Neural Computing and Applications (2023) 35:161–175
https://doi.org/10.1007/s00521-021-05750-z

S.I. : INFORMATION, INTELLIGENCE, SYSTEMS AND APPLICATIONS

Classifying MOOC forum posts using corpora semantic similarities: a study on transferability across different courses

Anastasios Ntourmas[1] · Sophia Daskalaki[1] · Yannis Dimitriadis[2] · Nikolaos Avouris[1]

Received: 29 January 2020 / Accepted: 16 January 2021 / Published online: 4 February 2021

Abstract
Information overload in MOOC discussion forums is a major problem that hinders the effectiveness of learner facilitation by the course staff. To address this issue, supervised classification models have been studied and developed in order to assist course facilitators in detecting forum discussions that seek for their intervention. A key issue studied by the literature refers to the transferability of these models to domains other than the domain in which they were initially trained. Typically these models employ domain-dependent features, and therefore they fail to transfer to other subject matters. In this study, we propose and evaluate an alternative way of building supervised models in this context, by using the semantic similarities of the forum transcripts with the dynamically created corpora from the MOOC environment as training features. Specifically, in this study, we analyze the case of two MOOCs, in which the models that we built are classifying forum discussions into three categories, course logistics, content-related and no action required. Furthermore, we evaluate the transferability of the derived models and interpret which features can be effectively transferred to other unseen courses. The findings of this study reveal the main benefits and trade-offs of the proposed approach and provide MOOC developers with insights about the main issues that inhibit the transferability of these models.

Keywords Massive Open Online Courses · Supervised modeling · Transferability · Forum discussion classification

1 Introduction

Massive open online courses (MOOCs) provide a wide range of opportunities in online learning. The potential of MOOCs is based on their unique power to be massive and accessible worldwide [1]. Due to their popularity, these online courses have become rich sources of available data that can be used for learning analytics purposes [2]. Therefore, a growing body of learning analytics research for MOOCs has been performed in recent years [3–6]. The main contribution of this line of research is to provide MOOC designers and instructors with insights in order to better redesign and improve their courses so as to promote effective learning.

A major component of any MOOC is its discussion forum, which is also a rich source of data. Within the discussion forum, learners can be facilitated by their peers or by the course instructional staff via asynchronous communication and active involvement in discussions [7]. One of the main challenges that MOOC instructors face in

The original version of this article was revised: Due to open choice cancellation.

✉ Anastasios Ntourmas
a.ntourmas@upnet.gr

Sophia Daskalaki
sdask@upatras.gr

Yannis Dimitriadis
yannis@tel.uva.es

Nikolaos Avouris
avouris@upatras.gr

[1] Electrical and Computer Engineering Department, University of Patras, Patras, Greece

[2] School of Telecommunications Engineering, Universidad de Valladolid, Valladolid, Spain

order to effectively facilitate learners is the *bandwidth problem*. The more learners participate within the forum the more difficult it is for instructors to keep track of the forum discussions and, thus, to provide prompt support to them [8]. MOOC instructors usually hire teaching assistants [9] for their courses or seek high-performing learners and invite them to facilitate their peers voluntarily as community teaching assistants (CTA) [10]. The main task of CTAs is to keep track of the forum discussions, perform timely interventions so as to facilitate learners with their problems and reduce the workload of the instructor.

Learners make use of the forum in different ways, escalating the *bandwidth problem* [11], despite the presence of course facilitators. Learners may participate in the forum so as to (a) pose content-related questions (problems related with the course material), (b) ask questions related to logistics (e.g., technical problems) or (c) discuss on topics that do not require facilitation (e.g., socializing) [12]. An important issue that MOOC facilitators face is the time and energy they have to spend in searching for posts that necessitate their intervention [13]. Usability issues of the MOOC forum platform were found in [14] to have a negative effect on this task, mostly due to the lack of sufficient navigation facilities. These issues may in fact have an adverse impact on the quality of support that MOOC facilitators provide to learners [14, 15] and advocate for the pressing need of developing tools that could significantly improve the course facilitators' experience in the forum by providing a "bird's eye" view of learners discussions [16].

Within the MOOC forum context, several automated decision support tools have been deployed that provide visual feedback to MOOC facilitators [17, 18] or recommendations about discussions in which their intervention is needed [19, 20]. From the learner perspective, recommendation systems about discussions that match learners' interests have been deployed [21] and tools that assist learners in finding potential peer facilitators [22]. Recommendation tools require the training of classification models, supervised or unsupervised, from data that derive most commonly from a specific course domain. Several studies claim that the type of interactions that take place between the participants of a MOOC forum [10] and the characteristics of the social networks that are formed by them [23] may depend on the course subject matter. This may imply that in MOOCs of different subject matter, the training data of the classification models may differ considerably. An important issue that has been studied by Moreno-Marcos et al. [24] is the transferability of such classification models to other course domains. The process of building a classification model needs a considerable effort by MOOC developers and the transferability of these models is a challenging and still open issue. Building a separate classification model for every new course may not be a viable solution for MOOC developers, so it is important to obtain models that could be extended beyond the course in which they were built, so as to ensure that these models are sustainable [25]. This issue was also stated in the literature review of Gašević et al. [26], where they revealed that few contributions actually evaluate the impact and transferability of such models in different contexts.

In this work, we address the *Transfer Learning* problem of automated message classification in MOOC forums by developing two supervised classification models, each one for a course of a widely different subject matter. Then, we evaluate their transferability by performing a cross-course evaluation process, where each classifier will be used to perform classifications on the unseen messages of the other course's dataset. To build the classifiers, we make use of the available data derived from the course environment (text documents) in order to construct two corpora per course, one related to the course's content (CR corpus) and one related to course's logistics (LR corpus). This approach was inspired by the work of Shatnawi et al. [27], where they proposed a way of constructing a *course domain ontology*, to be used to extract the training features for their classification models. In our study, the features that will be used for training the classifiers, will be the semantic similarities of the forum messages with each of the two corpora. We mainly focus on these specific features in order to interpret if they can be considered as adequate indicators in distinguishing forum messages. The main idea is that when the classifiers will be transferred to another course, these features will be extracted from relevant corpora of the new course. So, it is worth examining to what extent such features can be described as *course-independent*. The preliminary findings of this case study reveal the potential and the main issues derived from this approach that should be considered for our future research. Therefore, this study may provide MOOC developers with insights that may assist them in the future development of tools that will support course facilitators.

The rest of the paper is organized as follows. In the next section, relative studies that have addressed the problem of transferability in automated message classification in MOOC forums are presented. In Sect. 3, we present the context of our research. In Sect. 4, the definition of the methodology for this study is provided. In Sect. 5, we present results derived from our research and we perform an interpretation of our findings. Finally in Sect. 6 we provide the conclusions of this work and the implications for future research.

2 Related work

In this section, we present research studies in the field of machine learning, whose goal was to develop automatic decision models within the MOOC forum context. We distinguish these studies according to the machine learning

approach (unsupervised and supervised approach) that was followed.

2.1 Unsupervised machine learning approaches

In their work, Attapatu and Falkner [28] developed a framework for labeling MOOC forum discussions automatically via a topic modeling approach. Specifically, they used the text corpora from the weekly lectures of two MOOC offerings and the number of each week as label in order to train a Naive Bayes classifier. Then, by extracting the topics via Latent Dirichlet Allocation from the test MOOC discussion dataset, they tried to label the discussion according to the respective week that its topic may belong to. The evaluation of their model revealed promising results and they claimed that such model can provide adequate recommendations to learners for content-related discussions of their interest. Despite the fact that they did not mainly focus on addressing the transferability issues of their models, we consider important the fact that they used lecture transcripts of the MOOC and followed an automatic approach of extracting their features. Therefore, in our study, we further investigate this approach by examining how the material of different course domains can affect the performance of such models.

