Comparative study of two different MOOC forums posts classifiers: analysis and generalizability issues

Anastasios Ntourmas
Department of Electrical and Computer Engineering
University of Patras
Patras, Greece
a.ntourmas@upnet.gr

Nikolaos Avouris
Department of Electrical and Computer Engineering
University of Patras
Patras, Greece
avouris@upatras.gr

Sophia Daskalaki
Department of Electrical and Computer Engineering
University of Patras
Patras, Greece
sdask@upatras.gr

Yannis Dimitriadis
School of Telecommunications Engineering
Universidad de Valladolid
Valladolid, Spain
yannis@tel.uva.es

Abstract— **Massive Open Online Courses (MOOCs) offer a wide range of opportunities for learning. Their growing popularity has resulted in a large amount of data being available for learning analytics purposes. A major problem of MOOCs is the overwhelming number of posts in their discussion forums. The forum is a key part of the learning process within a MOOC, so this information overload affects negatively the participants' learning experience. Automatic classification of the posts can help searching of relevant information for both the learners and teaching assistants. In this study, we address this problem by building two multiclass classification models, using natural language processing techniques, that classify the posts according to a three-category coding scheme. Each model was created with data derived from a MOOC of different subject matter. The main goal was to evaluate each model's accuracy along with its generalizability to courses of different subject matter. This study contributes to the line of research for automatic classification of forum discussions, ultimately aiming at the development of tools that may assist participants while searching in the forum. Furthermore it provides insights on the main issues that inhibit generalization of classifiers created for a specific subject matter and investigate how their linguistic features relate to this inhibition.**

Keywords— MOOC, learning analytics, discussion forums, classification of posts,

I. INTRODUCTION

Massive Open Online Courses (MOOCs) are intended to offer a wide range of open access educational opportunities to the public. This potential is grounded in their unique strengths of being massive and open [1]. As a result nowadays, MOOC platforms attract thousands of learners with varying motivations and goals [2]. These goals may align with or diverge from the standards set by the instructor, in the latter case participation patterns may not be as expected [3]. Because of their popularity, MOOCs constitute rich sources of learner data available in high volumes for analytic purposes [4]. Moreover, recent years have witnessed a growing body of learning analytics research for MOOCs, including analysis of learners' participation [5, 6] and performance [7, 8], investigation for the factors that lead to learners' retention [9] and automated student modeling [10]. All of these studies highlight the importance of maintaining the dynamics within the MOOC platform, in terms of learner activities, at a high level.

The discussion forum is an indispensable part of a MOOC environment and at the same time an important source of textual data. Prior work has shown that active participation in discussion forums may relate to better course achievements [11, 12]. Participants are expected to use it regularly in order to provide or receive support for challenges encountered during the learning process [13]. Requests for support from learners are satisfied almost exclusively through peer interactions or interactions with the course instructional staff through the discussion forum. While most discussions in the forums are related to the course content, it is very probable that participants may pose questions related to course logistics (e.g. resolving technical problems) or more scarcely open discussions with other participants on topics not related to the course's content (e.g. community building). Due to the massive scale of MOOCs, these different usage forms of the forum by learners may result to information overload [14]. In conjunction with the lack of assistance in choosing posts to read and participate [15], this problem may result in navigation problems that affect negatively their learning experience.

The problem of information overload in MOOC discussion forums has been stated by a number of studies and it is generally accepted that the number of participants is an important factor that contributes to it [16]. We recently investigated the main issues on the discussion forum of the OpenEdX MOOC delivery platform. Our study showed that the forum of such a major MOOC platform is plagued by usability issues in terms of navigation that may scale up in MOOCs with very high participation. It was also found that such issues affect negatively the quality of support that the instructional staff provides to learners. In addition, it has been detected that a significant number of posts created by learners in MOOC forums is often not related directly to the course material [16]. The different nature of the forum posts, in conjunction with the high participation in MOOCs, result to an overwhelming accumulation of data in the forums and to navigation problems for both learners and instructors [17]. As noted in [18] in many MOOC discussion forums there are insufficient search facilities; and due to the large number of

978-1-7281-4959-2/19/$31.00 ©2019 European Union

threads and posts created, users face difficulties in handling the information provided. It is evident therefore that there is need for new methods and tools that could help resolve such issues and facilitate user navigation and experience within the discussion forum. Then learners could be able to locate posts of their interest easily and remain motivated and engaged with the course [19]. Moreover, in order to improve the quality of support that the instructional staff provides to learners, facilitation should be provided for assisting them in locating those posts that seek their intervention [17].

