Journal of Ambient Intelligence and Humanized Computing (2021) 12:9977–9989
https://doi.org/10.1007/s12652-020-02747-9

ORIGINAL RESEARCH

Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis

Nicola Capuano[1] · Santi Caballé[2] · Jordi Conesa[2] · Antonio Greco[3]

Received: 14 July 2020 / Accepted: 21 November 2020 / Published online: 23 December 2020

Abstract
Massive open online courses (MOOCs) allow students and instructors to discuss through messages posted on a forum. However, the instructors should limit their interaction to the most critical tasks during MOOC delivery so, teacher-led scaffolding activities, such as forum-based support, can be very limited, even impossible in such environments. In addition, students who try to clarify the concepts through such collaborative tools could not receive useful answers, and the lack of interactivity may cause a permanent abandonment of the course. The purpose of this paper is to report the experimental findings obtained evaluating the performance of a text categorization tool capable of detecting the intent, the subject area, the domain topics, the sentiment polarity, and the level of confusion and urgency of a forum post, so that the result may be exploited by instructors to carefully plan their interventions. The proposed approach is based on the application of attention-based hierarchical recurrent neural networks, in which both a recurrent network for word encoding and an attention mechanism for word aggregation at sentence and document levels are used before classification. The integration of the developed classifier inside an existing tool for conversational agents, based on the academically productive talk framework, is also presented as well as the accuracy of the proposed method in the classification of forum posts.

Keywords Massive open online courses · Neural networks · Text mining · Conversational agents

1 Introduction

Since 2007, massive open online courses (MOOCs) are continuously becoming more and more widespread. MOOCs are currently offered by various organizations around the world and can count on millions of students, thousands of courses and hundreds of educational institutions (Siemens 2013). However, in order to maximize the effectiveness of such kind of educational tool, plenty of technological and pedagogical problems still need to be addressed. In particular, due to the imbalance between the number of students and the available instructors, teacher-guided instructional scaffolding may be very limited, since the instructors are not able to dedicate to students the time they would require, and the interaction is limited to a simple one-way transfer of information.

Discussion forums are among the most popular interaction tools offered by MOOCs, often used by students to create a sense of belonging and better understand course topics (Capuano and Caballé 2015). Thus, MOOC forums are massively participated with a great amount of highly unstructured information on student influence and academic progresses, which can hamper rather than encourage the sense of community (Agrawal et al. 2015). Being the instructors completely unable to effectively moderate the forums, students who try to better understand course topics may not receive the support they need, thus favoring low academic satisfaction and high drop-out (Yang et al. 2015).

✉ Nicola Capuano
nicola.capuano@unibas.it

Santi Caballé
scaballe@uoc.edu

Jordi Conesa
jconesac@uoc.edu

Antonio Greco
agreco@unisa.it

[1] School of Engineering, University of Basilicata, Viale Dell'Ateneo Lucano, 10, 85100 Potenza, Italy

[2] Faculty of Computer Science, Multimedia and Telecommunications, Universitat Oberta de Catalunya, Rambla Poblenou, 156, 08018 Barcelona, Spain

[3] Department of Computer and Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, SA, Italy

The automatic analysis of discussion forum posts may help the instructors to plan useful interventions aimed at improving the quality of the course based on students' feedback. Indeed, various studies demonstrate that instructors consider important to understand, at an adequate level of abstraction, what happens in their forums (Hollands and Tirthali 2014; Tomkin and Charlevoix 2014). To meet this increasingly common requirement, novel methods for the automatic categorization of the MOOC forum posts have been recently proposed, as summarized in the next section.

In this paper, a multi-attribute text categorization tool for MOOC forum posts, based on natural language understanding (NLU) methods is presented. Such approach allows to extract useful insights from student posts and classify them with respect to the following six attributes: (i) intent (the aim of the post), (ii) domain (subject area of the post), (iii) topics (learning concepts the post is about within the domain), (iv) sentiment (the affective polarity of the post), (v) confusion (level of confusion expressed by the post) and (vi) urgency (how urgently a reply to the post is required). The overall procedure of this tool follows three sequential steps: first, a forum post is partitioned into a sequence of tokens, each transformed in a vector projecting it into a continuous space; second, the resulting vectors representing post words are combined to obtain a single vector representing the whole post; finally, in the classification step, a class is associated to such vector.

Each predefined attribute for the posts to be classified (namely, intent, domain, topic, sentiment, confusion, urgency) is handled as a separate classification task. The classifier is based on an attention-based hierarchical recurrent neural networks, where a bidirectional recurrent network is used for word encoding, then a local attention mechanism is applied to detect the post's words relevant for the meaning of each sentence and aggregate them in a single vector. The same process is applied at a global level to aggregate the sentence vectors. Eventually, a soft-max layer is applied for normalizing the classifier output into a probability distribution.

The information extracted from the previous classification process could be adopted by instructors for planning their interventions as well as an input for conversational software agents to engage students in guided, constructive, natural language interactions (Caballé and Conesa 2018). The goal of such agents is to promote useful peer discussions, namely argumentation, clarifications and mutual explanations, started and encouraged by the agent intervention and contribution to collaborative activities.

In the most common approaches (Kumar and Rosé 2011), the agent intervention is generally triggered by the detection of specific keywords. The introduction of automatic post categorization can be used to generate more targeted and timely interventions so improving their overall effectiveness (Capuano and Caballé 2019). Following this intuition, we have also explored and described how to integrate the defined categorization method within an existing instructional tool for conversational agents, based on the academically productive talk framework (Demetriadis et al. 2018).

The paper is organized as follows: in the next section the state of the art on MOOC forum post analysis is summarized, and the paper is contextualized within the relevant literature; in Sect. 3, the proposed approach is presented as well as the classifier architecture. In Sect. 4, the experimental results achieved by the proposed approach on the Stanford MOOCPosts dataset (Agrawal et al. 2015) are presented and compared with related methods. Then, in Sect. 5, the integration of the defined text categorization tools within an existing tool for conversational agents is discussed. The paper concludes by summarizing the main ideas and outlining on-going work.

2 Related work

This section analyzes the most recent and significant works related to this research. In particular, the first subsection describes other NLU-based approaches aimed at extracting useful information from MOOC forum posts (some of them are also used for comparison in Sect. 4). Instead, the second subsection describes examples of experimental educational services based on extracted information. Thus, while the first subsection provides the background for the defined approach, the second subsection contextualizes the application to conversational agents described in Sect. 5.

2.1 Analysis of MOOC forum posts

The problem of MOOC forum post analysis is not new although much of the early research on this topic has targeted the use of structured data. For example, in Yang et al. (2015), a classification model using discussion forum behavior and clickstream data is applied to identify posts that express confusion. Until now, few studies have also investigated the use of NLU in analyzing forum posts. For example, Agrawal et al. (2015), defined an instructional approach based on machine learning and NLU which detects confusion in forum posts; the tool was trained with 30,000 anonymized forum posts collected from eleven Web classes.