In another study, Ezen-Can et al. [29] employed unsupervised machine learning techniques to automatically analyze learner dialogues, with the ultimate goal to allow massive-scale automated discourse analysis to enhance learner support. In their methodology, they used the k-medoids clustering algorithm in order to cluster similar dialogue acts and via a qualitative analysis they interpreted the characteristics of the extracted clusters. The results of their study were encouraging and suggested that their approach can contribute to the potential development of adaptive real-time support systems for learners. Liu et al. [30] followed a semi-automatic annotation approach, combining both unsupervised and supervised machine learning techniques. From the content of the lecture slides and posts, a set of relevant words was determined by manually listing words that appear most frequently and extracting those that are specific to the course. This approach was performed manually to train several classification models. The main goal of the models was to calculate the relevance of the learner speech acts to the annotated corpus and provide predictions according to the types of discussions on a MOOC dataset. Results revealed that random forest using the randomization technique provided the most satisfactory results, but the main findings of their analysis were that the extracted topics did not cover all possible forum interactions.

A limitation of the aforementioned studies [29, 30] is that they tried to extract categories automatically from the forum transcripts, so as to perform a more accurate clustering. It was shown that this approach may result in not covering all possible interactions within the forum [30]. In our study, this problem is addressed at a higher level, where forum discussions are differentiated according to the most appropriate MOOC actor that should intervene. Most concretely, moderators should primarily take action in discussions related to course logistics; course facilitators such as CTAs or instructors should intervene in content-related discussions, while community building discussions should be self-regulated by learners. Furthermore, a recommendation tool based on this classifier may especially benefit course facilitators since they might deal only with discussions of their own interest and, thus, the *bandwidth problem* might be mitigated.

2.2 Supervised machine learning approaches

Several studies have tried to address the *Transfer Learning* problem in discussion forums, but their results revealed the difficulty of such task via a supervised approach. In their study, Almatrafi et al. [31] address the problem of information overload in MOOC discussion forums from the course facilitators' perspective and propose a supervised model that can identify urgent posts. The main goal was to help course facilitators prioritize their responses so that they make prompt interventions. To train their model they used data derived from three different course domains: Humanities, medicine, and education, and they extracted the main linguistic features from each domain via several feature extraction methods. To evaluate the transferability of their models, they excluded one of the three courses from the training set and used it as a test dataset. The results of the model performance were moderate (lowest Kappa: 0.58, highest: 0.64) in identifying urgent posts on the unseen domain and, thus, illustrating the difficulty of performing such task via supervised machine learning approaches. In another work, Boyer and Veeramachaneni [32] investigated the performance of supervised models built from data of previous offerings of the same course with the goal to predict learners that are about to drop-out. To train their models they used activity data derived from the MOOC participants. Their evaluation results revealed that their models had a poor performance when transferred to other domains and that further research should be performed across different course domains in order to interpret the parameters that can resolve transferability issues of the models.

On the other hand, several studies revealed promising results in terms of resolving transferability issues of supervised models. In their study, Whitehill et al. [33] developed a multinomial logistic regression model in order

to estimate the probability that a learner will drop-out based on features derived from the learners' event history within the MOOC platform. They used clickstream and grade history data derived from 10 different MOOCs of the Coursera platform in order to train their model. To evaluate the transferability of their model, they built their classifier from the course with the most training data and used it to perform predictions on the other nine. The performance of the classifier indicated that their model could be generalized to other course domains. Despite the fact that the forum transcripts were not used in this research for the training of the model, this study provided evidence that supervised machine learning approaches can resolve transferability issues for a specific prediction objective. Kizilzec and Halawa [34] used features from learners' interaction data with video, assignments, and forum transcripts from 20 MOOCs with the goal to train a logistic mixed-effects (hierarchical) model that will predict learners' attrition in the course. Their results indicate that employing data from a large number of available courses could improve the transferability of the model to another course domain.

Based on the studies presented in this section, we may observe that more promising results were produced by studies where a high number of MOOCs was available for training [33, 34], but still the transferability was not resolved adequately in widely different course domains. The solution of increasing the amount of training data in supervised classification tasks seems to be a reasonable approach, but on the other hand it needs more computational resources and much more human effort in order to manually label the data. Therefore, further research is still necessary regarding transferability between courses of different subject matter, so as to better interpret why classifiers have a poor performance when transferred, while requiring less computational and human resources. Another important factor that hinders the effectiveness of the supervised models when transferred to a new course domain, claimed by Kidzinsk et al. [35], is that several training features are bound in a specific course domain terminology. They claim that in order to achieve transferability, the predictive power should be reinforced with course independent variables.

Taking into consideration these observations of Kidzinsk et al. [35], we contribute in this research line by investigating if the semantic similarities with the CR and LR corpora can be considered as course independent training features. This goal is based on the fact that for each new course, a relevant CR and LR corpus may be dynamically created without requiring significant computational and human resources, and from these corpora the features will be extracted. In our case, study the evaluation will be performed between two different MOOCs in terms of subject matter, a Technology and a Humanities course.

3 The current study

3.1 Context of the study

In this study, we used data derived from two MOOCs that were offered in 2017 on *Mathesis*, a major Greek MOOC platform based on OpenEdX technology. The first MOOC, *Introduction to Python* (PY course), was an introductory course to computer programming via the Python programming language. The second, *World History: Man versus Divine* (WH course), aimed at introducing learners to the history of Asian religions during the world history's second circle. The duration of the courses was 6 and 9 weeks, respectively. The data retrieved per course consisted of the anonymized transcripts from the corresponding discussion forum.

The architecture of the discussion forum for the two courses is presented in Fig. 1. Discussions were organized in weeks, i.e., each week the forum participants could create their own threads (Level 1). Within any thread, participants could create new posts (Level 2) and they could provide their replies to a post (Level 3). In addition, there were two different types of threads, those created by the course staff and those created by the learners. Threads created by the course staff usually included multiple posts, each of them initiating a different discussion. On the other hand, threads created by learners were mostly related to a single question that initiated just one discussion. In particular, for the threads created by the course staff, we consider as a forum discussion the set of messages that consists of an initial *starting post* and its corresponding

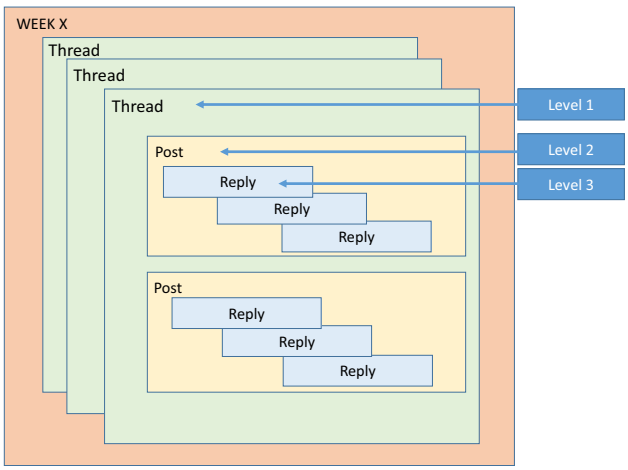

Fig. 1 Discussion forum architecture of PY and WH courses

replies. For the threads initiated by learners, we consider as forum discussion the whole thread itself. In this case, the *starting post* is the initial thread message and the corresponding replies are the posts and replies that follow within the thread.

We consider the *starting post* of a discussion to be a crucial part of the analysis, since as Wise et al. [36] stated, "a starting post reflects the primary intention of the discussion initiator and sets a direction for the content of all subsequent replies". Thus, the classification performed in this case study was based on the discussions' *starting posts*. More specifically, we attempt to solve a multi-class classification problem that will classify *starting posts* according to the following coding scheme:

- No Action Required (NAR) posts.
- Content-related (CR) posts.
- Logistics-related (LR) posts.

These categories correspond to different actors, that might be more appropriate to intervene, as explained in Sect. 2.1. For the LR discussions, we expect the platform personnel to provide assistance, for the CR discussions the teaching assistants should intervene and for the NAR discussions we assume that the community of learners might self-regulate these discussions, without any need for action by the platform personnel or the course facilitators. Such an attempt could further reduce the workload of Teaching Assistants by keeping track of the forum discussions and mainly focus on CR discussions.