Machine Learning (ML) approaches have been explored mainly towards the development of new tools that can assist MOOC participants. Such approaches aim to analyze discussion forum data and provide automated decision support on a specific context. From the instructors' perspective, research has focused on classification models that can assist their interventions in discussion forums [28, 29]. From the learners' perspective, such models have focused more on addressing specific learner needs (e.g. recommendation systems, etc.) [38]. Such studies suggest that ML can aid in resolving the issue of data overload in MOOC forums and assist MOOC designers in developing useful supportive tools for their participants.

In this study, we focus on the overwhelming amount of data in MOOC forums and develop supervised ML models that can automatically classify discussions within the forum. Our previous studies [8, 20] have shown that the type of interactions taking place between the participants of a MOOC forum may depend on the course's subject matter. Building upon these studies, we aim at constructing classification models for MOOCs of different subject matters. We trained two separate models, one for each different course, and by analyzing the linguistic features extracted per model, we then study their generalizability. The generalizability of such models is a key factor for the future development of such supportive tools for learners or instructors [21]. Building a classification model is a time-consuming process and requires a lot of effort. It was observed that the reliability of a classifier, built from a specific course's data, is questionable on subsequent versions of the same course [32]. It is evident that building a separate classifier for every new course is not a viable solution for MOOC developers. It is therefore important to investigate how classification models built with data related to a specific subject matter (e.g. mathematics) can be expanded to other MOOCs of different subject matter (e.g. humanities). This study aims at providing such insights that will assist the future development of supportive tools for MOOC participants. The goal of such tools will be to improve the participants' experience within the MOOC environment and, by extension, to improve their learning process.

II. Literature Overview

A lot of research has been conducted in the field of automatic understanding of discussions taking place in MOOC forums. Towards this goal, unsupervised machine learning approaches have given promising results. In their studies, Ezen-Can et al. [22, 23] followed unsupervised machine learning techniques to automatically understand the nature of discussion forum posts and dialogues, with the ultimate goal to enable massive-scale automated discourse analysis and mining for better supporting learning. Results showed that such techniques hold promise for future development of tools that could support automatic understanding of dialogue topics within the forum and for building adaptive dialogue systems. In another research, Liu et al [24] followed both supervised and unsupervised methods to automatically annotate forum discussions. Results revealed that this approach can simplify the work of teaching assistants in a MOOC and assist them in finding discussions that seek their interventions. In their study, Attapatu and Falkner [25] also developed an open framework to automatically generate and label discussion topics from MOOCs. Their study showed that in their context of study the results were quite promising but the generalization of their findings need further investigation with other MOOCs. Many studies have also investigated the field of understanding and analyzing the participants' behavior within the MOOC forum through clustering techniques [26, 27].

Supervised machine learning approaches have also provided important insights in the field of automatic classification of forum discussions. In their study, Chaturvedi et al [28] propose various prediction models designed to capture unique aspects of the MOOC forum discussions with the goal of automating instructor interventions. A similar approach has been followed by Kumar et al [29], where they developed a binary prediction model from data derived from 14 MOOCs and tested it on 61 MOOC datasets. Their goal was to classify discussions and predict if an intervention from the instructors is needed or not. A slightly different approach was followed by Ramesh et al [30], where they investigated the topic modeling of forum discussions through a seed LDA approach. In the field of linguistic modeling, Cui and Wise [31] investigated the main linguistic features of forum discussions that are related or not to the course's content by developing a binary classification model. Results showed that linguistic modeling is a promising method and can contribute in finding content-related threads more effectively. In another study, Wise et al [32] address the problem of overload and chaos created in MOOC forums by using a binary classification model that would classify discussions as related to the course material or not. They specifically built their model from data derived from a statistics MOOC and studied its generalizability on a subsequent version of the same course. Results showed that the model demonstrated questionable cross-course reliability.