In Pousada et al. (2017), several machine learning approaches, including Neural Network and Random Forest, were trained and validated to automatically classify new posts from in-class forums into three degrees of emotion, namely positive, negative and neutral. A large training set was prepared and classified by experts according to the emotional moves found in the text. The resulting classified posts, graphically depicted, served as a functional tool for

instructors to identify and measure the overall online classroom feeling as the course goes by, and to help lecturers make the appropriate teaching decisions to turn emotionally negative learning situations into positive.

In Caballé et al. (2009), random forest, support vector machine (SVM) and Naïve Bayes classifiers were trained on posts annotated with urgency and sentiment labels. The learned models were evaluated on posts collected from different courses and demonstrated a good cross-domain classification accuracy. To gain invariance from biases among different courses, in Wei et al. (2017), a long short term memory (LSTM) neural network, pre-trained on the posts collected from a course, has been fine tuned to recognize sentiment on different courses; the achieved results demonstrate that the neural network was able to generalize on the new courses after a transfer learning with few labelled posts.

In Almatrafi et al. (2018), different linguistic feature sets were used in combination with post metadata (i.e., post type, number of reads, likes, etc.) to identify posts requiring urgent response in MOOC discussion forums. Different classification techniques were used on these data including Naïve Bayes, SVM, random forests, AdaBoost and logistic regression. For the same task, more recently, an NLU-based approach has been proposed in Alrajhi et al. (2020) that uses the results of the post classification with respect to other features (like sentiment and confusion) as an additional input to detect the post urgency.

Most of the work reported for post MOOC analysis is based on text categorization. Older applications in this field use generic classifiers such as Naïve Bayes, SVM, logistic regression, etc. for statistical classification of vectorial document representations. Recently, however, the continuous development of deep learning methods, combined with refined document representation techniques, has made it possible to reach greater accuracy for this task. The main deep learning models used for text classification are largely based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Jang et al. 2020). Following the latest developments in the field, an RNN-based approach was adopted in the present work (as detailed in the next section).

2.2 Use of MOOC forum post analysis results

The feedback extracted by automatic tools for MOOC forum posts categorization can be exploited in various ways. In Wen (2014), the trending opinions of the students on course tools, lectures and assessments were automatically extracted from the posts. The analysis of the results demonstrated a strong correlation between the measured sentiment and the number of students who abandoned the course. In Agrawal et al. (2015), information retrieval was applied to suggest video-clips useful to clarify doubts of the students, detected through an automatic system able to extract the level of confusion from a post.

In Guitart and Conesa (2016), the most relevant topics discussed within the forums, and the opinions and problems students had with such topics, were extracted and displayed graphically over a classroom dashboard that provides support for teachers' decision-making. This information allows teachers to easily monitor classrooms to find out in real-time whether there is a peak of difficulty for students in a given topic as well as negative opinions of students on any aspect, providing a way to detect and correct risky situations. Other systems facilitate the teachers' work by classifying the forum messages by several characteristics, such as confusion, urgency and sentiment polarity (Wei et al. 2017), forecasting the success of students taking forum messages into account (Mongkhonvanit et al. 2019), finding out and making explicit the learning resources mentioned within the MOOC messages (An et al. 2019), or identifying potentially urgent student messages that require immediate attention from teachers (Sun et al. 2019).

A research field which can greatly benefit from the analysis of forum posts in MOOCs settings is the area of pedagogical conversational agents (a.k.a., chatbots), which have been designed and successfully developed to support effective interaction and learning in these settings (Caballé and Conesa 2018). In Ferschke et al. (2015a) a pedagogical agent for MOOCs was developed to encourage students reticent to ask for help and to verify that all the requests were answered. As a result, the agent was designed to support help seeking in discussion threads by means of a social recommendation algorithm that selects potential help providers from a pool of student peers who are then invited to participate in the thread, thus increasing the probability that help requests are met with a satisfactory response.

In Ferschke et al. (2015b), a conversational computer agent was designed to support chat activities in a MOOC context. It facilitates the formation of ad-hoc study groups, where students are matched and then taken to a chat room where they can work on a synchronous collaboration basis supported by the agent, which appears as a regular participant in the chat. Therefore, conversational agents that support chat-based small-group learning activities appear to have a direct application in MOOCs, thus providing large cohorts of students with the pedagogical benefits of synchronous collaboration (Dyke et al. 2013). Eventually, such agents can increase students' engagement and minimize dropout rates while amplifying the tremendous support resources that students can offer to each other by themselves (Fersche et al. 2015a).

Targeting the research topics outlined in this section, this paper extends and improves the work described in Capuano and Caballé (2019). In particular, the enhancement of our classifier architecture and its evaluation are presented in

the next two sections. Finally, a practical application to real world on how to use the information captured by the classifier within the academically productive talk framework for conversational agents is provided.

3 The defined approach

The aim of text categorization (TC) is the assignment of specific categories to free-text documents after a careful analysis of their content. Machine learning techniques can be very useful to automate this task (Manning et al. 2008). Such methods allow to learn a model which can predict the categories of arbitrary documents, thanks to a supervised learning carried out with a training set of labelled documents. Being $D = \{d_1, \cdots, d_n\}$ a set of documents (posts) to be classified and $C = \{c_1, \cdots, c_m\}$ a set of classes, the goal of TC is the learning of a classification function Φ defined as follows (Sebastiani 2002):

$$\Phi : D \times C \rightarrow \{T, F\} | (d_i, c_j) \mapsto \Phi(d_i, c_j) \tag{1}$$

which assigns the value T (true) if $d_i \in D$ is classified in $c_j \in C$, F (false) otherwise.

In the proposed TC approach, each attribute (intent, domain, topic, sentiment, confusion, urgency) is handled as a separate classification task. In a first pre-processing step, the text composing a forum post is divided into a sequence of tokens (words), each transformed in a vector projecting it into a continuous space. This process, named word vectorization, is described in the first sub-section. Then, vectors representing post words are combined to obtain a single vector representing the whole post and a class is associated to such vector. This step is performed with an attention-based hierarchical recurrent neural networks as described in the second sub-section.

3.1 Word vectorization

TC algorithms represent the documents with a vector of attribute values, belonging to a fixed common set of attributes; the number of elements in the vector is the same for each document (Cichosz 2019). Among the possible representations, the most widely used is the bag of words (BOW), which encodes each document d in a vector $\boldsymbol{d} = (w_1, \ldots w_{|T|})$ where T is the set of terms appearing at least once in the training documents and each $w_i \in [0,1]$ represents how much the term $t_i \in T$ is relevant for d.

In the simplest implementation, w_i is the term frequency (i.e., the number of occurrences of the term t_i in d). The term frequency inverse document frequency (tf-idf) function is often used as an alternative way to calculate term weights as follows:

$$w_i = TF(t_i, d) \cdot log \frac{|D|}{DF(t_i)}. \tag{2}$$

where $TF(t_i, d)$ is the term frequency while $DF(t_i)$ is the domain frequency of t_i (i.e., the number of training documents where t_i appears at least once).

