International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Machine Learning for Knowledge Construction in a MOOC Discussion Forum

Yassine Benjelloun Touimi, Abdelladim Hadioui, Nour-eddine El Faddouli, Samir Bennani

Abstract: ***discussion forums are spreadly employed as learning tools in online courses, particularly in the Massive open online course (MOOC). Learners share opinions, express needs, and seek tutoring, and participate in discussions in the online forum. However, learner's workstation generates massive information due to the number of MOOC participants, making it difficult to identify relevant information that can help and answer questions during the MOOC. Identifying and extracting knowledge from a MOOC discussion forum requires learner's engagement in a collaborative and informative learning environment that enables knowledge exchange and information sharing. In this article we offer a new approach to explore forums, interactions and collaboration of learners online, in a knowledge building process, by an extraction framework and presentation of knowledge based on the characteristics of the text written in the learners' messages during the training. Our proposal consists in combining the pretreatment of the natural language by the TF-IDF metric, and the embedding of the words by Word2Vec, and then we will use the machine learning algorithm SVM for a semantic classification according to the analysis interactions model. Thus, we will apply the transformations and pretreatments on the messages posted in the forums by the participants in the MOOC, then the Word2Vec to represent each word as a vector, which will be concatenated to the features of the context TF-IDF. These vectors will form the data input of our Learning SVM machine algorithm, which aims to establish semantic relationships between concepts. The knowledge is then expressed as ontology for a representation of knowledge and an enrichment of our model.***

Keywords: MOOC, Forum, Interactive Analysis Model, TF-IDF, Word2Vec, ontology, Support Vector Machine.

I. INTRODUCTION

Massive open online courses [1] have become popular over the past decade and have delivered broad learning opportunities at the expense of traditional classroom environments.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Yassine benjelloun Touimi*, Department of Computer Science, Mohammadia School of Engineering, University Mohammed V, Morocco. E-mail: Yassine.benjelloune.touimi@gmail.com
Abdeladim Hadioui, Department of Computer Science, Mohammadia School of Engineering, University Mohammed V, Morocco.. E-mail: ahadioui@gmail.com
Nourredine Faddouli, Department of Computer Science, Mohammadia School of Engineering, University Mohammed V, Morocco.. E-mail: n.faddouli@gmail.com
Samir Bennani, Professor and head, Lab in Mohammadia School of Engineering, and Vice President in the University Mohammad V Morocco.

MOOCs have had a significant impact on the characteristics of education, from open source (tools and learning platforms) to open content (open learning and learning content) to the open online learning experience (open online course) valid since the 2000s.

The MOOC approach has become popular and it reaches a broad community of learners, through its openness it can lead a group of learners to study together regardless of their original culture and social background. By comparing with traditional courses, MOOCs cover a very wide scale, and are globally accessible without any restrictions on participation in the session.

The different MOOCs are described in the literature: cMOOCs are based on connectivity and networking, however xMOOCs are based on behaviorist approach.The process of learning in connectivity occurs when the learner feeds his knowledge by making connections with the collective knowledge of the community. These connections are established in a context of conceptual and social / external interactions. Connectivists state that learning is not simply a transfer of knowledge from teacher to learner and does not take place in a single environment, but they state that knowledge is transformed and transferred by the interaction of especially in a Web environment [13]. Among the cMOOCs learning tools include the discussion forums that play a crucial role in the learning process.

Forums are not used consistently in all MOOC conceptions. For example, some MOOCs depict interaction 's in forums, as a complement of the course, while others do not. Nevertheless, as in most online courses, MOOCs generally rely on the discussion board to replace conversations that would normally take place in the presential classroom. A discussion forum (newsgroup, online forum, etc.) is defined as an online means that furnishes students with the ability to submit messages, respond to messages, categorize messages, and view responses to messages. The messages and the answers are organized in threads, so one can envisage a forum like a store to post messages online [3] [4].

Moreover, in a MOOC, the discussion forum is invested by great importance from students, as it is the only channel that provides support, to answer questions, ask the tutor to intervene and connect with colleagues. As a result, MOOC forums tend to offer a rich and comprehensive process of learners' learning processes, interactions and discussions. However, the huge volume of data produced during activities in the MOOC forums, makes that information ambiguous,

and posts in the forum difficult to effectively sort and semantically, which complicates the task of the tutor to answer questions.

MOOCs face the challenge of exchanging and investigating in a discussion forum, extracting relevant data, and sharing among learners. This task is greedy and tedious in resources when reading, coding and analyzing this textual data. In sum, learners do not benefit from the relevant knowledge generated during forum interactions by posts and comments.

In terms of qualitative research, the volume of interactions in MOOC forums makes reading, coding, textual data analysis tedious, which implies significant latency, and resource intensity.

However, the occasion for collaborative learning in the online discussion is revealed by the sharing of constructive activities of shared knowledge, due to interactions that facilitate involvement in the learning of their peers in the group, possess experiences and different origins.

The collaborative forum provides a common platform for students to share knowledge that are being in construction and reflection [17]. In addition, the asynchronous format of communication is recognized as an effective tool for creating a criticizing community, in which participants formulate their conceptions, and exchange their ideas and evaluate the beliefs and practices of other learners [18].

