). <https://doi.org/10.3991/ijet.v14i08.9964>
- Larkin-Hein, T.: On-line discussions: a key to enhancing student motivation and understanding? In: 31st Annual Frontiers in Education Conference. Impact on Engineering and Science Education. Conference Proceedings, F2G-6-F2G-12, 2 May 2001. <https://doi.org/10.1109/FIE.2001.963720>
- Mayoraz, E., Alpaydm, E.: Support Vector Machines for Multi-class Classification (1999). <https://www.cmpe.boun.edu.tr/~ethem/files/papers/iwann99.pdf>