The main issue of BOW is that this representation completely loses word order, grammar and context. Word embeddings (WE) allows to overcome this limitation, being a context-sensitive method. This technique requires that the terms are represented as dense vectors projected into a continuous vector space. The position of a word vector in the space is learned from training documents and it is based on the terms surrounding the considered word. The main advantage of the WE approach is the capability to learn semantic similarities between words: terms with similar meanings are characterized by closer representations.

The authors of Mikolov et al. (2013) demonstrated that semantic and syntactic patterns can be reproduced using vector arithmetic. For example, it is possible to obtain a vector representation of "Sister" by subtracting the one representing the word "Man" from the one indicating "Brother" and then adding the representation of "Woman". Once WEs are learned, a document can be represented by aggregating (e.g., summing or averaging) word vectors of included terms to obtain a single vector representation (Le and Mikolov 2014). In the proposed approach, a WE aggregation strategy based on neural network, strictly coupled with the classification step, is proposed and discussed in the next sub-section.

For our purposes, we have used pre-generated WEs provided by spaCy[1]: a Python tool for NLU. The WEs have been obtained by training a CNN on large natural-language text corpora: the Universal Dependencies and the WikiNER corpora (Nivre et al. 2016) for Italian as well as the OntoNotes (Pradhan and Ramshaw 2017) and the Common Crawl[2] corpora for English.

3.2 Text categorization model

Independently from the specific document representation (BOW or WE), the classifier able to determine the attributes of interest for the post must be trained with samples encoded through the chosen representation. This task can be carried out with several methods, such as decision trees, neural networks, Bayesian classifiers, and support vector machines. In this paper we adopt a hierarchical attention-based recurrent neural networks approach that represents an end-to-end solution integrating both WE aggregation and classification steps.

[1] https://spacy.io/

[2] https://commoncrawl.org/

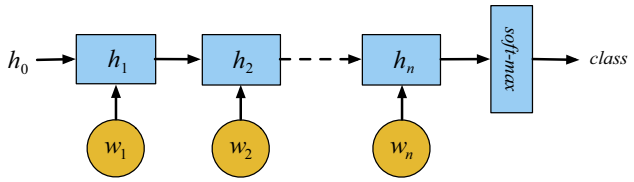

Fig. 1 Architecture of a standard RNN for document classification

A recurrent neural network (RNN) is a neural network model where connections between nodes form a directed graph along a temporal sequence. The basic premise of RNNs is to parse items forming an input sequence (e.g., WE of tokens forming a text), one after the other, updating a hidden state vector to represent the context of prior input. The hidden state (memory) is used, together with input vectors, to classify each token in light of its context.

Figure 1 shows the architecture of an RNN for document classification. Starting from an initial hidden state h_0 (usually void), for each WE w_i with $i \in \{1, \ldots, n\}$ composing a text to analyse, a new hidden state is generated according to the following equation (Goodfellow et al. 2016):

$$h_i = \tanh\left(b + Wh_{i-1} + Uw_i\right), \tag{3}$$

where parameters b (bias vector), U (input-hidden connections) and W (hidden-hidden connections) are learnt by the RNN on the training set with algorithms based on gradient descent (Goodfellowet al. 2016). When the last word w_n is consumed, the last hidden state h_n (summarizing the whole text) is used for text classification according to the following equation:

$$class = \text{softmax}\left(c + Vh_n\right), \tag{4}$$

where parameters c (bias vector) and V (hidden-to-output connections) are also learnt on the training set, and *softmax* is a function aimed at non-linearly normalizing the output into a probability distribution over the set of classes highlighting the largest values (Capuano and Caballé 2019).

The sequential nature of RNNs is often not enough to characterize natural language. According to Yang et al. (2016), a better text representation can be obtained by introducing attention mechanisms, where the output is not a function of the final hidden state but rather a function of all hidden states (general attention) or a subset of them (local attention). In the same reference, a mixed local/global attention mechanism mimicking the structure of the document was proposed, following the intuition that the parts of a document are not equally relevant for a specific classification task.

Such model uses a bidirectional RNN (where hidden states depend on both previous and next states) at a word level with a local attention mechanism to extract the words that are important to the meaning of each sentence, and then aggregate the representation of those words to form a single vector representing each sentence. Then, the same process is applied at a global level to aggregate the sentence vectors, thus obtaining a single document vector that is used for classification. In Yang et al. (2016), it was demonstrated that such a network, also referred as hierarchical attention network (HAN), outperforms other text classification methods by a substantial margin.

Figure 2 shows the HAN architecture. Given a document composed of m sentences, w_{ij} denotes the WE that represent the j-th word of the i-th sentence with $i \in (1, \ldots, m)$ and $j \in (1, \ldots, n)$. For each sentence, the HAN generates a sentence vector s_i through two subsequent steps. In a *word encoding* step the hidden state h_{ij} is calculated for each word w_{ij} summarizing the information of the i-th sentence. It is made of two components: the first depending on the preceding states and calculated according to (3) and a specular one dependent on the subsequent states. A gating mechanism is used to regulate the flow of information according to Bahdanau et al. (2015).

In a subsequent word attention step, a vector s_i is obtained for each sentence as a weighted sum of hidden states h_{ij} with $j \in (1, \ldots, n)$, where the weights α_{ij} are aimed at identifying the most informative words of a sentence, as follows:

$$s_i = \sum_{j=1}^{n} \alpha_{ij} h_{ij}; \alpha_{ij} = softmax\left(u^T \tanh\left(c + Vh_{ij}\right)\right), \tag{5}$$

where u is a "context vector" learned during the training process as described in Yang et al. (2016) and the importance α_{ij} of a word w_{ij} is measured as the normalized similarity between u and the output of the j-th unit of the i-th RNN, calculated according to Eq. (4).

The same process is then iterated at the sentence level with a sentence encoding step aimed at obtaining the hidden states h_i from the corresponding vectors s_i with $i \in (1, \ldots, m)$, followed by a sentence attention step aimed at generating the document vector v as weighted sum of the sentence vectors. The vector v is a high-level representation of the whole document (forum post) and is used as input for the classification step that is performed according to the Eq. (4) with v replacing h_n. The class corresponding to the higher value of the obtained probability distribution is then returned as the classification output and the related probability as the related confidence score.