Peer student interactions forms a potent tool that provides feedback and interaction to help learners in knowledge building processes. Students have different profiles (academic, cultural, nationalities), which vary the views in the online discussion forum. In addition, the forums provide the opportunity to produce the scaffolding of thought [19], which develops a collaborative and individual reflection of learners. Our approach in this work is to develop an automatic framework based on the Learning machine, in order to extract and represent the knowledge of the discussion forums in the MOOC environment.

In the following section we will discuss the state of the art of the various works carried out in the field of MOOC discussion forums.

In the 3rd chapter we will apply a pretreatment performed for the corpus of the text written in the threads messages in the forum. The 4th chapter presents our framework, and its components and the machine learning algorithms used for the classification as well as the evaluation of our classifier. Then we will build ontology for the representation of the knowledge extracted in chapter 6. The last section will put light on work in progress and perspectives.

II. STATE OF ART

With technological progress, new forms of group work have emerged, particularly in the field of distance learning. E-learning has improved teaching conditions by overcoming the spatial and temporal barrier. The e-learning, has several tools to improve the conditions of the course such as discussion forums, emailing lists, wikipedias ... .etc [5].

With the emergence of E-learning, researchers have examined the variable aspects of asynchronous discussion forums, focusing on research areas such as the role of the tutor in online forums [6], forum structures, the use of the social network to better understand patterns of interaction [7], and the impact of forums on the learning process [8].

However, the tutor's role remains paramount in an LMS platform [6], as a facilitator who helps learners to choose their course, and allows them to express themselves in online discussions and he still plays the role of moderator that synthesizes, critiques and structures the content, while managing the deadlines to complete the activities.

In addition, the tutor can still play the role of expert and bring his pedagogical experience and emotional support that is necessary to avoid abandonment during learning [31]. But the feeling of isolation that the learner feels is very common and is an important factor among many that generates its abundance in online education.

In order to free the learner from this feeling of isolation [9] which is reflected in the abundance of distance education, the tutor is then called upon to play a very important role and to ensure a good support for the learners to help them to be motivate and learn more effectively [32]. In addition, MOOCs allow geographically dispersed groups to collaborate and learn autonomously, which accentuates the phenomenon of isolation and disconnection similar to those experienced in the distance learning environment.

In addition, MOOCs allow geographically dispersed groups to collaborate and learn autonomously, which accentuates the phenomenon of isolation and disconnection similar to those experienced in the distance learning environment.

Otherwise, students must also ensure a reciprocal technical arrangement to access tools, materials and learning activities. The feeling of abundance is identified as related to the feelings expressed by the learners' posts in the forums based on the analysis of the feelings of the discussions in the MOOC forums [11].

while several studies have been conducted on discussion forums, few of them have dealt with the specificity of MOOCs discussion forums. The analysis of the MOOC forums, showed the primary role of the discussion forums as a complement to the course, and allows to analyze the interactions of the learners to classify the active and inactive learners, which explains the phenomenon of abundant and retention [12].

In a MOOC context, previous work [14] showed the manner a forum should be implemented, but these studies were limited to the presence of the learner in a learning space, and not to contributions to the exchange in the forum to achieve the objectives of the course.

All these attending and monitoring needs of the MOOC [10], show the importance of a tool for communication and flexible information exchange between the tutor and the learners on the one hand and between the learners on the other hand. But the number of messages is proportional to the number of learners. Knowing that the number of learners registered in a MOOC has reached significant numbers, the exploitation of massive data in the exchanged messages becomes impracticable.

The learners in MOOC forums freely express their messages, and have no uniform syntactic representation or semantic linking the different information in the posted content [33],

Which requires that functionalities be included in the asynchronous communication tools of the MOOCs, to facilitate the search for knowledge semantically. Due to these constraints related to an online MOOC learning environment, this article is essentially aimed at giving tutors the opportunity to understand, and to analyze learners' knowledge-seeking behavior based on cognitive theory.

Specifically, we aim to make up a framework, based learning theory. This framework will serve to identify the cognitive level of students based on the information released in the posted messages.

In the state of the art of MOOCs, the tutor noted that the discussion forum formed a means of communication during the course, and was a source of many discussions. The forum is considered a tool for expression, explication, and relation, which allows participants to share their homework's, thoughts, and remarks.

The tutor also noted that it was motivating to see the number of messages and answers increase by time. In this perspective, we propose a framework for knowledge construction based on social interactions in a MOOC discussion forum.

In previous research, discussion forums have been considered as a means of building knowledge, using the pedagogical taxonomic approach to model messages and identify types of interactions in conversations. The messages are analyzed based on Bloom's taxonomy level [34] to categorize the scientific discourse. The results of this research show that interactions within MOOC discussion forums are a learning process at particular levels of cognitive learning.

The results also imply that different types of forum interactions have characteristics relevant to particular learning levels, and that the number of higher level cognitive learning incidents increases as the course progresses. The results of this research show that interactions within MOOC discussion forums are a unique learning process at particular levels of cognitive learning.

In this research, we investigate the application of cognitive learning theory to forum posts (threads, messages, and comments) as instances of the socioconstructivism process [35] and learning.