3.2 Research questions

In this study, we make use of the available documents that exist within the MOOC environment in order to construct two corpora per course that would contribute to the feature extraction of the multi-class classification problem that we attempt to solve. The first corpus derives from the course material documents (CR corpus) and the second one from the logistics-related documents provided by the MOOC platform (LR corpus). The features that will be extracted per *starting post* will be its semantic similarity with the CR corpus (CR similarity) and the semantic similarity with the LR corpus (LR similarity). Through this approach, we look for features that may sufficiently differentiate the *starting posts* of each category, based on corpora that can be easily derived from each course material. Furthermore, the main goal is to examine to what extent a supervised model trained with these features can be transferred to a new course of a different subject matter.

To address this goal, we calculate the CR and LR semantic similarities for the *starting posts* per course. Then, we investigate if these features can help in identifying the category that each *starting post* belongs to, via a statistical analysis that will be described in Sect. 4. Next, we develop a supervised model per course, trained with the corresponding semantic similarities of its posts, and we evaluate the predictive performance of the models on their respective test datasets. Finally, we perform a cross-course evaluation of the models in order to assess their transferability. To sum up, the main research questions that we address in this case study are the following:

- **RQ1:** *Do the starting posts that belong to the LR, CR and NAR categories can be differentiated according to their semantic similarities for each course?*
- **RQ2:** *Can the semantic similarities be used to build a reliable supervised model for each course that classifies starting posts according to these categories?*
- **RQ3:** *Can the supervised models, that were trained from the semantic similarities of the starting posts with the CR and LR corpus, be reliably transferred to another course of a substantially different subject matter?*

4 Methodology

In this section, we describe the dataset that we used and the methodology that we followed for the development of the supervised models for the two MOOCs. We present the way we performed the normalization of the forum transcripts, the feature extraction methods that we followed for each approach, the classification algorithm that we used in order to build our models and finally the evaluation process. In Fig. 2, we present a visual representation of the steps that will be described in this Section.

4.1 Dataset description

The dataset used for the current analysis included the forum transcripts of each course. The structure of the dataset contained the following fields:

- **id:** the id of the forum message
- **type:** if it is a thread or a post message
- **author:** the author of the message
- **body:** the content of the message
- **category:** the category derived from the coding of the transcripts to be described in Sect. 4.2

The total number of entries was 5134 for the PY and 5611 for the WH dataset, which correspond to the total number of posted messages in the discussion forum of each course. As described in Sect. 3.1, the analysis mainly focused on the messages that initiated a new discussion in the forum

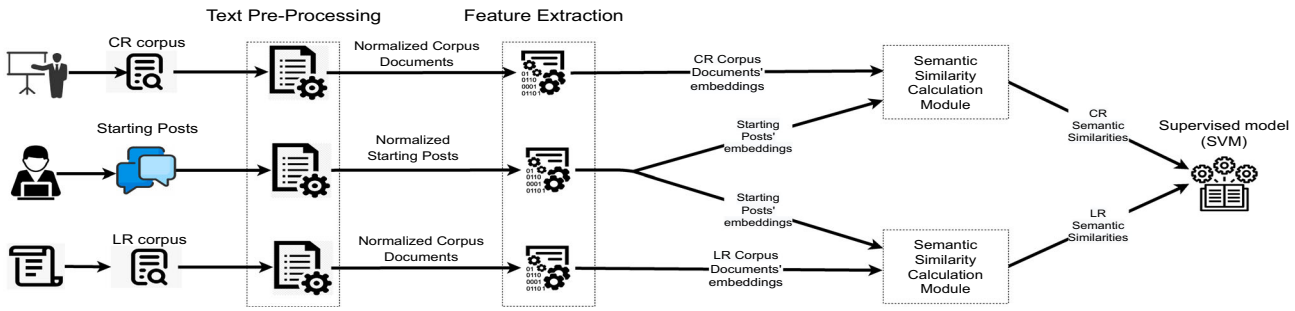

Fig. 2 Procedure of building the supervised classifier per course

(*starting posts*). These messages were used for the classification process.

4.2 Coding of the forum transcripts

According to the forum dataset of the two courses, there were 980 and 997 in total *starting posts* in PY and WH, respectively. Two coders performed the manual labeling of the *starting posts* for both courses, according to the following instructions:

- **CR:** problem related to the course material.
- **LR:** problem related to the course logistics.
- **NAR:** discussion related to community building, so No Action Required.

Their results were evaluated by using Cohen's kappa (k). Cohen's kappa is a chance-corrected measure for inter-rater reliability that accounts for the possibility of chance agreement between the coders [37]. The kappa coefficient was $k = 0.83$ for our data, suggesting quite a high inter-rater reliability. Furthermore, in order to achieve an absolute consensus, the two coders debated their disagreements until no further dispute existed and a kappa coefficient of 1 was reached. Based on this labeling the distribution of the different post categories is shown in Fig. 3. Apparently, the majority (app. 60%) were CR posts around 30% were LR posts, while just 10% were labeled as NAR posts.

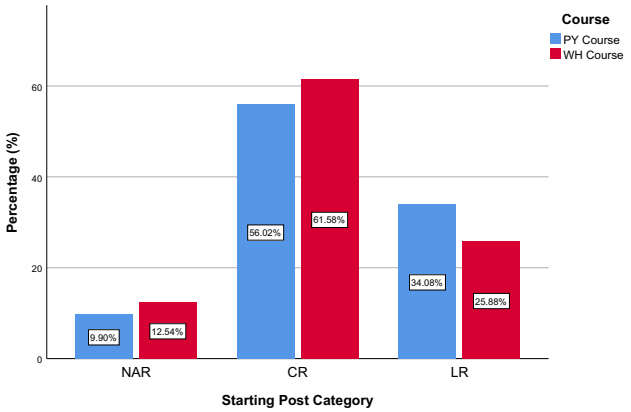

Fig. 3 Distribution of *starting posts* categories in each course's forum data after the coding process

4.3 Pre-processing of the forum transcripts

In the text-preprocessing stage (Fig. 2), several actions were performed in order to remove any noise or unnecessary information that existed within the forum transcripts. Firstly, we substituted detailed or textual information of a specific type with shorthand transcripts, as shown in Table 1. There was no substitution for the Python code transcripts due to the importance of this feature for the calculation of the semantic similarities with the CR corpus of the PY course. The next step was to normalize the forum transcripts. All punctuation, special characters and stop words were removed and the rest of each transcript was lemmatized. These were the basic normalizations that were performed. In order to further examine which normalizations could further improve the performance of the classifiers, via trial-and-error, we experimented with several additional normalizations. From this approach, it was found that the removal of verbs increased the classifiers' performance significantly, so we included it in our normalization procedure. For the natural language processing techniques that were described, the spaCy Python module was used due to its good performance [38] and the fact that it supports natural language processing methods for the Greek language.

Table 1 Substitutions of specific textual data

Data	Shorthand
Link to an online resource	[URL]
Attached image	[IMG]
Reference to the name of another user	[USER]
Reference to a video lecture	[VIDEO]
Reference to a book resource	[BOOK]

4.4 Feature extraction

For the context of this study, the feature extraction method that could be used should had been appropriate for the calculation of the semantic similarities between normalized query documents. For this reason, the Word2Vec word embedding technique was implemented within our methodology. Word2vec is a two-layer neural net that processes text documents and extracts a set of vectors, which are the distributed numerical representations of word features [39]. This method was appropriate for the context of this study due to its utility to group the vectors of similar words, in terms of context, into vector space together. Each vector, used to represent a word of a document, is called *neural word embedding*. In our procedure of building the supervised models (Fig. 2), we use Word2Vec in order to extract these features from each course's corpora and its forum *starting posts*.

4.5 Building corpora and Semantic Similarity Extraction

In order to construct the CR corpus per course, we used the corresponding video transcriptions which were included in PDF documents within the MOOC environment. Each video transcription was related to a specific week's video lectures. As a result, the final CR corpus contained so many documents as the number of weeks of the respective course. For the construction of the LR corpus, we retrieved all the available PDF files that provided information related to the logistics of the course (course rules, submissions, assignments, etc). It should be noted that due to the fact that both courses were offered in the same MOOC platform, the files related to logistics were exactly the same and, thus, both courses shared the same LR corpus. All the corpora were preprocessed according to the way we presented in Sect. 4.3 and the document embeddings of each corpus were extracted based on the feature extraction method that was employed (Sect. 4.4). In Table 2, the characteristics of the final corpora are presented.