4 Experiments and evaluation

Based on the document representation models and text categorization approaches defined in the previous section, a multi-attribute text categorization tool for MOOC forum

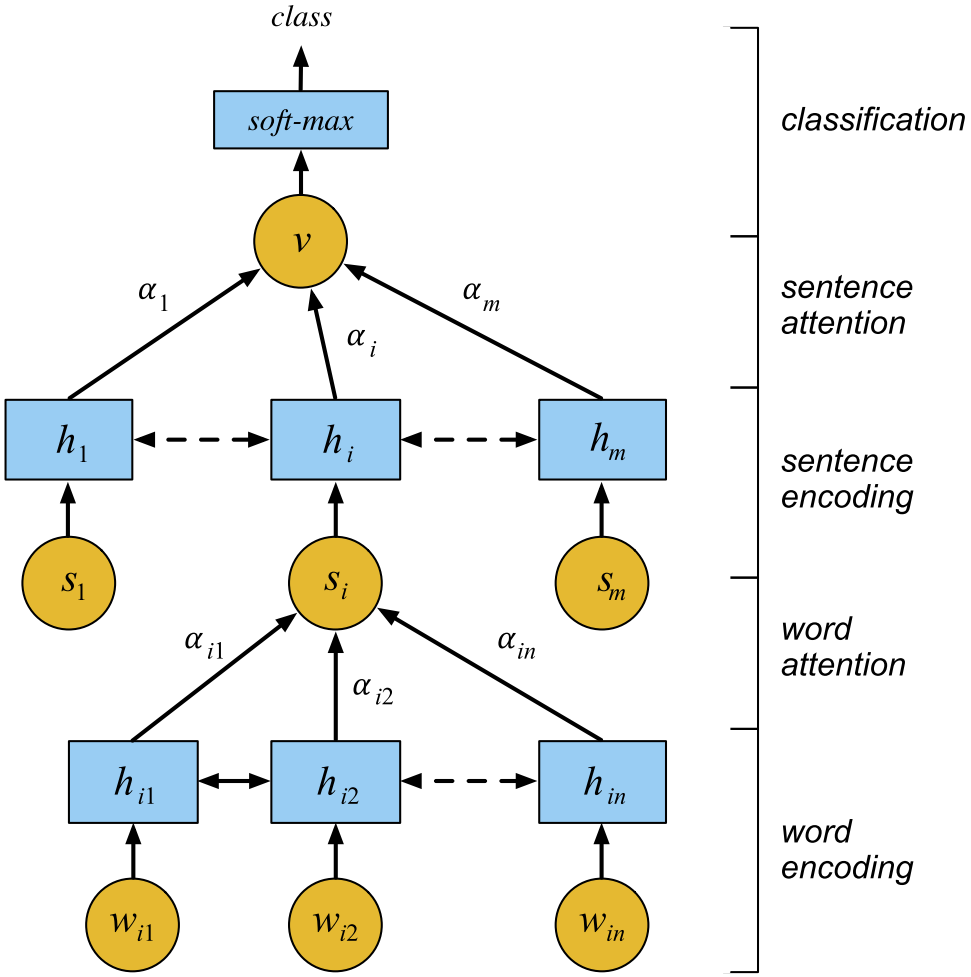

Fig. 2 Architecture of the adopted HAN for document classification

Table 1 Attributes and relative values used in the proposed multi-attribute text categorization tool for MOOC forum posts

Attribute	Meaning	Categories
Intents	The general aims of the post	Three dimensions: question (help seeking), answer (help giving), opinion
Domain	Main educational domain concepts involved in the post	Three dimensions (according to Stanford MOOCPosts): humanities, medicine, education
Topic	Main domain topics involved in the post	Eight dimensions (according to Stanford MOOCPosts): statistics, psychology, economics, health, scientific writing, statistics in medicine, emergency management, math teaching
Sentiment	The affective polarity of the post	Positive, negative, neutral
Confusion	The level of confusion expressed by the post	Low, medium high
Urgency	How urgently a reply to the post is requested	Low, medium high

posts was developed in Python. To allow external applications to use the classifier (see next section) it has also been equipped with a Web service interface. Given an input string representing a forum post, the developed tool is able to detect the information summarized in Table 1. The defined models have been trained by using the Stanford MOOCPosts[3] data set containing 29,604 anonymized learner forum posts from 11 Stanford University public online classes (Agrawal et al. 2015) within the Humanities, Medicine and Education domain areas.

For each forum post, three human annotators provided the following attributes: question (yes/no), answer (yes/no),

[3] https://datastage.stanford.edu/StanfordMoocPosts/

opinion (yes/no), sentiment (from 1 to 7), confusion (from 1 to 7), urgency (from 1 to 7). With respect to sentiment, a low score (close to 1) indicates a negative polarity while a high score (close to 7) indicates a positive one. As for confusion and urgency, a low score (close to 1) indicates a post that is not confused/urgent while a high score (close to 7) indicates a very confused/urgent post.

To match these attributes with the ones we are interested in, we included the first triplet (question, answer, opinion) in our attribute intents, while the values for sentiment, confusion and urgency have been discretized in three categories (positive/negative/neutral for sentiment, low/medium/high for confusion and urgency). The dataset discretization process worked as follows: scores below 3 were mapped to the negative/low class, scores above 5 were mapped to the positive/high class while remaining scores were mapped to the neutral/medium class. Moreover, the educational domain and the related topics were extracted from the additional information in the dataset, namely the course title and domain area.

To deal with the limited amount of data, we performed a 4-fold cross-validation. In particular, we divided the dataset in 4 subsets of equal size. For each step, we used the k-th subset with $k \in \{1, \ldots, 4\}$ as validation set, and the remaining subsets to train the classifier. As performance metrics for all attributes, we used average precision, recall and f-score, defined as follows:

$$prec = \frac{TP}{TP + FP}; rec = \frac{TP}{TP + FN}; F = 2 \cdot \frac{prec \cdot rec}{prec + rec}, \quad (6)$$

where TP is the total number of true positives (correctly predicted labels), FP is the total number of false positives (wrongly predicted labels) while FN is the total number of false negatives (correct but unpredicted labels) (Sokolova and Lapalme 2009).

Therefore, the classifier has been trained four times on 22,203 items, by using the remaining 7401 for validation. To prevent overfitting, we adopted as regularization rules a batch size increasing from 4 to 32 samples per iteration and a fixed 20% dropout rate. The average results among the 4 validation steps and obtained over 10 training epochs are summarized in Table 2 and compared with previous results obtained on the same dataset with the following networks:

- bow-ff: the documents are encoded with BOW and categorized with a fully connected 2-layers feed forward neural network;
- cnn-we-ff: the documents are represented by averaging word vectors learned from the training set through a CNN and categorized with a fully connected 2-layers feed forward neural network;

Table 2 Results achieved by the proposed method and comparison with other approaches

Attribute	Architecture	Loss	Precision (%)	Recall (%)	F-score (%)
Intents	bow-ff	0.122	79.98	73.33	76.51
	cnn-we-ff	0.087	81.2	70.20	75.34
	HAN	0.132	83.39	70.87	76.62
Domain	bow-ff	0.076	87.34	81.20	84.16
	cnn-we-ff	0.046	83.85	81.92	82.87
	HAN	0.081	87.03	82.64	84.78
Topics	HAN	0.173	84.78	65.66	74.00
Sentiment	bow-ff	0.645	88.62	86.07	87.33
	cnn-we-ff	0.051	85.43	85.43	85.43
	convL	0.245	85.25	86.84	86.04
	HAN	0.115	88.34	88.29	88.31
Confusion	bow-ff	0.081	87.19	84.25	85.70
	cnn-we-ff	0.045	84.33	81.92	83.11
	convL	0.369	80.05	80.35	80.20
	HAN	0.103	87.25	85.56	86.40
Urgency	bow-ff	0.098	84.05	75.75	79.69
	cnn-we-ff	0.056	80.45	73.95	77.06
	convL	0.474	76.90	76.33	76.61
	HAN	0.103	82.95	76.40	79.54

- convL: the documents are represented as sequences of word vectors, a CNN layer is used to extract local features, a LSTM layer is used to extract long-term semantic dependencies and a fully connected layer is then used for classification.