We define the discussion forum exchange process that is realized on three levels:

Thread: the message containing the initial topic of the forum.

Post: the message in the Thread.

How: the message that answers a post or another how.

Specifically, we seek to develop a model to identify the cognitive learning level of a data exchange, when interacting in a discussion forum. The exchange of information in a discussion area is considered to be a cognitive learning process [16].

One approach is to consider the interpretation of information exchanges in a discussion forum as part of the learning process [15], so we will define a framework based on the methodology of cognitive learning according to the needs of the learner information, and attributes inherent to the knowledge extraction process.

the empirical studies have shown that participants use online discussion forums such as a knowledge sharing medium [17], so the interactions in a forum is a form of knowledge sharing through online conversation, what qualifies it learning process.

In this context, we will use the approach of the interaction analysis model [20], which allows analyzing the knowledge building processes, in order to identify the key terms of each of the five levels of knowledge.

III. INTERACTION ANALYSIS MODEL

The content analysis is defined as a research methodology that relies on procedures to draw valid conclusions from a text [36]. The interaction analysis model [20] has examined the constructivist knowledge creation phases, while the categories of collaborative behavior [21] present situations of collaborative learning. The interaction analysis model examines the transition between critical thinking phases (critical thinking) such as negotiation, to illustrate the construction of knowledge.

The Interaction Analysis Model was developed to analyze the process of constructing participants' knowledge in a learning environment that can reach a higher level of critical thinking across the different phases of interactions with their peers.

In other words, content analysis studies allow observations on, cognitive and social interactions among learners, levels of participation, collaborative activity among learners, and the level of knowledge building among the learners.

Several researches have been carried out in interaction analyzes, such as those conducted in online asynchronous project pedagogy environment for college students [22].

According to [20], the notes written by students in the course are classified in Phase I and Phase II, while students are considered to have low mental functions, and notes listed in Phase III, Phase IV, and Phase V represent students with high mental functions.

This research revealed that almost 86% of the notes [20] [22] residing in Phases I and II of the students in the course, are qualified low levels depending on the grid coding scheme. In the grid, the five phases of knowledge construction are necessary for the co-construction of knowledge in the collaborative learning, therefore a very low percentage of high levels of notes written by learners will involve an unsatisfactory quality of learners' knowledge co-construction.

In this research, we will classify the messages in relation to the five phases, by doing a similarity mapping with a set of keywords characterizing the phases of the interaction analysis model, and the words composing the text of the messages posted in the asynchronous discussion forum[23].

As an example, for phase 1 of haring / comparisons of information, we find keywords such as, 'what', 'who', 'is what', 'in my opinion', 'I am agree', 'good idea', 'more', 'besides', 'add', etc. .....

For phase II of dissonance, we have the key words: 'I do not agree', 'and if we did that', 'idea'.

The phase III concerns the negotiation and co-construction of new knowledge:

'We can do it', 'we doubt', 'we are on', 'can be', 'it's ok', etc.

We find in phase VI the tests and attempts to build new knowledge with keywords like 'I think', 'sample', 'data', 'application.

In phase V, we have the declarations and applications of new knowledge built with the terms: 'synthesis', 'association', 'combination', 'formula', 'equation', 'calculation', 'approach', 'new way'.

Based on the previous work [29] [30], for each phase of the analytical model, we will build a dictionary of keywords, which will be the subject of reference for the comparison of similarity with the words of the text of the messages posted. Key words will be represented by Word2Vec vectors including the skip-gram algorithm we will have a presentation of each word by vector. This vector presentation will make it easier to measure the similarity of words in posted messages.

TABLE I: Categories and indicators of the interaction Analysis model values

Phase	Categories	Indicators
1	Shares and comparisons	A. statement of opinions B. statement of agreement C. Examples of Corroborations D. Request and answer questions to clarify the details of the statements E. Definition, description, and identification of a problem
2	Dissonance	A. Identifying and stating areas of disagreement B. Asking and answering questions to clarify the source and extent of disagreement C. Restating the participant's position, and possibly advancing arguments or considerations in its support by references to the participant's experience, literature, formal data collected or proposal of relevant metaphor or analogy to illustrate point of view
3	Negotiation & Co-Construction	A. Negotiation or clarification of the meaning of terms B. Negotiation of the relative weight to be assigned to arguments C. Identification of areas of agreement or overlap among conflicting concepts D. Proposal and negotiation of new statements embodying compromise, co-construction E. Proposals integrating or accommodating metaphors or analogies
4	Testing Tentative Constructions	A. Testing the proposed synthesis against 'received fact' as shared by the participants and/or their culture B. Testing against existing cognitive schema C. Testing against personal experience D. Testing against formal data collected E. Testing against contradictory testimony in the literature
5	Statement & Application of Newly Constructed Knowledge	A. Summarization of agreement(s) B. Applications of new knowledge C. Metacognitive statements by the participants illustrating their (cognitive schema) has changed as a result of the interaction

IV. PROPOSED APPROACH

To achieve our main goal, which manifests in building and sharing new knowledge in a discussion forum, as shown in Figure 2, our model is presented by a system composed of seven steps:

- Constitution of the dictionary of the IAM model.
- Pretreatment of the text of the messages.
- Analysis of the content of the messages.
- Extraction of concepts.
- Extraction of semantic relations.
- Construction of ontology.
- Representation of knowledge.