The different approaches followed by all these studies varied in terms of performance. There was limited investigation on how the linguistic features, extracted for the models, could be related to the classifier's performance. More discussion was held on other numeric features, like number of votes, length of threads etc. In this study, we develop two separate multiclass classification models with textual data derived from two MOOCs of different subject. The main goal of our research is not only to investigate their performance in terms of accuracy but to study the main issues that prevent them from being fully generalizable to other MOOCs of different subject matter, with main focus on the linguistic features of these models.

III. THE CURRENT STUDY

A. Context of study

For our study, we used data derived from two MOOCs offered in 2017 on the mathesis.cup.gr platform, a major Greek MOOC platform based on OpenEdX technology. The first course, 'Introduction to Python' (PY course), was an introductory course to computer programming through Python. The second course, 'World History: Man versus Divine' (WH course), aimed to introduce learners to the history of Asian religions during the Second Circle of World History. The duration of the two courses was 6 and 9 weeks respectively. The two courses were different in terms of subject domain. The PY course was related to technology and WH to humanities. The study was performed on the anonymized discussion forum data.

The discussion forum of both courses was organized according to the three-level architecture shown in Figure 1.

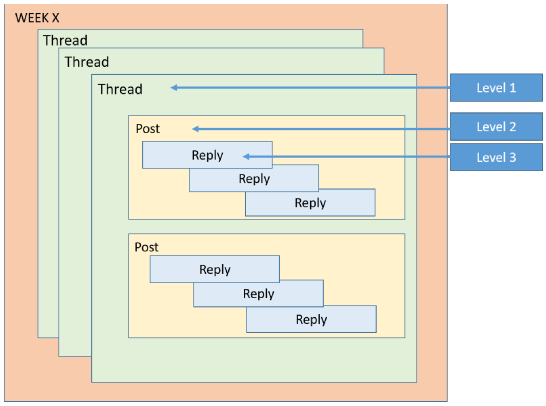

Figure 1: Discussion forum architecture of PY and WH courses

As shown in Figure 1, the discussions of each course were organized in weeks. For each week, participants could create their own threads (Level 1). Within each thread, they could create new posts (Level 2) and within each post they could post replies (Level 3), thus forming discussions. In this particular platform's forum there were two different types of threads. The ones created by the course staff and those by the learners. Threads created by the course staff usually comprised of multiple posts each initiating a different discussion. On the other hand, threads created by learners were mostly related to a single question that usually initiated just one discussion.

Within the context of our study, for the threads created by the course staff, we consider as forum discussion the group of an initial *starting post* and its corresponding replies. For the threads created by learners, we consider as forum discussion the whole thread itself. In this case, the *starting post* is the thread's textual theme and the corresponding replies are the posts and replies that follow. Moreover, we consider the *starting post* of a discussion as an important unit of our analysis. Inspired be the approach followed in Wise et al [32], we assume that the starting post reflects the primary intention of the discussion initiator and sets a direction for the content of all subsequent replies. So, using only the initial *starting post* of a discussion, we attempt to classify each discussion into one of three categories:

a) Non content-related (NCR)

b) Content-related (CR)

c) Course logistics-related (CLR)

This coding scheme was chosen because it can differentiate those discussions that require intervention by either the instructor (CR) or by the platform's administrators (CLR) from the discussions that are not relevant to the course material (NCR).

B. Research issues

According to our coding scheme, the first goal was to explore whether the textual data that appeared in each of the *starting posts* categories was somehow related to the course domain. The courses in our study were different in terms of subject matter and it was worth exploring whether the linguistic features extracted from the *starting posts* of each course could be used for their classification. So the first research question of our study was the following:

***RQ1**: Do the starting posts that belong to the different categories of the coding scheme carry any distinct linguistic features that are related to the course's subject matter?*

By answering positively to this first research question then we would be able to know if it was feasible to create a post classifier for each course.