The first two models have been described in Capuano and Caballé (2019). Even if they were experimented on the same dataset, it should be noted that the post topic was not detected so no comparison is possible for this attribute. The third model, described in Wei et al. (2017), was only experimented for detecting sentiment, confusion and urgency seen as binary classification problems (thus supporting just positive and negative classes). To make the results comparable to those described below, the model was reimplemented in TensorFlow[4] and the last layer was converted to support three-fold classification (thus adding the neutral class).

As it can be seen in Table 2, the f-score obtained by the proposed model is in the range between 74.00% and 88.31% (HAN rows in the table) that is better than that obtained with the preceding methods in all cases but for urgency (where bow-ff results as the most accurate method). With respect to training extent, Fig. 3 demonstrate that, in almost all cases,

[4] https://www.tensorflow.org/

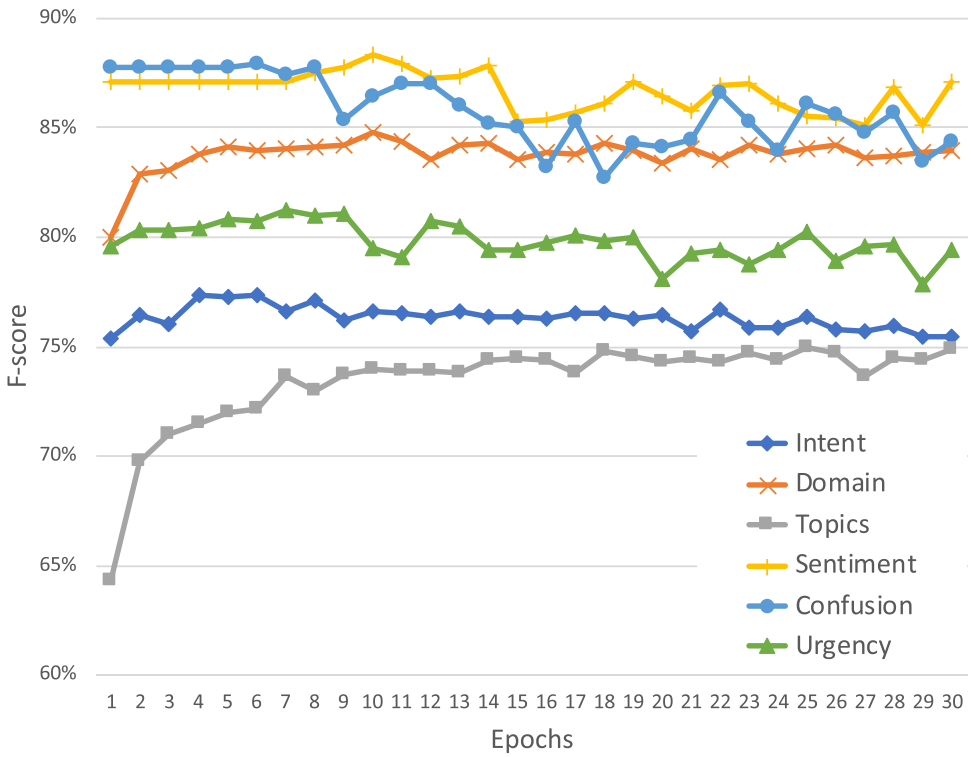

Fig. 3 Validation f-score measured per epoch for each attribute

10 epochs are enough for the network to learn an effective representation. Additional training steps do not allow to achieve a higher performance but, conversely, cause the classifier to overfit training data degrading performance on the validation set.

As reported in Sect. 2, other approaches besides those already considered have been proposed and experimented on the same dataset. However, these approaches are hardly comparable with the proposed one as they deal with binary classification and are specialized just on one attribute. For example, the model proposed in Almatrafi et al. (2018), specialized for recognizing urgent posts, achieves an f-score of 88% by relying on selected linguistic features in conjunction with post metadata as the number of reads and likes. In Alrajhi et al. (2020), a macro-averaged f-score of 84.5% is obtained for the binary classification of post urgency using the value of dataset attributes other than urgency (i.e., sentiment, confusion and intent) as additional features than the text of the post.

Both of these approaches achieve higher performance on urgency than the models shown in Table 2 but use more and higher-level information than the analyzed end-to-end solutions for solving a simpler problem (binary rather than three-folded classification on just one attribute). So, the proposed model seems to offer a good compromise between the quality of the result and the information required to obtain it. On the same information, in fact, it is more performing than the competitors. On the other hand, when information is increased (both in quantity and quality) it seems possible (albeit for specific and simplified tasks) to obtain better performance.

5 Application to conversational agents

Inspired by the research activities of the teachers' community in defining methods to encourage and support effective classroom discussions, we explored the possibility to use information extracted from forum posts with conversational agents (CAs) as an agile and efficient tool for scaffolding productive peer discussions (Caballé and Conesa 2018). As previously explained, CAs are systems used for the interaction with students within synchronous and asynchronous collaboration tools. The aim of these tools is the promotion of constructive discussions and interactions by applying specific dialogue patterns. In this section, we describe an application scenario fostering the usage of the defined text categorization tools with CAs. This application scenario focuses on the design of agent-based dynamic interventions deriving from the academically productive talk (APT) discourse framework (Demetriadis et al. 2018) extended with our multi-attribute post categorization approach presented above.

5.1 Academically productive talk

The APT framework emphasizes the key role of social interaction in inducing beneficial mental processes. In contrast

Table 3 Examples of APT-CA interventions

APT move	Example
1. Add-on	[Student], would you like to add something to what your partner said about [Concept]?
2. Agree–disagree	[Student], what do you think of what [Partner] said about [Concept]? Do you agree or disagree?
3. Verify-and-clarify	[Student], do you agree with the following statement: [Concept1] [verb expression] [Concept2]? Why?
4. Build-on-prior-knowledge	[Student], do you think [Concept1] is somehow related to [Concept2]? How?