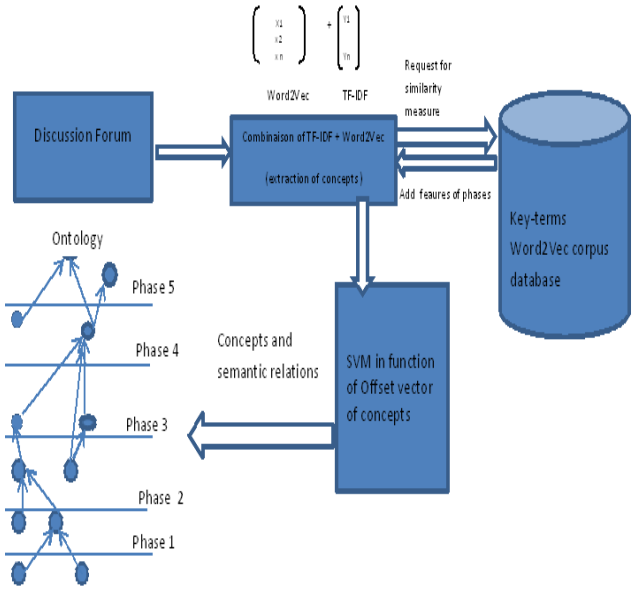

Figure 1: Knowledge building framework in forum discussion

A. Building the IAM Model Dictionary:

Beforehand we built our dictionary of words of the IAM model, collecting all the verbs, words, corresponding to the five phases of the model. The dictionary is stored in a database.

Each word in the dictionary will have a glossary that explains in detail the meaning of the word, and a set of synonyms that facilitates comparison with other words.

This corpus forms the basis of our study, and each word in the dictionary will be presented by a Word2Vec [37], to calculate the similarity with the words of the messages written by the learners.

B. Preprocessing the message data

In this step, the text is cleaned by pretreatment of data collected from messages written in natural language on the MOOC course website. To clean the data, we use the Natural Language Toolkit (NLTK) [38], to process data from text. NLTK offer a widely used dictionary which is the WordNet. Subsequently, the text processing will be done by libraries, for the creation of tokens, the creation of links, the display and the analysis (Tokenization, stemming, tagging, and parsing).

For cleaning and pretreatment of data, we perform the following steps for each message in the discussion forum: remove tags, remove punctuation, remove stop words, put words in lowercase, remove non-ASCII code words , apply the stemming, on the other hand we keep all the URLs published in the messages.

The terms used in the written communication are indicators on the behavior of the learner, his feelings on a subject, on which topics he is interested, his organization, his analysis of the information. The content analysis approach provides views on the structure of the message in a MOOC discussion forum.

In addition to the thematic classification of forum messages, content classification is used as the basis for semantic analysis. In the content analysis approach, we use the TF-IDF metric [39].

C. Analyzing the content of messages

The TF-IDF metric (Term-Frequency - Inverse Document Frequency) [39] is based on the similarity indicator of the inverse of the frequency of the document, which means the inverse of the proportion of document that contains the term candidate at the logarithmic scale. In terms of calculating process weights, post messages must be analyzed and constructed in key term vectors.

The key-term vectors will be used to calculate the terms weights $(W_{i,j})$ using the TF-IDF weight metric (frequency of terms / frequency of the document inverse).

The following formula measure the weight Wi,j of the terms i in the posted messages:

$$W_{i,j} = \frac{f_{i,j}}{max_z f_{z,j}} * \log(\frac{D}{d_i})$$

$f_{i,j}$: frequency of term i in the posted messages j

$max_z \ f_{z,j}$: maximum frequency among all z keywords that appear in the message posted j.

D : is the total number of messages posted that can be recommended by a user, and di is the number of messages posted that contains the term **i**.

After calculating all the terms weights of all the terms in a posted message, the term weight vector i will be:

$$W_i = [W_{i,1}, W_{i,2}, W_{i,3}, \ldots\ldots\ldots\ldots, W_{i,n}\ ];$$
$$j = 1, \ldots . n \ posts$$

After endowing each word with a weight vector, we use the approach based on the prediction Word2Vec to represent words semantically. Word2Vec is a prediction-based approach that practices on a large corpus of text [40].

Word2Vec is a single layer neural network [40], consisting of two models: skip-gram and CBOW (continious-bag-of-words). In the case of CBOW, we give a context and we predict the target word, while in skip-gram for a target word we predict its context.

In this article, we will use skip-gram which will take the target word W_i as input, and the output is its context $\{W_1, W_2, \dots W_c\}$ defined in a window of size C.

After training our model by skip-gram, we add to each vector Word2Vec, its vector weight, which enriches the modeling of the word with a new feature. The vector of each word is a new vector enriched semantically.

D. Concepts extraction

During this step, we propose an algorithm for extracting key concepts.

The semantic model of the vector of each word of the posted messages will be the input of our algorithm of extraction of the concepts.

During the construction of the corpus of the project, for each keyword we put definitions in dictionary form. So for each word we create a corresponding synonym set.