For the CR corpora (PY and WH), we merged each week's video transcriptions into one single document and, thus, the CR corpora consisted of 6 (PY) documents and 9 (WH) documents, respectively. The two CR corpora were different in terms of size due to the different number of weeks between the courses and the higher duration of the WH video lectures. The final length of the vocabulary that was finally used after the preprocessing for the CR corpora was approximately three times larger for the WH CR corpus. On the other hand, the final LR corpus was very small, as shown in Table 2, due to the fact that the logistics instructions were included in 1-2 pages each. The data that were used to construct the LR corpus were related to five different categories: Assignment instructions, Submission rules, Peer Assessment rules, Certificate instructions and General information about the course material. Therefore, the LR corpus was made of these 5 documents.

Table 2 Characteristics of the PY, WH and Logistics corpora

	PY Corpus	WH Corpus	Logistics Corpus
# of documents	6	9	5
# of sentences	2679	8208	46
# of tokens	55115	121195	1142
Final Vocabulary	13540	33221	129

The features that were used for the training of each supervised model were the semantic similarities of each *starting post* with the respective CR and LR corpora per course. For the calculation of the semantic similarities, there was a need to reduce the dimensionality between the different documents, in terms of size, due to the fact that *starting posts* had a smaller length than each document of a corpus. To address this need, we used the *Cosine Similarity* method. *Cosine Similarity* is a metric that is used to determine how similar the documents are irrespective of their size. Mathematically, it specifically measures the cosine of the angle between two vector representations projected in a multi-dimensional space. This approach was followed in order to calculate the semantic similarity of each *starting post* with each document of a corpus.

The last goal was to extract the semantic similarity of every single *starting post* with the set of documents that a corpus is comprised of. For this purpose, we calculated the average of the semantic similarities of each *starting post* with each document as a corpus. This calculation was performed via the "Semantic Similarity Calculation Module" (Fig. 2), whose inputs are the document embeddings

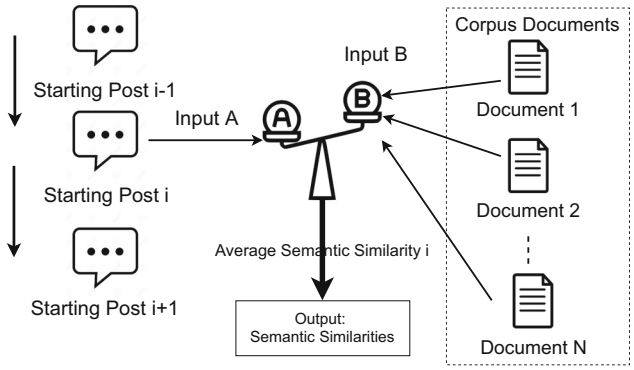

Fig. 4 Semantic Similarity Calculation Module

of a given corpus and the course's *starting posts*. The inner structure of this module is presented in Fig. 4.

4.6 Training and evaluation of the supervised models

In the training phase of the supervised models, we had to perform a multi-class classification to construct the classification models for each course. The classification algorithm that we used was the support vector classifier (SVC). Specifically, we built an one-vs-the-rest (OvR) classifier using the winner-takes-all strategy by employing a radius basis function (RBF) kernel. SVC is used in support vector machine (SVM) classification and attempts to match the data, returning a "best fit" hyperplane dividing or categorizing the data and the OvR is a suitable classification approach for multi-class classification due to its interpretability [40]. SVMs are considered one of the most appropriate methods that can be employed in text categorization problems due to their acceptable performance in comparison with other classification methods [41]. The training data were split into 75% for training and 25% for testing while performing stratified sampling. Following this classification approach for each course, we created two separate models (PY-SVC and WH-SVC) and evaluated each one on the PY and WH test datasets correspondingly.

The next step was to evaluate the transferability of both classifiers. Specifically, we used the classifier (SVC model) of the PY course to perform predictions on the WH *starting posts* and vice versa. For this study, we call this approach as a cross-course evaluation. Through this approach, we would be able to investigate the performance of these models on new unseen data derived from a new course, which in our case can be considered as totally different in terms of context, and thus, we would be able to interpret the performance of these models when transferred.

5 Results and discussion

5.1 RQ1: Interpretation of the extracted semantic similarities with each corpus

As described in Sect. 3, the first goal was to examine for each course through a statistical analysis whether the *starting posts* of each category can be differentiated effectively using the semantic similarities with the two corpora. Toward this goal both graphical and statistical tests have been employed.

In Fig. 5, the scatterplots of semantic similarities of each category's posts with the two corpora are shown for the two courses. It can be observed that the majority of CR

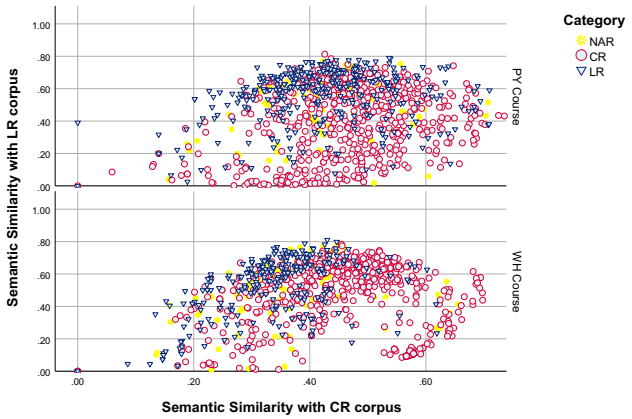

Fig. 5 Semantic similarities of *starting posts* with CR and LR corpora for both courses

posts (red circles) aggregate to the lower right while the LR posts (blue triangles) aggregate to the upper left side. The NAR posts (yellow stars) tend to spread across, almost like noise. Obviously, it will be impossible to differentiate all three categories, however, since only the CR and LR posts require action, it is worth investigating whether the extracted similarities could define meaningful decision boundaries, through a classification model, at least for these two classes.

In order to evaluate and quantify the differences between extracted similarities with the two corpora of each course, we proceeded with a comparative statistical analysis for the mean semantic similarity scores, as calculated for the different post categories of the two courses (Table 3). Based on this analysis, we first compared the mean similarity scores for the three post categories within each course. The comparison was performed using ANOVA separately for the CR and LR Corpus. According to these tests, it can be seen that for both MOOCs there are significant differences between the mean semantic similarity scores of the different post categories either with the CR Corpus ($F(2,977) = 14.30$, $p \ll 0.01$ for PY course and $F(2,309) = 140.51$, $p \ll 0.01$ for WH course) or the LR Corpus ($F(2,278) = 62.18$, $p \ll 0.01$ for PY course and $F(2,321) = 20.84$, $p \ll 0.01$ for WH course). The differences in the mean semantic scores between post categories are shown also graphically (Fig. 6) along with the 95% confidence intervals. Obviously LR posts have the least mean similarity scores with the CR Corpus for both courses and CR posts have the least mean similarity scores with the LR Corpus. Furthermore, in the post hoc analysis (Table 3), one-to-one comparisons revealed that there were significant differences ($p \ll 0.01$ for both corpora) between the CR and LR posts for both courses, while this is not the case when comparing NAR and LR mean similarities scores with the CR Corpus ($p = 0.583$ and $p = 0.117$