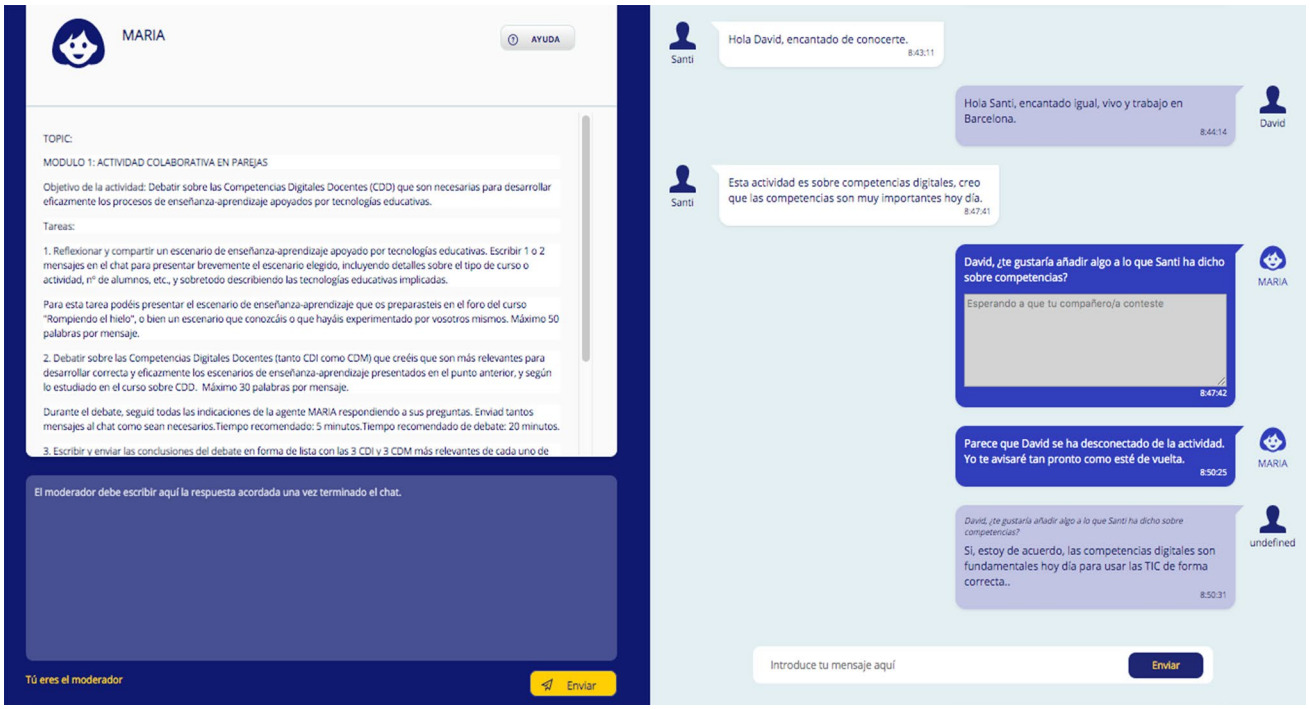

Fig. 4 Example of a real APT-based CA intervention (blue bubbles, in Spanish) in a MOOC forum (Caballé et al. in press)

to other popular classroom discourse frameworks, APT prioritizes reasoning over correctness and attempts to orchestrate student-centered discussions (Michaels et al. 2010). According to this framework, the instructors should provide dynamic support via facilitative conversational moves, in order to promote the participation and the orchestration of constructive discussions. The aim of APT is to model and trigger appropriate forms of peer dialogue. This way, the APT proposes a set of moves designed to stimulate productive forms of peer dialogue. Examples include add-on (e.g., "Would you like to add something to …"), building on prior knowledge (e.g., "Do you think concepts A and B are related to each other …"), verify and clarify (e.g., "Why do you think that?"), etc. (see Table 3).

In addition, CA interventions follow the definition of an associated pattern. A pattern is related to an event or a combination of events, that happen in a forum and could be interesting for the agent to identify and analyze as a possibility for triggering an intervention. A sample APT-based CA intervention is shown in Fig. 4.

Typically, a pattern is intended as a combination of something uttered by the humans along with contextual information of what is going on in the chat environment (Tegos et al. 2019). Within this research, the patterns of interest can be categorized as follows: (a) static patterns, which are related to one or more events independent of the dynamic evolution of peer interaction, and (b) dynamic patterns, that can be associated to contextual events arising from the analysis of peer utterances and the extraction of certain key concepts, defined by the teacher when designing the conceptual links, i.e., the agent domain ontology (Demetriadis et al. 2018).

5.2 Extended APT with post categorization

While APT considers certain moves that can stimulate constructive forms of peer dialogue, so that fruitful moves can be dynamically generated, further attributes, such as

intents, topics, sentiment, urgency and confusion should be also detected in the forum as well as in individual posts. To this aim, the proposed method can give an important contribution. Figure 5 depicts an example of the interaction of a learner in a discussion thread, moderated by a CA fostering constructive discussions according to the results of the proposed multi-attribute post categorization tool. In the first post Mario requires support on a specific course topic. The method recognizes his intent (and the other attributes) and determine the related topics which can help the student. Among the results, the categorization tool retrieves a positive assessment of another student (Laura) on the same subject and the chatbot solicited her feedback through a contribute linking move.

Then, Laura replies in the third post to Mario following the request of the chatbot. The intent and the topics of the posts correspond to the values of the previous post, so the answer is on the same topic; however, the system is able to understand that the reply is not completely clear, since a high level of confusion is detected. Therefore, the chatbot asks again for new interventions, with another contribute linking move. As a consequence, Anna writes a fifth post giving a pertinent (right intents and topics), clear (low confusion) and appreciated (positive sentiment) answer.

However, in the sixth post Mario asks for further support on the same topic, according to the detected intent. In this case, the negative sentiment and the high levels of urgency and confusion suggest disappointment and frustration. To provide immediate help and alleviate dissatisfaction, the chatbot performs an affective move and suggest a direct link to the available resources of the course on the topic.

This is a simple example of a potential application with the attributes automatically extracted by the proposed method within a CA framework. This information can then be used as input for CAs in order to involve students in guided and constructive interactions carried out in natural language within the discussion forums (see Fig. 4). The goal of these agents would be to promote helpful and constructive discussions between colleagues, including argumentation and mutual clarifications and explanations. These discussions, based on the extracted information, could be initiated and encouraged by informed agents' actions, improving the effectiveness and timeliness of the interventions (Toti et al. in press). Further applications may include the extension of the tool to learning domains, which only requires an additional training step. In fact, in Wei et al. (2017) the authors proved that transfer learning techniques relying on very few labelled additional examples extracted from the new course can be successfully used to extend the capabilities of the tool to new courses. We expect that similar considerations hold for intent, confusion and urgency.

As a final remark, it is worth to separately discuss the recognition of domain topics. Indeed, domain concepts and their links must be modelled with different techniques, such as knowledge graphs or lightweight ontologies (Capuano et al. 2011; Capuano et al. 2009), since the CA and the categorizer need to determine the relevant concepts and their relations. In this way, the categorizer can be trained to recognize these concepts. NLU methodologies like named entity recognition (Lee et al. 2006) may help to reduce the time necessary for labelling the new samples.