Our system can automatically classify messages based on phases knowledge, by indicators of the interactions model analysis of learners in the online discussion forum using Word2Vec method. A vector space model involves two steps: the creation of term weights and cosine similarity.

The algorithm for extracting the key concepts of the discussion forum parameterized by the cognitive phases of the IAM model is detailed below.

ALGORITHM 1: Extraction Algorithm

{Sd}= Set of IAM model terms //Extracted by the TF-IDF measure

{T}=Set of terms in the posted messages

CC={ Set of concepts extracted in accordance with the phase}

Syn_i={Set of synonyms}

For each term ti in T

For each Sdi in Sd

Syn_i=find_SynonymSet(ti, Sdi) //we use Word2Vec and by comparing the vectors //by 4 similarity with value 1

If Syn_i!={}

CC={ti,Syn_i, phase} // addition of the corresponding phase to the vector of the

//candidate concept which increases the semantics of the concept

end if

{T}={T}-Syn_i

End for

End for

E. Algorithm for extracting semantic relations

Earlier work extends and refines the semantic hierarchy, which is built manually using other resources (e.g., Wikipedia) [24]. However coverage is limited by the scope of resources.

Several other works depend on lexical patterns, which suffer from deficiencies because patterns cover only a few small proportions of complex linguistic circumstances [25].

In addition, the similar distribution methods [26] are based on the assumption that the term can only be used in contexts where its hypernyms and hyponyms are used.

This logic is effective in the case of extraction of named entities. However it is not rational in the case of massive data of web mining with broad semantics. In short, all these methods do not use words effectively in a semantic way.

As the main components of ontologies, semantic hierarchies were studied by many researchers. Concept hierarchies are based on manually built semantic resources such as WordNet [27]. Some hierarchies have solid structures of high precision, but their gate is limited to grain concepts (e.g. "Ranunculaceae" is not included in WordNet).

We have made the same observations, that almost half of the hypernyms-hyponyms relations are absent in several languages such as in the Chinese language which suffers from several deficiencies in its semantic thesaurus.

Therefore, we will have a huge need for resources to add the resources manually built. For example the famous ontology YAGO [28] links the categories in Wikipedia to WordNet, but the scope remains limited to Wikipedia.

All these limitations, in the extraction of semantic relations, led us to embed all words by Word2Vec. The words embedded [37] retain interesting linguistic regularities, capturing a considerable amount of syntactic / semantic relations.

Taking the most salient example: v (king) - v (queen) = v (man) - v (woman), indicating that the offset vector present the common semantic relation between two pairs of words.
We observe that these same properties apply to some hypernyms-hyponyms relations.

The embedded offset vectors are computed between a sample of word pairs of hypertext-hyponyms randomly selected, and we measure their similarity [41]: v (shrimp) - v (prawn) = v (fish) - v (goldfish).

We notice that the word can be linked to its hypernyms by using the offset vectors of words embedded. However, the difference between "carpenter" and "laborer" is distant from the offset between "gold fish" and "fish" , "Indicating that hypernym and hyponym relationships, are more complicated than offset processing. These remarks are also valid for all types of relationships: meronyms, synonyms, antonyms,..etc .

Our proposal in this article consists in transforming the words of the posted messages, enriched by the TF-IDF features, and the level of the construction phase of the knowledge, and then we will make use of the machine learning method of SVM [42], in order to retrieve the semantic relations between the concepts extracted from the text.

SVM [42] is a classification method that is particularly well suited for processing massive data such as text and images.

The support vector machines (or separators wide margin) are a set of learning techniques to solve problems of discrimination, that is to say, to decide which class a sample, or regression, and to predict the numerical value of a variable. In our case study we will extend our relations to include the relations meronym, synonym, homonym, etc.

Within the SVM, the inclusion of non-linearities in the model is by the introduction of non-linear kernels, which is a projection to a new dimension, however the use of kernels does not fundamentally alter the nature of SVM [43].

This kernel is a function F that associates with any pair of observations (X, Y) a measure of their "reciprocal influence" calculated through their correlation or their distance. Typical examples of kernel are polynomial kernel.

In our case study, the SVM method will take as a core the measure of similarity between the enriched word vectors, by the method of “Extended Gloss Overlaps as a Measure of Semantic Relatedness” [44].

Concepts are commonly represented by a dictionary of the meanings of words. Each of them has a definition or glossary that briefly describes their meaning. Our measure determines the level of relativity between two concepts by counting the number of shared words (overlap) in the words describing the meaning of the concepts, as well as in the glossaries of words that are related to these concepts according to the dictionary. These related concepts are explicitly encoded in WordNet relationships, but can be found in any dictionary by synonyms, antonyms, references.

This measure is based on a function that accepts two glossaries of the candidate words as input, finds the overlapping sentences of the glossary between the terms and returns a score.

As an illustration, we suppose that a set of relations S = {gloss, hypernym, hyponym}, compose our set of relations in pairs = (gloss, gloss), (hype, hype), (hypo, hypo), (hype, gloss), (gloss, hype)} then the relativities between sets of synonyms is calculated by the following formula:

$$\text{relatedness}\,(A,B) = \sum_{k=0}^{n} \text{score}\,(R1(A), R2(B))$$

$$\begin{aligned} &relatedness(A,B) \\ &= score(gloss(A), gloss(B)) + score(hype(A), hype(B)) \\ &+ score(hypo(A), hypo(B)) + score(hype(A), gloss(B)) \\ &+ score(gloss(A), hype(B)) \end{aligned}$$

The function of relatedness [44] accepts as input two glossaries, find the overlapping sentences between them and return a score.