Table 3 Comparing Mean Similarities with CR and LR Corpora for the two courses

	PY Course	WH Course	Test of Mean Differences
NAR posts	97 (9.9%)	125 (12.5%)	
CR Corpus	0.436 ± 0.106	0.365 ± 0.130	4.34 (<0.0005)**
LR Corpus	0.512 ± 0.175	0.496 ± 0.214	0.60 (0.552)
CR posts	549 (56.0%)	614 (61.6%)	
CR Corpus	0.464 ± 0.110	0.464 ± 0.118	0.02 (0.986)
LR Corpus	0.407 ± 0.217	0.435 ± 0.223	-2.15 (0.032)
LR posts	334 (34.1%)	258 (25.9%)	
CR Corpus	0.423 ± 0.117	0.338 ± 0.098	9.63 (<0.0005)**
LR Corpus	0.551 ± 0.168	0.527 ± 0.184	1.66 (0.098)
Total	980 (100%)	997 (100%)	
CR Corpus	0.447 ± 0.114	0.419 ± 0.128	5.205 (<0.0005)**
LR Corpus	0.467 ± 0.209	0.466 ± 0.216	0.012 (0.991)
Test for Equality of Means between categories			
CR Corpus	14.30(<0.0005)**	140.51[a] (<0.0005)**	
LR Corpus	62.18[a] (<0.0005)**	20.84[a] (<0.0005)**	
Post Hoc Tests (Multiple Comparisons): Mean Similarity w/ CR Corpus			
NAR vs CR	0.061	<0.0005**,[b]	
NAR vs LR	0.583	0.117[b]	
CR vs LR	<0.0005**	<0.0005**,[b]	
Post Hoc Tests (Multiple Comparisons): Mean Similarity w/ LR Corpus			
NAR vs CR	<0.0005**,[b]	0.013[b]	
NAR vs LR	0.015[b]	0.423[b]	
CR vs LR	<0.0005**,[b]	<0.0005**,[b]	

***Highly significant differences

[a]Based on the Welsh robust test of Equality of Means, due to significant differences among variances [b]

[b]Based on Tamhane's T2 test, due to significant differences among variances

for the two courses) or with LR Corpus ($p = 0.015$ and $p = 0.423$, accordingly). Specifically, CR posts have significantly higher scores with the CR corpus and, accordingly, the LR posts have significantly higher scores with the LR corpus for both courses. This is a very promising finding since it indicates that CR and LR posts on the average give higher scores with their respective corpus, thus they may be differentiated according to their semantic similarities.

On the other hand, our statistical analysis verified that the semantic similarities of NAR posts insert noise to the classification problem. In fact, it was found that the mean similarity scores of NAR and LR posts are not significantly different for any of the two corpora. The picture is approximately the same for both courses. Additionally, for the case of the PY course, the mean similarity scores with the CR corpus between CR and NAR posts were also found to have no significant differences ($p = 0.061$). Given this analysis it becomes quite apparent that distinguishing NAR posts from the other two categories for the PY course, using only their similarity scores with the two corpora is a very difficult task and the outcome completely random. On the contrary for WH course, there are no significant differences between the mean similarity scores of NAR and CR posts with the LR Corpus.

To better understand the similarity values for the NAR posts, we selected all NAR *starting posts* with semantic similarity higher than 0.4 with the CR or LR corpus and tried to identify why their similarities were that high. The reading of those transcripts revealed that they included terms related to logistics, while according to their content they were not related to problems with logistics. The coders, during the coding process described in Sect. 4.2, were prompted to label as LR all *starting posts* that were related to problems with logistics. In the NAR posts with semantic similarity higher than 0.4, learners were seeking information regarding the difficulties of the course (assignments, MOOC procedures) and general information about the course, which could not be considered as "urgent" by the course staff. Therefore, a number of content-related and course logistics-related terms were accumulated within these posts and resulted in such high similarity

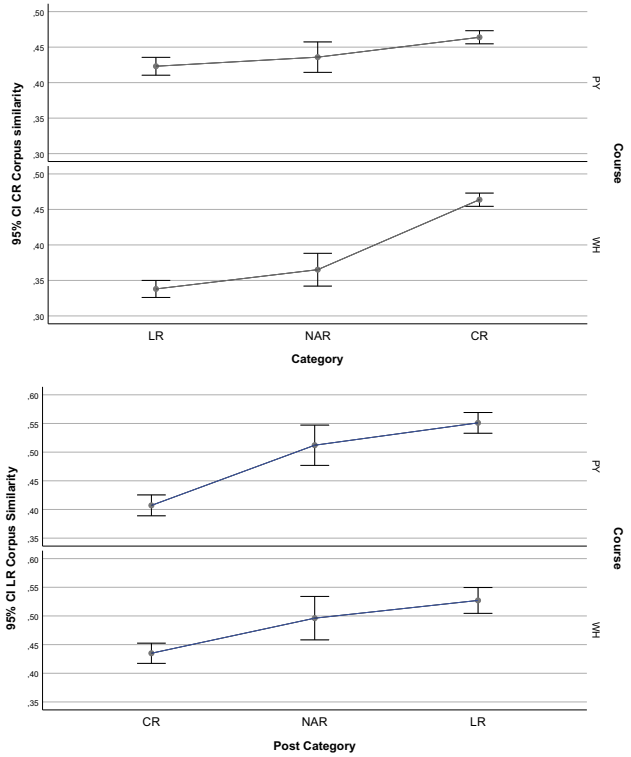

Fig. 6 Mean semantic similarity scores for the CR, LR and NAR post categories with 95% confidence intervals

measures. So this problem was in fact related with the labeling procedure initially followed for the *starting posts*.

Further investigation was performed on the problematic behavior of the NAR posts similarities, by exploring the weekly distribution of these posts for the two courses. In Fig. 7, we present the number of NAR posts per week for both courses. It can be observed that for the PY course, the majority of NAR posts were posted during the first week of the course and a much smaller number (less than 10) during the following weeks. On the contrary for the WH course, the NAR posts were more uniformly distributed across the duration of the course. It should be noted that despite the fact that the duration of the courses was 6 and 9 weeks, respectively, the forum posting activity continued in the following weeks after the end of the course's schedule. To interpret this posting activity, from the qualitative analysis discussed above, it was found that for the PY course, in the first week learners tended to introduce themselves in relevant threads that were created for this reason. This explains the high NAR posting activity during the first week. On the other hand, the different pattern observed in the NAR posting activity for the WH course very possibly is related to the different subject matter and different objectives between the two courses. The PY course can be described as a skill-oriented course, while WH offers a theoretical foundation and, thus, each course attracts learners with different motivations.

Addressing our first research question, the results of the statistical analysis revealed that by using the semantic similarities with the CR and LR corpus, we could sufficiently differentiate the CR and LR posts. Due to their significant differences, it seems that a decision boundary can be formed and thus lead to the creation of a classification mode. On the other hand, the similarity scores of the NAR posts did not have significant differences with the LR posts for both courses, and with the CR posts for the PY course. This indicates that these scores may insert noise to the classification problem that we will attempt to deal with in the following sections.

5.2 RQ2: Building and evaluation of the SVC models

In this part of our study, the main goal was to use the semantic similarities as training features in order to train SVC models, one per course and evaluate their performance on their corresponding test datasets. The decision boundaries formed with the help of these models are presented in Figs. 8 and 9 on the scatterplot of the semantic similarities. It can be observed that, for both courses, the classification algorithm was able to split the area into two (or three for WH) regions, in which the CR *starting posts* (yellow area) and LR *starting posts* (brown area) belong.