The second research question of our study was the following:

***RQ2:** Can these linguistic features be used to build a model that reliably classifies starting posts of the discussion forum in each course, according to the proposed coding scheme?*

If it is feasible to create a reliable classification model for each course separately, then it would be worth considering to investigate the cross-course reliability. So the last research question of our study was the following:

***RQ3:** Can a classifier that was built from data of a technology course provide accurate classifications to the starting posts of a humanities course, and vice versa?*

The goal of this research question was to investigate the reliability of these cross-course predictions and investigate if generalizability can be achieved.

IV. METHODOLOGY

In this section, we present the step-by-step methodology that was followed. Firstly, the coding process is described and then the pre-processing of the textual data. We describe the feature-extraction method we performed, the classification algorithm we used to build our model, and the way we performed the cross-course evaluation.

A. Coding of the transcripts

For the purposes of our study we isolated all *starting posts* from the discussion forum data of each course. The coding task was performed by two coders. The principal investigator discussed the coding scheme with the coders, who then coded all *starting posts* for both courses. Their results were evaluated

for using Cohen's kappa (*k*). Cohen's kappa is a chance-corrected measure for interrater reliability that accounts for the possibility of chance agreement between coders [33]. For our coding scheme the data at hand gave k=0.83, indicating quite a high reliability. The goal however, was to achieve absolute consensus for the coding of the transcripts before proceeding to the development of the classification models. For this purpose, the two coders carried a discussion and justified the coding of each *starting posts* until there was no more disagreement and a kappa coefficient of 1 was achieved.

B. Pre-preprocessing of the transcripts

After coding all transcripts, the next step was to pre-process the textual data. The preprocessing stage comprised three actions. First, we replaced specific textual information in the transcripts with shorthand, as shown in Table I. Next, we removed any punctuation or special character from the transcripts. Then we normalized the textual data. For this we lemmatized the transcripts and removed all stop words. For the lemmatization we used the Greek language lemmatizer developed by Petasis et al [34].

TABLE I. SUBSTITUTIONS OF SPECIFIC TEXTUAL DATA

Data	Shorthand
Link to an online resource	[URL]
Attached image	[IMG]
Reference to the name of another user	[USER_REF]
Code posted within the starting post (Python Course only)	[PYTHON_CODE]
Reference to a video lecture	[VIDEO_REF]
Reference to a book resource	[BOOK_REF]

C. Feature extraction

The next step was to extract the main features from the normalized textual data. For this purpose we followed the bag-of-words approach [35], which despite being a unigram representation, it pervades text classification and often achieves high performance. Through this approach, each *starting post* is represented as a vector of indicator variables, one for each word that appears in the training data, i.e. the *n*-th indicator has the value of 1 if the *n*-th word in the vocabulary is present in the *starting post*, 0 otherwise.

This method provides us with the main features that feed the classification model.

D. Classification modeling

In this case study we followed a supervised ML approach. The coding scheme consisted of three different categories and in order to build the classification models for each course, we had to follow a multiclass classification approach, while for the classification algorithm we chose Support Vector Machines (SVM).

SVM are supervised machine learning algorithms used for classification, regression, novelty and outlier detection [36]. We chose SVM instead of the popular Multinomial Naïve Bayes (MNB) method since MNB classifiers are considered to be less accurate compared to the SVM when applied to text categorization problems [37]. For each one of the courses in our study we developed an SVM classification model and each model was built to automatically predict the class of a *starting post* when reading it. The data was split into 75% for training and 25% for testing. We also split the data by category in order to have the same distribution of categories in both the training and test set (stratified sampling).

E. Cross-course evaluation

The last step of our study was to evaluate the cross-course reliability of the classifiers that were created for the two courses. For this purpose we used the classifier created from the training data of the PY course to classify the starting posts of the WH course (PY to WH) and afterwards we followed the same approach inversely (WH to PY). The results of this approach was expected to provide us with insights about the cross-subject applicability of this type of classification and by extension, whether our models presented any overfitting.

To better interpret any generalizability issues of the classifiers, we selected a random sample of misclassified *starting posts* (25% of them) and investigated the reasons that led to such misclassifications by focusing again on the main linguistic features of each classifier.