Since in our case we are with several types of relations and, the distance between 2 vectors word2vec is between -1 and 1, we opt for the creation of a multitude of class, so we will have k classifiers to classify the word pairs.

F. Multi-class SVM for classification of semantic relations

The methods of the multi-class support vector machines [45] reduce multi-class problems to a composition of several two-way hyper planes to draw the decision boundaries between the different classes [11]. These methods decompose the set of examples in several subsets each representing a binary classification problem. For each problem, a separation hyper plane is determined by the binary SVM method.

Multi-class SVMs are distinguished by two approaches: One-on-One (1vs1) and One-on-All (1vsR). In our case studies we opted for the approach of One against all (1Vs 1). The One-on-One approach is a special case of the decomposition methods populated by Dietterich et al. [29] to solve problems with several classes. This approach requires the construction of $K * (K-1)/2$ SVM each separating a pair of classes (i,j) from those existing. During classification, an input vector x is presented to all built classifiers. The output of each SVM provides a partial vote concerning only the pair of classes (W_i, W_j). Considering that each SVM calculates an estimate $P_{i,j}$ of the probability:

$$P_{i,j} = P(x \subseteq W \,|x, x \in W_i \cup W_j)$$

Then the simplest classification rule can be written:

$$argmax_{1 \le i \le K} \sum_{j \# i} [P_{i,j} > 0.5]$$

Thus, we will build our ontology structure based on the scores of the measurement distances between the different offset vectors of the word pairs.

TABLE II: Classification of relations between vectors offset based on SVM multiclass

Concept sCandidats	Distance with offset C1-C3	Distance Between C1-C4			Distance between Cn-Cm	Class of relations
Offset C1-C2						hypernym
Offset C2-C3						Meronym
Offset C3-C4						
Offset Ck-Cl						Synonym

Avec $i = 1, \ldots\ldots, n-1$.

According to Table 2, we construct for each relationship type classifier (hypernym, hyponym, meronym, etc ...).

G. ontology enrichment by new concepts

In this phase, we enrich our ontology, by designing an enrichment algorithm based on multi-class SVM and the relation of relativity between concepts. We introduce enrichment algorithm by adding new concepts:

ALGORITHM 2: Extraction Algorithm

Input: 1 sentence , Word_target W

Output= {list of concepts and relations}

W ← tokenization + elimination of word space + POS tagging + lemmatization
{list of words}

For each Wi

Ci ← set of the context of the word with a window of 4 words surrounding

Vector(ci) ← Word2Vec (context of Ci) est {C1,C2,C3, …}

Result_concept_link={empty}

Link← concept

if Wi exist in {onotlogy} so exit

Score= get_Relatedness by SVM multiclass

Result← {concept,link} + Result

Reurn Result
End for
End

V. RESULTS AND PERFORMANCE EVALUATION

In order to make a best experimentation of our automatic extraction system, we will use a MOOC forum of our system in Mohammadia School of Engineering.

The Input of our system is composed of messages and threads exchanged during the classes of a course, which have been treated initially in first time.

In the phase of pre-treatment, we will reach a file with the candidate concepts.

Our main purpose is the creation of a knowledge ontology based on four levels. Hence our automatic extraction system will produce a key concepts and semantic relations between them.

The automatic extraction system considers the five phases, as rules for the qualification of the actor's knowledges. The rules are the five phases of the analytical Model of interactions.

Subsequently, we will define tree category of actors based on their knowledge classification by the system. Each category of learners depends on the phases of knowledge reached.

Hence we can classify the messages according to the phases defined in the interaction analyses model:

TABLE III: classification of the messages

Category	Phases	Number of messages
Category 1: Advanced	Phase 1	4000
	Phase 2	2000
	Phase 3	2000
	Phase 4	1000
	Phase 5	200
Category 2: intermediate	Phase 1	5000
	Phase 2	2000
	Phase 3	1000
	Phase 4	100
Category 3: beginners	Phase 1	2000
	Phase 2	1000
	Phase 3	100

According to table III, the majority of learners reach the tree first phases as seen by the number of messages classified in those categories: Shares and Comparisons, Dissonance, Negotiation & Co-Construction.

The number of messages classified in the phase 1, 2, 3 is 19100 messages posted, and present 93 % of the total messages which is a good indicator of communications between learners and sharing ideas and thoughts.

the numbers of messages classified in phase 4 is 1100 messages which present 5 % of all messages posted in MOOC forum. This small number reflect the poor number of learners in the Testing Tentative Constructions phase. Hence, learners can't test the knowledge constructed in the first three phases.

The same observation for the phase 5, we notice the weak number of messages posted, which present 1% of all messages. We can conclude that a few learners in this MOOC course has construct a new knowledge and assimilate them.

On the other hand, we notice that category of advanced learners sends most messages in the MOOC forum, and reach the fifth phase of construction knowledges process with 45% of messages posted.