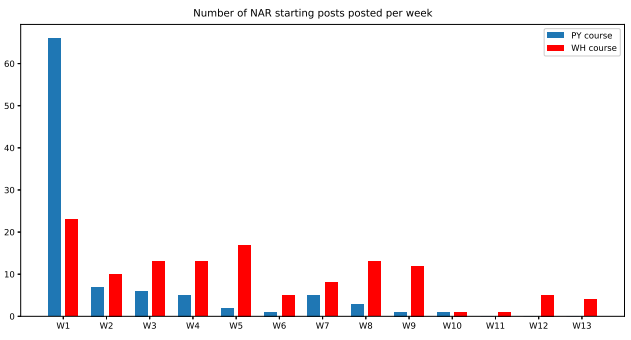

Fig. 7 Distribution of NAR *starting posts* per week for the PY and WH course

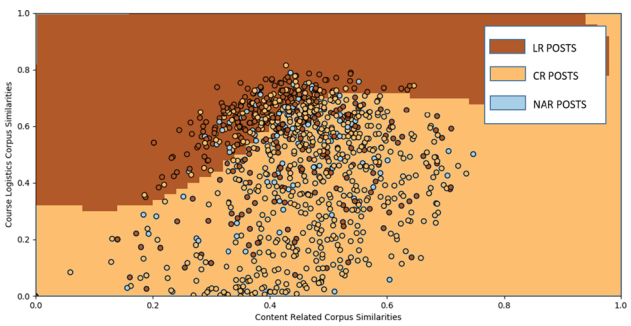

Fig. 8 SVC decision boundaries (RBF kernel) for the PY course

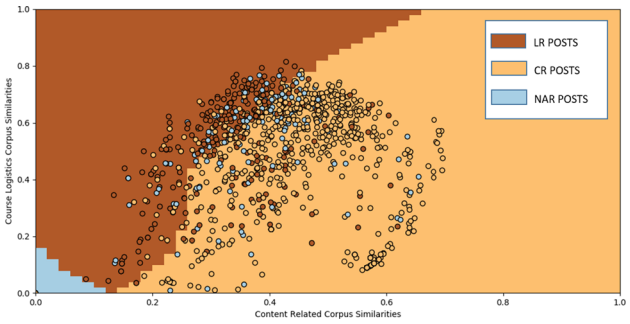

Fig. 9 SVC decision boundaries (RBF kernel) for the WH course

For the PY course, the algorithm failed to assign a region for the NAR posts, while for WH it only assigned a very small region (blue). It can also be observed that there are errors in both regions that lead to misclassifications.

The evaluation metrics of the SVC models as measured from their corresponding test datasets are presented in Table 4. To assess the evaluation metrics, we used the agreement measures that were proposed by Landis and Koch [42]. According to Table 4, it is observed that both classifiers performed *substantially good*, while the WH-SVC scored slightly higher than the PY-SVC model. To better interpret the performance of the models, the corresponding confusion matrices are presented in Tables 5 and 6. Both classifiers achieved higher correct predictions on the CR *starting posts*, 87.6% (PY) and 91.6% (WH). In terms of predicting LR *starting posts*, the PY-SVC model had a *moderate* performance by predicting 42.9% of them correctly, while similar performance was observed for the WH-SVC which predicted 49.2% correctly. On the other hand, both models performed poorly in predicting the NAR *starting posts* correctly, since both models had 0% accuracy.

These findings verify the results of the statistical analysis that we performed in Sect. 5.1. In fact, both classifiers could adequately classify CR discussions, while a moderate performance was observed for the classification of the LR posts in both courses. To further explore this moderate performance, we performed a qualitative analysis on the CR corpora for the misclassified LR posts as CR. Specifically, we investigated the documents of the CR corpus with

Table 4 Evaluation metrics for the SVC classifiers

Metric	PY-SVC classifier	WH-SVC classifier
Accuracy	0.64	0.69
Precision	0.57	0.60
Recall	0.64	0.69
F1 Score	0.59	0.63

Table 5 Confusion Matrix of PY-SVC classifier

	Predict:NAR	Predict:CR	Predict:LR	Total
Label: NAR	0	18	6	24
Label: CR	0	120	17	137
Label: LR	0	48	36	84

Table 6 Confusion Matrix of WH-SVC classifier

	Predict:NAR	Predict:CR	Predict:LR	Total
Label: NAR	0	23	8	31
Label: CR	0	141	13	154
Label: LR	1	32	32	65

which the misclassified posts had high similarity score. It was found that within the CR corpus of each course, the instructor was making often reference to the goal of the corresponding lecture, thus using terms like "*In this video we will...*" or "*In this lecture we will...*". Terms like "*video*" or "*lecture*" were also found within the LR corpus, but they were mostly related with the problems and the quality of the video lectures. It was also found that in several LR posts from the PY course, learners were having submission problems and they were including their code within the post, so it is possible that this fact contributed to increased CR similarity scores.

The most significant problem that was observed, was the poor performance in classifying NAR starting posts. The fact that there was no decision area for the NAR posts in PY and a small one in WH course also verifies the results of our statistical analysis. A possible reason of the classifiers' poor performance may be related to the absence of a NAR corpus. To classify NAR posts, we expected that the similarities with both corpora would be low, but the evaluation results confuted our initial hypothesis. Therefore, it was not feasible to build a NAR corpus due to the fact that NAR posts do not require intervention by the course staff so they may be related to any community-building topic of discussion. On the other hand, there may be several ways to alleviate this issue. First of all, NAR posts are the minority of the total dataset with approximately 10% of all *starting posts* per course. Thus, the trade-off can be described as small, because in a real case scenario, the course facilitators and moderators would be burdened with a relatively small additional number of posts. Furthermore, the accumulation of NAR posts in the first week (Figure 7) for the PY course may help in locating these posts within the forum and excluding them from the classification process.

To address our second research question, it can be deduced that semantic similarities can be used to train adequately a supervised classification model and the evaluation results were found to be promising. The results revealed that further research could be performed on ways to differentiate CR from LR posts even more effectively. For the case of NAR posts, despite the poor performance of the SVC models in classifying them, the trade-off was relatively small and their occurrence through the weeks was found to be predictable for the PY course. Thus, it is possible via custom approaches to exclude a number of these posts so as to increase the performance of the classifiers.

5.3 RQ3: Transferability of the SVC models

To evaluate the transferability of the SVC models, we used the PY-SVC model to classify the WH *starting posts* using the semantic similarities scores calculated with the WH corresponding corpora and vice versa. The evaluation of this cross-course transfer is presented in Table 7 and the confusion matrices in Tables 8 and 9.

It can be observed that both models performed *substantially good* [42] when performed classifications of new and unseen *starting posts*. In fact, we observe that the performance of the "foreign" classifiers (Table 7) on the average is slightly worse than the performance of the "domestic" ones (Table 4), recording just minor losses due to transferability. The positive outcome is that both models preserved their *substantially good* performance via the usage of semantic similarities with their own corpora as training features.

The confusion matrices in Tables 8 and 9 also reveal several noteworthy results. Despite the fact that both models misclassified all NAR *starting posts* of the other course, the PY-SVC classifier predicted correctly 79% of the WH CR *starting posts* and the WH-SVC classifier 93.6% of the PY CR posts. For the LR *starting posts* of each course, the corresponding correct predictions were 65.1% for the WH dataset and only 27.2% (*fair* performance) for the PY dataset.

Table 7 Evaluation metrics for cross-course application of SVC models

Metric	WH-SVC to PY dataset	PY-SVC to WH dataset
Accuracy	0.62	0.65
Precision	0.56	0.59
Recall	0.62	0.65
F1 Score	0.54	0.61

Table 8 Confusion Matrix: WH-SVC on PY dataset

	Predict:NAR	Predict:CR	Predict:LR	Total
Label: NAR	0	82	15	97
Label: CR	1	514	34	549
Label: LR	1	242	91	334

Table 9 Confusion Matrix: PY-SVC on WH dataset

	Predict:NAR	Predict:CR	Predict:LR	Total
Label: NAR	0	60	65	125
Label: CR	0	485	129	614
Label: LR	0	90	168	258

In order to visualize the fair performance of the WH-SVC in classifying LR starting posts when transferred to the PY dataset, we placed the decision boundaries of the WH-SVC on the PY similarities scatterplot (Fig. 10). Obviously, due to the different curvature of the two decision boundaries, the majority of the LR posts are now found within the CR decision area (yellow color), thus resulting to misclassifications. For the same reason, the WH-SVC has much higher accuracy (93%) in classifying the CR posts. On the other hand, when the PY-SVC decision boundaries are placed on the similarities scatterplot of the WH course, a more balanced outcome is achieved in terms of classification errors among CR and LR posts (Fig. 11). From a statistical point of view, the observed level of accuracy in post category predictions can be explained using the results shown in the last column of Table 3. Testing the mean similarity differences between courses it turns out that for the CR posts and LR posts their mean similarities with their corresponding corpora are not significantly different ($t(1158) = 0.02$, $p = 0.986$ for CR posts and $t(590) = 1.66$, $p = 0.098$ for LR posts); while their mean similarities with cross-corpora are significantly different ($t(1161) = -2.15$, $p = 0.032$ for CR posts and