V. RESULTS

A. Data description

According to the forum data of each course and our definition of *starting posts* within the discussion forums, there

TABLE II. TOP TEN LINGUISTIC FEATURES PER CATEGORY FOR EACH COURSE

Category	PY course	WH course
Non Content-related	hello, programming, IMG, USER_REF, computer, URL, teacher, page, learn, post	support, IMG, summer, listen, hello, USER_REF, internet, course, university, history
Content-related	IMG, idea, chapter, file, USER_REF, way, return, PYTHON_CODE, problem, error	BOOK_REF, USER_REF, god, egypt, religious, IMG, zoroastrism, jew, spain, refer
Course logistics-related	technical, type, appear, help, material, IMG, evaluation, course, submit, way	forward, IMG, empty, dimension, submit, grade, platform, center, history, low

were 980 and 997 total *starting posts* in PY and WH courses, respectively. All starting posts were labeled during the coding process and Table III presents the cross classification of the *starting posts* based on category and course content.

According to Table III, there are not noticeable differences in the distribution of the NCR and CR *starting posts* between the two courses, while the difference between the CLR ones was approximately 9%.

TABLE III. PERCENTAGES OF STARTING POSTS CATEGORIES IN EACH COURSE'S FORUM DATA

Category	PY course	WH course
NCR Not Content-related	9.89%	12.53%
CR Content-related	56.02%	61.58%
CLR Course logistics-related	34.08%	25.87%

B. Extraction of the linguistic features

After preprocessing the textual data, as described in section IV.B, we implemented the bag-of-words approach to extract the main linguistic features from the *starting posts* of each course. The number of features extracted from the PY and WH courses were 727 and 1342, respectively. All features were sorted according to their corresponding coefficients. In Table II we present the top ten linguistic features that correspond to each category for each course. The interpretation of these linguistic features will be performed in the 'Discussion' section.

C. Building the classification models

For the development of the classification models, we used the linguistic features extracted from each course's textual data and created an SVM classifier for each one of the courses. Table IV presents the evaluation metrics for the classification performance on the test data for each SVM classifier. For the assessment of the evaluation metrics we use the agreement measures of Landis and Koch [39].

TABLE IV. EVALUATION METRICS OF EACH SVM CLASSIFIER

Evaluation Metric	PY course SVM Classifier	WH course SVM Classifier
Accuracy	0.69	0.75
Precision	0.68	0.74
Recall	0.69	0.75
F1 Score	0.68	0.75

The evaluation metrics reveal that the SVM classifier of the PY course performed substantially well. On the other hand, the SVM classifier of the WH course performed even better. We present a quantitative representation of these metrics in Tables V and VI, as multiclass confusion matrices for each classifier's evaluation process.

From the confusion matrix of the PY course (Table V) it is evident that most misclassifications occurred for the categories CLR and CR (68 total). On the other hand, in the confusion matrix of the WH course the misclassified transcripts were approximately equal in proportion for the three categories.

TABLE V. CONFUSION MATRIX OF PY COURSE SVM CLASSIFIER

	Predict: NCR	Predict: CR	Predict: CLR
Label: NCR	14	6	4
Label: CR	3	106	28
Label: CLR	2	35	47

TABLE VI. CONFUSION MATRIX OF WH COURSE SVM CLASSIFIER

	Predict: NCR	Predict: CR	Predict: CLR
Label: NCR	10	15	6
Label: CR	12	131	11
Label: CLR	4	14	47

D. Cross-course reliability

Evaluating the two SVM classifiers indicates that their classification ability on the test dataset of each course was substantially good. To investigate the generalizability of the models, firstly we used the PY course SVM classifier to make predictions on the labeled dataset of WH course. Next we followed the same approach inversely. In Table VII we present the evaluation metrics that derived from the cross-course evaluation approach we followed.

TABLE VII. CROSS-COURSE EVALUATION METRICS

Evaluation Metric	PY classifier to WH dataset	WH classifier to PY dataset
Accuracy	0.50	0.53
Precision	0.59	0.61
Recall	0.50	0.53
F1 Score	0.50	0.55

At a first glance, cross-course evaluation metrics seem to be moderate (0.40-0.60). It is evident that almost half of the classifications were wrong. In Tables VIII and IX we present the confusion matrices of the two classifiers using a cross-course evaluation process.