6 Conclusions and further work

In this paper we demonstrated the effectiveness of the proposed multi-attribute text categorization tool, designed for the analysis of MOOC forum posts. The method, experimented on a widely used existing dataset of annotated posts, is able to detect intents, domain, topics, sentiment, confusion and urgency of forum posts achieving a f-score between 74 and 88%. Simulation results are better when compared to those obtained with other approaches on the same dataset. The attributes can be exploited by instructors to plan their interventions as well as input for autonomous agents aimed at engaging learners in guided discussions. To this end, we also described the integration of the classifier inside an existing tool for conversational agents, based on the academically productive talk framework.

Future directions for the research in this field are surely the usage of the analysis of the performance of the proposed tool in real MOOC environments and its improvement according to the experimental results. As for the first point, these experiments are already ongoing within the purposes of the colMOOC project[5] (see acknowledgements) aimed at developing and experimenting conversational agents and learning analytics tools in real learning contexts of MOOCs (Demetriadis et al. 2018). To this end, as part of the evaluation tasks of this project, a first experience with a MOOC course equipped with an innovative APT-based conversational agent system supporting synchronous collaboration in dyads was recently run with about 2000 students enrolled with preliminary evaluation results (Caballé et al. in press). Results and practical implications of this experience for the work presented here will be analyzed and reported in next stages of this research.

Regarding the second point, one of the known issues of the proposed tool is the unbalance between the represented categories. We conjecture that the accuracy may be further improved by extending the dataset with new real samples or through data augmentation techniques. An experiment aimed at integrating transfer learning techniques to facilitate the tool adaptation to different learning

[5] https://colmooc.eu/

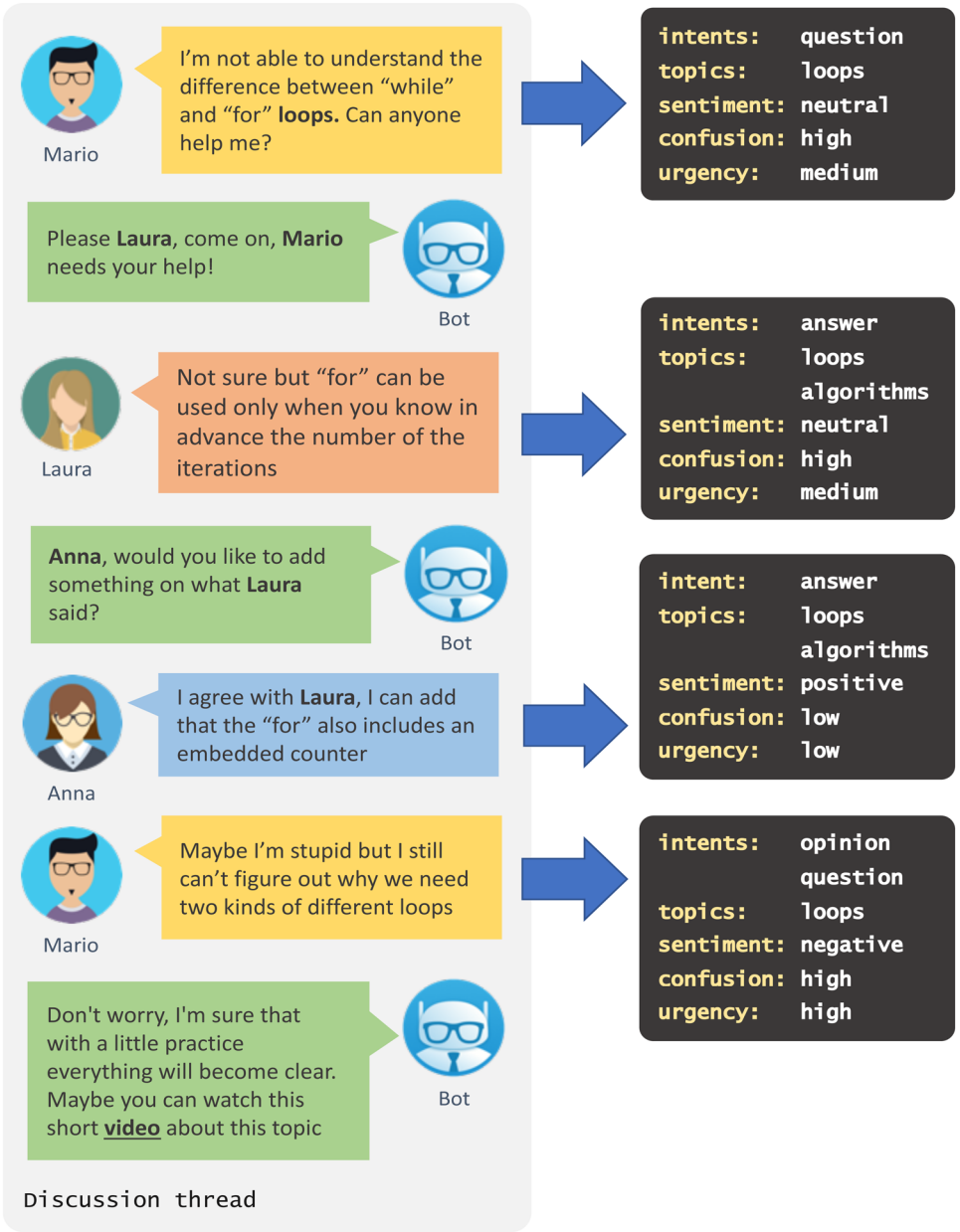

Fig. 5 Example of conversation encouraged by the proposed method. The extracted intent, topic, sentiment, confusion and urgency allows the chatbot to promote constructive discussions

domains has been already performed with encouraging results (Capuano in press). In addition, the adoption of advanced word representation models like contextualized embeddings (Devlin et al. 2019) as well as the integration of classical NLU techniques like part-of-speech tagging, named entity recognition, parsing, etc. is envisaged to try to further improve the classifier performance.

Funding Open access funding provided by Università degli Studi della Basilicata within the CRUI-CARE Agreement. This work has been supported by the project colMOOC "Integrating Conversational Agents and Learning Analytics in MOOCs", co-funded by the European Commission within the Erasmus + program (ref. 588438-EPP-1-2017-1-EL-EPPKA2-KA).

Compliance with ethical standards

Conflict of interest There is no conflict of interest nor any competing interest.

Availability of data and material The work uses referenced public datasets.

Code availability The source code of the developed system is not disclosed.