Furthermore, the second category is the intermediate learners which they reach the fourth phase, so they have the ability to test the new knowledges in real case.

The percentage of the messages posted by this category is important and positioned around 40%. The rest of the messages belongs to the beginners.

In addition, the advanced and intermediate have the big impacts on the construction knowledge process so they are more influencers than beginners. Hence they reach the phase four and five of the analysis model.

VI. CONCLUSION

In this article we have opted for the use of the interaction analysis model in a MOOC discussion forum, in order to represent the knowledge constructed by the learners.

A discussion forum in a MOOC environment is designed as a complement to the course and not just a tool for sharing and exchanging information. This article discusses the interest of online discussion forums in a MOOC, and analyzes the exchanges based on a message interaction model.

This model classifies the knowledge construction process in five cognitive building levels. Based on this approach, learner messages can be classified using similarity measures, by combining the TF-IDF method, with the Word2Vec vector representation.

Thus an algorithm was proposed, which focused on the extraction of concepts in a learning forum and the classification according to the abstraction levels of the IAM model.

Then we use the method of Machine Learning SVM, for the extraction of the semantic relations between the different concepts, in order to build our ontology by cognitive level.

Then we proposed a second enrichment algorithm of our cognitive ontology by new concepts extracted from the new messages posted.

However, online discussion forums in a MOOC environment generate massive data, which leads us to analyze these data by Big Data techniques [46] such as Map Reduce algorithms [47], SPARK ML [48].

In addition, we will exploit in our next work, the algorithm LSTM [49], for a better optimization of the data.

Another aspect that will be dealt with in our next work, which is very important, is the sentimental analysis of learners [50].

REFERENCES

1. Egloffstein, M., Ifenthaler, D., (2017). Employee Perspectives on MOOCs for Workplace Learning. TechTrends 61, 65–70. https//doi.org/10.1007/s11528-016-0127-3
2. Gautier, J.-M., Gayet, A.,(2017). De la Data aux Big Data: enjeux pour le Marketing client–Illustration à EDF. Statistique et Société 4, 49–53.
3. Kaiser, C. & Bodendorf, F. (2012). "Mining consumer dialog in online forums," Internet Research, 22, 275-297.
4. Prabowo, R., Thelwall, M., Hellsten, I., & Scharnhorst, A. (2008). Evolving debates in online communication: A graph analytical approach, Internet Research, 18, 520-540.
5. Wheeler, Steve; Yeomans, Peter; Wheeler, Dawn, The Good, the Bad and the Wiki: Evaluating Student-Generated Content for Collaborative Learning, British Journal of Educational Technology, v39 n6 p987-995 Nov 2008.
6. Margaret Mazzolini, Sarah Maddison, Computers & Education Volume 40, Number 3, 2003 ISSN 0360-1315 Publisher: Elsevier Ltd
7. Thomas, L 2002, 'Student Retention in Higher Education: The role of institutional habitus', Journal of Education Policy vol. 17, no. 4, pp. 423-432.
8. Hughes, D. L., & Chan, S. P. (2014). Participation in asynchronous online discussion forums does improve student learning of gross anatomy. Anatomical sciences education, 7(1), 71-76
9. Adamopoulos, P. (2013). What Makes a Great MOOC? An Interdisciplinary Analysis of Student Retention in Online Courses. In Proceedings of the 34th International Conference on Information Systems, ICIS. 2013.
10. Denzin, N. K. & Lincoln, Y.S. (Eds). (2003). Strategies of Qualitative Inquiry. London: SAGE Publication
11. Wen, M., Yang, D., Rosè, C.P.: Sentiment analysis in MOOC discussion forums: What does it tell us? In: Proceedings of Educational Data Mining (2014)
12. Wong, J.S., Pursel, B., Divinsky, A. and Jansen, B. J.(2015). An Analysis of MOOC Discussion Forum Interactions from the Most Active Users. 2015 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP15). Washington DC, p. 452-457. 31 Mar.-3 Apr
13. Mehmet.K, Hakan.A,5th World Conference on Learning, Teaching and Educational Leadership (WCLTA 2014), A Theoretical Analysis of Moocs Types From A Perspective of Learning Theories.
14. Kop, R., Fournier, H., & Mak, J. (2011). "A pedagogy of abundance or a pedagogy to support human beings? Participant support on massive open online courses," International Review of Research in Open and Distance Learning, 12, 74 93.
15. Bloom, B. S. & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook 1: Cognitive domain. New York: Longmans.
16. Schmeck, R. R. (1988). Learning strategies and learning styles (perspectives on individual differences. New York:Plenum Press.
17. Wu,D. Hiltz, S.R. Predicting learning from Asynchronous Online Discussions. 2004. http://www.sloanc.
18. Webb, K. & Blond, J. Teacher Knowledge: The relationship between caring and knowing. Teaching.
19. Johnson, D.W. and Johnson, R.T. Cooperation and the use of technology. In D.H. Jonassen (ed). Handbook of research for educational communications and technology. 1996. Simon & Schuster MacMillan.
20. Gunawardena, C. N., Lowe, C. A. and Anderson,T. Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. Journal of Educational Computing Research, Vol. 17, No. 4, 1997, pp. 397-431.
21. Johnson, D.W. and Johnson, R.T. Cooperation and the use of technology. In D.H. Jonassen (ed). Handbook of research for educational communications and technology. 1996. Simon & Schuster MacMillan
22. Zumbach, J., Reimann, P. & Koch, S. (2006). Monitoring students' collaboration in computer mediated collaborative problem-solving: Applied feedback approaches. Journal of Educational Computing Research, 35(4), 399-424
23. A.H. Biriyai 1 , E.V. Thomas, Online Discussion Forum: A Tool for Effective StudentTeacher Interaction, 2014.
24. M.Suchanek,G.Kasneci,G.WeikumYAGO: A Large Ontology from Wikipedia and WordNet, Journal of Web Semantics, Volume 6, Issue 3, September 2008, Pages 203-217.
25. A. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora, COLING, 1992, Volume 2: The 15th International Conference on Computational Linguistics
26. A.Lenci, G.Benotto, Identifying hypernyms in distributional semantic spaces, SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)
27. Miller, WordNet: a lexical database for English, Communications of the ACM, Volume 38 Issue 11, Nov. 1995.
28. Suchanek , YAGO: A Large Ontology from Wikipedia Andwordnet, Journal of Web semantics , 2008.
29. Quek, C. L. (2010). Analysing high school students' participation and interaction in an asynchronous online project-based learning environment. Australasian Journal of Educational Technology, 26(3). https://doi.org/10.14742/ajet.1078.
30. Skinner, Jason. "Interaction Analysis , Synchronous CMC , & a Multi-Modal Unit of Analysis." (2008). -Annual Proceedings-Anaheim: Volume – Citeseer
31. Gütl, Christian, Rocael Hernández, Vanessa Chang and Miguel Morales Chan. "Attrition in MOOC: Lessons Learned from Drop-Out Students." LTEC@KMO (2014).
32. Edmondson, Joanne H., and JoAnna White. "A Tutorial and Counseling Program: Helping Students at Risk of Dropping Out of School." Professional School Counseling, vol. 1, no. 4, 1998, pp. 43–47. JSTOR, www.jstor.org/stable/42731822.
33. Ming Zhou, Bejing (CN); Jizhou Huang, Chongqing (CN); knowledge extraction from online discussion forums,2010.
34. Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company.
35. From Constructivism To Social Constructivism: A Vygotskian Perspective on Teaching and Learning Science. Hodson, Derek; Hodson, Julie , School Science Review, v79 n289 p33-41 Jun 1998.
36. An Overview of Content Analysis Steve Stemler Yale University, Volume 7, Number 17, June, 2001, Practical Assessment, Research and Evaluation.
37. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. word2vec. https://code.google.com/p/word2vec/.
38. Bird, S. (2002). NLTK: The Natural Language Toolkit. ArXiv, cs.CL/0205028.
39. [39] Karen Spärck Jones, « A statistical interpretation of term specificity and its application in retrieval », Journal of Documentation, vol. 28, no 1, 1972, p. 11–21 (DOI 10.1108/eb026526.