TABLE VIII. CONFUSION MATRIX OF PY CLASSIFIER TO WH DATASET

	Predict: NCR	Predict: CR	Predict: CLR
Label: NCR	49	47	29
Label: CR	123	308	183
Label: CLR	22	60	176

TABLE IX. CONFUSION MATRIX OF WH CLASSIFIER TO PY DATASET

	Predict: NCR	**Predict: CR**	**Predict: CLR**
Label: NCR	38	27	32
Label: CR	88	203	258
Label: CLR	36	50	248

The confusion matrix of PY classifier (Table VIII) reveals that the misclassification rate on the WH dataset was high for all three categories. Specifically, in the case of the CR *starting posts,* the classifier predicted correctly approximately half of the posts (306 out of 614). The WH classifier's performance was slightly better but still achieving a moderate score. Another observation that is made from both confusion matrices is that both classifiers performed better in the case of CLR *starting posts.*

VI. DISCUSSION

In this research, we address the problem of the overwhelming aggregation of data in MOOC discussion forums by investigating the automatic identification of the discussions according to the transcripts of their *starting posts.* We used a three-category coding scheme in order to specify the type of the discussion through multiclass classification. Due to the fact that the type of interactions and dialogues that take place within the forum may depend on the course's subject matter [8, 20], we focus our research on two MOOCs in which their subject matter belongs to two different fields: technology and humanities. By creating two classification models one for each course, we perform a cross-course evaluation in order to investigate if generalization is feasible and what are the main factors that prevent it.

A. RQ1: Linguistic features of each course's starting posts

Before the creation of the classification models for each course, the main linguistic features of the forum *starting posts* were extracted. The features were related to each of the three categories of the coding scheme. According to Table II, it is evident that each of the three categories includes some unique linguistic features, for both courses.

For the *starting posts* of the NCR category*,* there are linguistic features that indicate greeting messages (e.g. *hello*) and others that indicate presentation of users' background (e.g. *programming*, *university*). Other features imply references to the teacher and the course (e.g. *computer*, *teacher*, *learn*, *course*) and a few that indicate social dialogues (e.g. *summer*, *internet*).

The CR category was associated with an adequate number of distinct linguistic features, mainly features that totally relate to the subject matter of the corresponding courses. In the case of the PY course, there are features that can be related to coding issues (e.g. *file*, *way*, *return*, *error*). It was observed that many learners posted their code (PYTHON_CODE) or the image of an error (IMG) in order to ask for help from other learners or the instructors. For the WH course, there were also unique features that were related to the course subject matter. In this case, CR *starting posts* included religion terms (e.g. *god*, *religious*, *jew*, *zoroastrism*) and related locations (e.g. *egypt*, *spain*). There were also a lot of references to book material (BOOK_REF). It can be deduced that the linguistic features of this category for each course are distinct for the corresponding subject matter.

Finally, the category CLR appeared to have also distinct linguistic features for both courses. A post on course logistics may relate to problems with the course environment, (either *technical* or related to the *platform*), the assignment submission process (*submit*, *grade*, *forward*, *type*), or even to the uploaded material (*material*). The top ten linguistic features in this category relate directly to a course's logistics and the terms are independent of the courses' subject matter.

To address our first research question, for each of the two courses, there are distinct linguistic features for each category and a classification modeling can be performed. For the NCR *starting posts*, there are common features between the two courses, but this seems reasonable due to the fact that they can be related to any subject.

B. RQ2: Performance of the classification models

The evaluation metrics (Table IV), resulting from the classifications of the two classifiers on their corresponding test datasets, showed that the PY classifier produced substantially good results. By looking at the confusion matrix of the PY classifier (Table V), the misclassifications that led to such evaluation metrics occurred mainly in the categories CR and CLR. From the 137 CR *starting posts* that the test dataset contained, 106 (77.37%) were classified correctly. There were also 84 CLR *starting posts* in which only 47 of them (55.95%) were correctly classified. Such accuracy rates imply that the PY classifier is not very successful. To further explore the linguistic features that led to wrong classifications, we followed a qualitative approach on a sample of misclassified *starting posts.* It was observed that a number of *starting posts* was related to the course's assignments. These posts were related to either submission problems or content-related questions about the assignments' exercises. The common linguistic features of such posts were words like 'test', 'submit', 'exercise' etc. The PY classifier had linked these features to the CLR category and this led to erroneous classifications. In Figure 2 there is a representative example of a misclassified post in the PY course.

```
Actual Label: Content related
Predicted Label: Course logistics
Document:-
(I have faithfully followed the professor's instructions in the video
and I have tried some of the above advice of our classmates, but i
can't get the correct value. Can someone help me out?
I know that it's not a rating test but i want to submit it on time.)
```