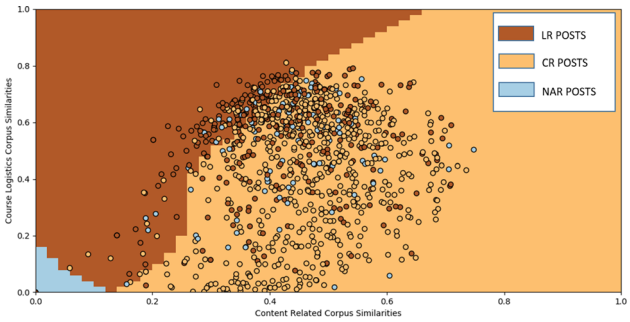

Fig. 10 WH-SVC decision boundaries placed on PY semantic similarities scatterplot

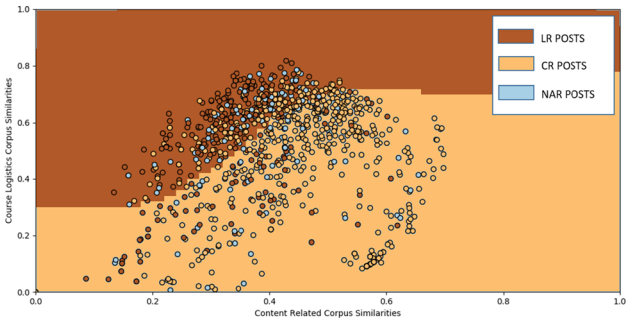

Fig. 11 PY-SVC decision boundaries placed on WH semantic similarities scatterplot

$t(585) = 9.63$, $p \ll 0.01$ for LR posts). Since the mean similarity of the LR posts with the CR Corpus is significantly less for the WH Course and the mean similarity of the CR posts with the LR Corpus is significantly more for the same course the classifier trained with the WH dataset will misclassify heavier the LR posts in the PY dataset and the classifier trained with the PY dataset will misclassify heavier the CR posts in the WH dataset.

To address our final research question, the findings of the cross-course evaluation revealed differences in terms of performance of the models in classifying *starting posts* when transferred. According to the evaluation metrics (Table 7), at first sight both models seem to have minor losses in terms of transferability. The confusion matrices (Tables 8 and 9), though, revealed that the accuracy of the WH-SVC cannot be considered as acceptable due its fair performance in classifying LR posts. On the other hand, a very promising accuracy was found for the PY-SVC when transferred and the performance of the model can be considered as acceptable. Finally, the accuracy results related to the NAR were poor for both courses.

6 Conclusion and future implications

In this case study, we addressed the problem of automated support of MOOC discussion forums. In particular, we focused on the feasibility of transferring such support between two MOOC courses of different disciplines (humanities and technology). We modeled the support of forum discussion as a multi-class classification problem. The classification schema that we used, involved classes of different MOOC actors that are responsible to intervene and support learners. We consider that a recommendation tool based on this schema may especially benefit course facilitators and moderators since they would be asked to deal only with forum posts of their own interest. Our study, that involved a natural language processing approach, revealed promising results in addressing transferability issues of the developed supervised models.

Our approach involved calculating the semantic similarity of each forum post to the text corpora that were dynamically created from the course instructional material on one hand, and instructions about course logistics, like assignment submissions etc., on the other. The developed classifiers were able to differentiate the posts of the content-related (CR) and logistics-related (LR) categories at a satisfactory level. In particular, it was found that despite the difficulties in classifying correctly the posts of the no action required (NAR) category, the results of our statistical analysis revealed that the *starting posts* belonging to the CR and LR categories had significantly higher similarities with their respective corpora. This observation explains the evaluation results of the SVC classifiers on their corresponding test data. Thus, the proposed technique, that was based on usage of automatically produced video lectures transcripts as instructional material for each MOOC, produced interesting results, that may help reducing the computational and human resources needed to build a supervised classification model. On the other hand, the results implied that further research should be performed on issues related to the classification of NAR that the proposed approach failed to address.

In order to tackle the main goal of our study, we performed a cross-course evaluation of the developed classifiers, i.e., we used the classifier developed for the technology course in the humanities MOOC and vice versa. The results of this cross-course evaluation were also quite promising, indicating that such models may be transferred across courses. Specifically, the technology (PY-SVC) classifier had approximately the same accepted performance on the new course data compared to its performance with the corresponding test dataset of the original course. On the other hand, the humanities (WH-SVC) classifier displayed better performance on the CR posts of the technology course (PY) dataset, but worse on the LR posts. This unbalanced behavior was interpreted via the visual framing of the WH-SVC decision boundaries on the PY similarity scores (Fig. 10). It was found that the different shape of the boundary between the CR and LR categories resulted in such an unbalanced performance. In fact, the different shape of the decision boundaries of the two models may be an interesting attribute that can be used in order to calculate the "semantic distance" between two courses. Thus, via the implementation of the proper adjustments to the decision boundaries of a model, such unbalanced behavior might be mitigated.

For our future research, we plan to extend the study to more courses, with different degrees of similarity. The objective is to further explore how the proposed approach behaves in course domains which are more similar in terms of subject matter, like for instance subsequent offerings of the same course. Taking into consideration the findings of this study, our goal is to further investigate how the decision boundaries of similar courses perform when

transferred and thus attempt to define a way of calculating the "semantic distance" between courses of different degrees of similarity.

An important issue that needs further exploration is related to the asymmetric CR and LR corpora. Particularly, the LR corpus was significantly smaller than the CR corpus per course and further research should be conducted in order to enrich it with more features. As we revealed in our previous study [43], LR linguistic features of the forum transcriptions were found to be transferable between the technology (PY) and humanities (WH) course. Therefore, a possible approach would be to construct the LR corpus from a large number of LR forum starting posts rather than use the available documents of the MOOC platform. Finally, another interesting future direction of research would be to investigate ways of preprocessing the text corpora of instructional material, in such a way to alleviate observed similarities across courses, and thus to improve the performance of the classifiers. Via this approach we would be able to define a corpus structure where the extracted semantic similarities could lead to more accurate predictions.

The exploration of how other classification methods (as e.g., LSTM neural networks, boosting, or cost-sensitive techniques) behave when applied on our classification problem may provide important insights in terms of improving the classifiers' performance. Moreover, an interesting future approach is to examine the performance of these models on courses whose posts are written in languages other than Greek in order to study the transferability of the proposed approach to different languages.

Finally, regarding the poor performance of both classifiers in classifying NAR posts, the employment of cost-sensitive techniques could potentially improve classification results. In this study, the dataset was imbalanced with respect to the three classes and this affected the performance of the classifiers. On the other hand, the penalties or costs regarding the misclassifications of each category were different, e.g., classifying a NAR post as CR does not have the same penalty as the opposite. Therefore, by employing cost-sensitive techniques, we could assign different costs to the different types of misclassification errors, aiming at minimizing the overall cost. These costs could then be taken into consideration during the training phase. The main focus of the current study was on the exploration and interpretation of the *starting posts*' semantic similarities behavior as training features and not on ways to increase the performance of the classifiers. We presented an extensive quantitative and qualitative analysis regarding these features in order to achieve this goal and the interpretation of the results provided us with relevant insights regarding their transferability. However, in our future research, we plan on investigating ways of enhancing the classifiers' performance within the same context using, e.g., other classifiers or cost-sensitive techniques.