40. Martindale, Colin,Cognitive psychology: A neural-network approach,PsycINFO Database Record (c) 2016 APA.
41. R. Fu, J. Guo, B. Qin, W. Che, H. Wang, and T. Liu. , Learning semantic hierarchies via word embeddings, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, page 1199--1209. (2014)
42. Cristianini, Nello; Shawe-Taylor, John,An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press. ISBN 0-521-78019-5. (2000).
43. Fradkin, Dmitriy; Muchnik, Ilya (2006). "Support Vector Machines for classification" (PDF). In Abello, J.; Carmode, G. (eds.). Discrete Methods in Epidemiology. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. 70. pp. 13–20.
44. by Satanjeev Banerjee , Ted Pedersen In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Extended Gloss Overlaps as a Measure of Semantic Relatedness, 2003.
45. Wang Z., Xue X.,Multi-Class Support Vector Machine. In: Ma Y., Guo G. (eds) Support Vector Machines Applications. Springer, Cham, 2014.
46. Elgendy N., Elragal A. (2014) Big Data Analytics: A Literature Review Paper. In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2014. Lecture Notes in Computer Science, vol 8557. Springer, Cham
47. Shim K. (2013) MapReduce Algorithms for Big Data Analysis. In: Madaan A., Kikuchi S., Bhalla S. (eds) Databases in Networked Information Systems. DNIS 2013. Lecture Notes in Computer Science, vol 7813. Springer, Berlin, Heidelberg.
48. Meng, Xiangrui et al. “MLlib: Machine Learning in Apache Spark.” J. Mach. Learn. Res. 17 (2015): 34:1-34:7.
49. Soutner D., Müller L. (2013) Application of LSTM Neural Networks in Language Modelling. In: Habernal I., Matoušek V. (eds) Text, Speech, and Dialogue. TSD 2013. Lecture Notes in Computer Science, vol 8082. Springer, Berlin, Heidelberg
50. Sunghwan Mac Kim,Rafael A Calvo, Sentiment Analysis In Student Experiences of Learning, Conference: Educational Data Mining 2010, The 3rd International Conference on Educational Data Mining,.