References

Agrawal A, Venkatraman J, Leonard S, Paepcke A (2015) YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips. In: Proceedings of the International Conference on Educational Data Mining. Madrid, Spain, pp. 297–304

Almatrafi O, Johri A, Rangwala H (2018) Needle in a haystack: identifying learner posts that require urgent response in MOOC discussion forums. Comput Educ 118:1–9

Alrajhi L, Alharbi K, Cristea A (2020) A multidimensional deep learner model of urgent instructor intervention need in MOOC Forum Posts. In: Proceedings of Intelligent Tutoring Systems 2020. Springer, Cham, Switzerland, pp. 226–236

An Y, Pan L, Kan M, Dong Q, Fu Y (2019) Resource mention extraction for MOOC discussion forums. IEEE Access 7:87887–87900

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). San Diego, CA, USA

Caballé S, Conesa J (2018) Conversational agents in support for collaborative learning in MOOCs: an analytical review. In: Proceedings of the 10th International Conference on Intelligent Networking and Collaborative Systems (INCoS). Springer, pp. 384–394

Caballé S, Conesa J, Gañan D (in press) Evaluation on using conversational pedagogical agents to support collaborative learning in MOOCs. In: Proceedings of the 15th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. Springer, Tottori, Japan

Caballé S, Lapedriza A, Masip D, Xhafa F, Abraham A (2009) Enabling automatic just-in-time evaluation of in-class discussions in on-line collaborative learning practices. J Dig Inform Manage 7(5):290–297

Capuano N (in press) Transfer learning techniques for cross-domain MOOC forum post analysis. In: Intelligent Systems and Learning Data Analytics in Online Education. Elsevier, Amsterdam, The Netherlands

Capuano N, Caballé S (2015) Towards adaptive peer assessment for MOOCs. In: Proceedings of the 10th International Conference on P2P, Parallel, GRID, Cloud and Internet Computing (3PGCIC 2015). IEEE Computer Society, Krakow, Poland. pp. 64–69

Capuano N, Caballé S (2019) Multi-attribute categorization of MOOC forum posts and applications to conversational agents. In: Proceedings of the 14th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC 2019). Antwerp

Capuano N, Dell'Angelo L, Orciuoli F, Miranda S, Zurolo F (2009) Ontology extraction from existing educational content to improve personalized e-Learning experiences. In: Proceedings of the 3rd IEEE International Conference on Semantic Computing (ICSC 2009). Berkeley, CA, USA

Capuano N, Gaeta M, Salerno S, Mangione GR (2011) An ontology-based approach for context-aware e-learning. In: 3rd IEEE International Conference on Intelligent Networking and Collaborative Systems. Fukuoka, Japan

Cichosz P (2019) Case study in text mining of discussion forum posts: classification with bag of words and global vectors. Appl Mathe Comput Sci 28(4):787–801

Demetriadis S, Tegos S, Psathas G, Tsiatsos T, Weinberger A, Caballé S et al (2018). Conversational agents as group-teacher interaction mediators in MOOCs. In: Proceedings of Learning With MOOCS (LWMOOCS). Madrid, Spain, pp. 43–46

Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Minneapolis, Minnesota

Dyke G, Howley I, Adamson D, Kumar R, Rosé C (2013) Towards academically productive talk supported by conversational agents. In: Productive multivocality in the analysis of group in-teractions. Springer, pp. 459–476

Ferschke O, Howley I, Tomar G, Yang D (2015) Fostering discussion across communication media in massive open online courses. In: Proceedings of the 11th International Conference on Computer Supported Collaborative Learning (CSCL). pp. 459–466

Ferschke O, Yang D, Tomar G, Rosé C (2015) Positive impact of collaborative chat participation in an edx mooc. In: 17th International Conference on Artificial Intelligence in Education (AIED). Springer, pp. 115–124

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, USA

Guitart I, Conesa J (2016) Adoption of business strategies to provide analytical systems for teachers in the context of universities. Intern J Emerg Technol Learn (iJET) 11(7):34–40

Hollands F, Tirthali D (2014) MOOCs: expectations and reality. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University, NY

Jang B, Kim M, Harerimana G, Kang S, Kim J (2020) Bi-LSTM model to increase accuracy in text classification: combining Word2vec CNN and attention mechanism. Appl Sci 1–14

Kumar R, Rosé C (2011) Architecture for building conversational agents that support collaborative learning. IEEE Trans Learn Technol 4(1):21–34

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014). Beijing, China

Lee C, Hwang Y, Oh H, Lim S, Heo J, Lee C et al (2006) Fine-grained named entity recognition using conditional random fields for question answering. Lect Notes Comput Sci 4182:581–587

Manning C, Raghavan P, Schutze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge

Michaels S, O'Connor M, Hall M, Resnick L (2010) Accountable talk sourcebook: for classroom that works. University of Pittsburgh Institute for Learning

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26:3111–3119

Mongkhonvanit K, Kanopka K, Lang D (2019) Deep knowledge tracing and engagement with MOOCs. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. pp. 340–342

Nivre J, de Marneffe M, Ginter F, Goldberg Y, Hajič J, Manning C, et al (2016) Universal dependencies v1: a multilingual treebank collection. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Portorose, Slovenia

Pousada M, Caballé S, Conesa J, Bertrán A, Gómez-Zúñiga B, Hernández E, et al (2017) Towards a web-based teaching tool to measure and represent the emotional climate of virtual classrooms. In: Proceedings of the 5th International Conference on Emerging Intelligent Data and Web Technologie. Springer, pp. 314–327

Pradhan S, Ramshaw L (2017) OntoNotes: large scale multi-layer, multi-lingual, distributed annotation. Handbook of linguistic annotation. Springer, The Netherlands, pp 521–554

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47

Siemens G (2013) Massive open online courses: innovation in education? Open Educational Resources: innovation, research and practice. Athabasca University Press, Vancouver, Canada, pp 5–16

Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manage 45:427–437

Sun X, Guo S, Gao Y, Zhang J, Xiao X, Feng J (2019) Identification of urgent posts in MOOC discussion forums using an improved RCNN. In: IEEE World Conference on Engineering Education (EDUNINE). IEEE, pp. 1–5

Tegos S, Psathas G, Tsiatsos T, Demetriadis S (2019) Designing conversational agent interventions that support collaborative chat activities in MOOCs. In: Proceedings of EMOOCs-WIP, pp. 66–71

Tomkin J, Charlevoix D (2014) Do professors matter?: using an a/b test to evaluate the impact of instructor involvement on MOOC student outcomes. In: Proceedings of the ACM Conference on Learning@Scale. New York, NY, USA

Toti D, Capuano N, Campos F, Dantas M, Neves F, Caballé S (in press) Detection of student engagement in e-learning systems based on semantic analysis and machine learning. In: Proceedings of the 15th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. Lecture Notes in Networks and Systems. Springer, Tottori, Japan

Wei X, Lin H, Yang L, Yu Y (2017) A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. Information 8(3):92

Wen M, Yang D, Rosè C (2014) Sentiment analysis in MOOC discussion forums: what does it tell us? In: Proceedings of Educational Data Mining

Yang D, Wen M, Howley I, Kraut R, Rose C (2015) Exploring the effect of confusion in discussion forums of massive open online courses. In: Proceedings of the 2nd ACM Conference on Learning@Scale. New York, NY, USA

Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the International NAACL-HLT 2016 Conference. San Diego, CA, USA