Acknowledgements This research is performed in the frame of collaboration of the University of Patras with online platform mathesis.cup.gr. Supply of MOOCs data, by Mathesis is gratefully acknowledged. Doctoral scholarship "Strengthening Human Resources Research Potential via Doctorate Research – 2nd Cycle" (MIS-5000432), implemented by the State Scholarships Foundation (IKY) is also gratefully acknowledged. This research has also been partially funded by the Spanish State Research Agency (AEI) under project Grants TIN2014-53199-C3-2-R and TIN2017-85179-C3-2-R, the Regional Government of Castilla y León Grant VA082U16, the EC Grant 588438-EPP-1-2017-1-EL-EPPKA2-KA.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Terras MM, Ramsay J (2015) Massive open online courses (MOOCs): Insights and challenges from a psychological perspective. B J Educ Technol 46(3):472–487. https://doi.org/10.1111/bjet.12274
2. O'Reilly UM, Veeramachaneni K (2014) Technology for mining the big data of MOOCs. Res Pract Assess 9:29–37
3. Kizilcec RF, Piech C, Schneider E (2013) Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Learning analytics & knowledge, pp 170–179. https://doi.org/10.1145/2460296.2460330
4. Kennedy G, Coffrin C, De Barba P, Corrin L (2015) Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance. In: Learning analytics & knowledge, pp 136–140
5. Liyanagunawardena TR, Parslow P, Williams SA (2014) Dropout: MOOC Participants' Perspective. In: European MOOCs Stakeholders Summit, pp 95–100
6. Hecking T, Chounta IA, Hoppe HU (2017) Role modelling in MOOC discussion forums. J Learn Anal 4(1):85–116. https://doi.org/10.18608/jla.2017.41.6
7. Kumar M, Kan MY, Tan BC, Ragupathi K (2015) Learning Instructor Intervention from MOOC Forums: Early Results and Issues. In: Educational data mining, pp 218-225
8. Wiley DA, Edwards EK (2002) Online self-organizing social systems: The decentralized future of online learning. Q Rev Distance Educ 3(1):33–46
9. Drachsler H, Kalz M (2016) The MOOC and learning analytics innovation cycle (MOLAC): a reflective summary of ongoing research and its challenges. J Comput Assist Learn 32(3):281–290. https://doi.org/10.1111/jcal.12135
10. Ntourmas A, Avouris N, Daskalaki S, Dimitriadis Y (2018) Teaching assistants' interventions in online courses: a comparative study of two massive open online courses. In: Pan-Hellenic conference on informatics, pp 288-293. https://doi.org/10.1145/3291533.3291563
11. Peters VL, Hewitt J (2010) An investigation of student practices in asynchronous computer conferencing courses. Comput Educ 54(4):951–961. https://doi.org/10.1016/j.compedu.2009.09.030
12. Brinton CG, Chiang M, Jain S, Lam H, Liu Z, Wong FMF (2014) Learning about social learning in MOOCs: From statistical

analysis to generative model. IEEE Trans Learn Technol 7(4):346–359. https://doi.org/10.1109/TLT.2014.2337900
13. Rowe M (2018) Operating at the Limit of what was Possible: A case study of facilitator experiences in an Open Online Course. Curric Teach 33(2):91–105. https://doi.org/10.7459/ct/33.2.06
14. Ntourmas A, Avouris N, Daskalaki S, Dimitriadis Y (2019) Evaluation of a Massive Online Course forum: design issues and their impact on learners' support. In: IFIP conference on human-computer interaction, pp 197-206
15. Ntourmas A, Avouris N, Daskalaki S, Dimitriadis Y (2019) Teaching Assistants in MOOCs Forums: Omnipresent Interlocutors or Knowledge Facilitators. In: European conference on technology enhanced learning, pp 236-250
16. Sharif A, Magrill B (2015) Discussion forums in MOOCs. Int J Learn Teach Educ Res 12(1):119–132
17. Fu S, Zhao J, Cui W, Qu H (2016) Visual analysis of MOOC forums with iForum. IEEE Trans Vis Comput Graph 23(1):201–210. https://doi.org/10.1109/TVCG.2016.2598444
18. Wong JS (2018) Messagelens: A visual analytics system to support multifaceted exploration of MOOC forum discussions. Visual Inf. 2(1):37–49. https://doi.org/10.1016/j.visinf.2018.04.005
19. Chandrasekaran MK, Kan MY, Tan BC, Ragupathi K (2015) Learning instructor intervention from mooc forums: Early results and issues. In: Educational data mining, pp 218-225
20. Chandrasekaran MK, Epp CD, Kan MY, Litman DJ (2017) Using discourse signals for robust instructor intervention prediction. In: AAAI conference on artificial intelligence, pp 3415–3421
21. Yang D, Piergallini M, Howley I, Rose C (2014) Forum thread recommendation for massive open online courses. In: Educational data mining, pp 257–260
22. Howley I, Tomar GS, Ferschke O, Rose CP (2017) Reputation systems impact on help seeking in mooc discussion forums. IEEE Trans Learn Technol 99(1):1–14. https://doi.org/10.1109/TLT.2017.2776273
23. Ntourmas A, Avouris N, Daskalaki S, Dimitriadis Y (2018) Comparative study of MOOC forums: Does course subject matter?. In: ICT in Education, pp 1–8
24. Moreno-Marcos PM, De Laet T, Muñoz-Merino PJ, Van Soom C, Broos T, Verbert K, Delgado Kloos C (2019) Generalizing predictive models of admission test success based on online interactions. Sustainability 11(18):4940. https://doi.org/10.3390/su11184940
25. Ferguson R, Clow D, Macfadyen L, Essa A, Dawson S, Alexander S (2014) Setting learning analytics in context: Overcoming the barriers to large-scale adoption. In: Learning Analytics And Knowledge, pp 251-253. https://doi.org/10.1145/2567574.2567592
26. Gašević D, Dawson S, Siemens G (2015) Let's not forget: Learning analytics are about learning. TechTrends 59(1):64–71. https://doi.org/10.1007/s11528-014-0822-x
27. Shatnawi S, Gaber MM, Cocea M (2014) Automatic content related feedback for MOOCs based on course domain ontology. In: Intelligent data engineering and automated learning, pp 27-35. https://doi.org/10.1007/978-3-319-10840-7_4
28. Atapattu T, Falkner K (2016) A framework for topic generation and labeling from MOOC discussions. In: Learning at Scale, pp 201-204. https://doi.org/10.1145/2876034.2893414
29. Ezen-Can A, Boyer KE, Kellogg S, Booth S (2015) Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. Learning Analytics & Knowledge, pp 416–150 https://doi.org/10.1145/2723576.2723589
30. Liu W, Kidzićski Ł, Dillenbourg P (2016) Semiautomatic annotation of mooc forum posts. In: State-of-the-art and future directions of smart learning, pp 399-408
31. 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. https://doi.org/10.1016/j.compedu.2017.11.002
32. Boyer S, Veeramachaneni K (2015) Transfer learning for predictive models in massive open online courses. In: Artificial intelligence in education, pp 54-63
33. Whitehill J, Williams J, Lopez G, Coleman C, Reich J (2015) Beyond prediction: First steps toward automatic intervention in MOOC student stopout. Educational data mining, pp 171–178. https://doi.org/10.2139/ssrn.2611750
34. Kizilcec RF, Halawa S (2015) Attrition and achievement gaps in online learning. Learning at Scale, pp 57–66. https://doi.org/10.1145/2724660.2724680
35. Kidzinsk L, Sharma K, Boroujeni MS, Dillenbourg P (2016) On Generalizability of MOOC Models. In: International educational data mining society, pp 406–411
36. Wise AF, Cui Y, Vytasek J (2016) Bringing order to chaos in MOOC discussion forums with content related thread identification. Learning Analytics & Knowledge, pp188–197https://doi.org/10.1145/2883851.2883916
37. Banerjee M, Capozzoli M, McSweeney L, Sinha D (1999) Beyond kappa: a review of interrater agreement measures. Can J Stat 27(1):3–23. https://doi.org/10.2307/3315487
38. Hernandez N, Hazem A (2018). PyRATA, Python Rule-based feAture sTructure Analysis. Language Resources and Evaluation. https://www.aclweb.org/anthology/L18-1330
39. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
40. Duan KB, Keerthi SS (2005) Which is the best multiclass SVM method? An empirical study. In: International workshop on multiple classifier systems, pp 278-285
41. Zhang T, Oles FJ (2001) Text categorization based on regularized linear classification methods. Inf Retr 4(1):5–31. https://doi.org/10.1023/A:1011441423217
42. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174. https://doi.org/10.2307/2529310
43. Ntourmas A, Avouris N, Daskalaki S, Dimitriadis Y (2019) Comparative study of two different MOOC forums posts classifiers: analysis and generalizability issues. In: International conference on information, intelligence, systems and applications, pp 1-8. https://doi.org/10.1109/IISA.2019.8900682