Figure 2: Example of misclassified starting post in the PY course

On the other hand, the WH classifier seemed to be more accurate according to its evaluation metrics (Table IV). This can be verified by the WH classifier's confusion matrix (Table VI). It can be seen that the number of misclassified *starting posts* is quite small for every category.

To address our second research question, we observed that the classifiers displayed different levels of accuracy. The WH classifier seemed to be much more accurate in its predictions, while the PY classifier was not as accurate in predicting the CR and CLR categories correctly, which resulted in lower evaluation metrics. From our qualitative approach, we concluded that a possible factor that led to the poor reliability of the PY classifier may have been the way we preprocessed the initial textual data and the feature extractor approach we performed. Firstly, by removing the stop words and by lemmatizing the datasets, all syntactic information of the transcript was lost. Secondly, the bag-of-words approach is a unigram feature extractor method, which also does not provide any syntactic information about the transcripts.

C. RQ3: Generalizability of the models

In the last part of this study, we investigated the generalizability of the classification models we built for each one of the courses. According to the evaluation metrics of this cross-course evaluation (Table VII), neither PY nor WH classifier achieved acceptable accuracy. In fact, evaluation reveals that sometimes more than half of the classifications made were wrong. The confusion matrices of the two classifiers that relate to the cross-course evaluation (Tables VIII and IX) reveal that in terms of NCR and CR *starting posts*, both classifiers failed to make a sufficient number of correct classifications. An example of a misclassified transcript by the PY classifier is presented in Figure 3.

```
Actual Label: Content related
Predicted Label: Not content-related
Document:-
(Is there a relationship between the middle road of the Buddha
and the principle of mediocrity in the Aristotelian ethic?)
```

Figure 3: Example of a misclassified starting post from the cross-course evaluation process (PY classifier)

In Figure 3, it is observed that the linguistic features of the CR *starting posts* in WH course are NCR features of the PY classifier. The classifier was created by unigram features and the fact that CR unigrams don't appear in the transcript, results in classifying it as NCR.

On the other hand, for the CLR category, both classifiers performed better in terms of accuracy. The PY classifier predicted correctly 176 CLR *starting posts* (68.21%) out of 258 and the WH classifier predicted 248 (74.25%) out of 334. It seems that the unigram features of the CLR *starting posts* are more generalizable than the unigrams of the other two categories.

To address our last research question, it was observed through the cross-course evaluation process that neither the PY classifier nor the WH classifier can provide acceptable classifications for the other course. The only remarkable observation was that in the case of course logistics, both classifiers performed much better. These results imply that unigram feature extraction methods cannot provide sufficient data to build a generalizable classifier across courses of different subject. Syntactic information is a key element of the forum transcripts and should be implemented in the linguistic features of the model.

VII. Conclusion and Future Research

In this study, we attempted to address the problem of information overload [14] in MOOC discussion forums by developing classification models that automatically identify discussions according to their content. By building classification models for two MOOCs of different nature, we investigated their performance in each course and their generalizability. The results of this study provide insights about the main issues that prevent such models of being generalizable and suggest that alternative approaches in terms of feature extraction and text pre-processing should be implemented. Within a specific subject matter, unigram feature extractors may be adequate to build a classification model, but our results showed that they prevent it from being generalizable. The fact that bag-of-words feature extraction method removes the syntactic information from the textual transcripts, means that the model's classifications will be based mostly on unigram features. This was verified by our models' cross-course classifications.

Such study can contribute to the potential development of tools that will assist learners and instructors to navigate through the MOOC discussion forum more effectively and by extension improve their learning experience. Such tools may provide recommendations to users related to discussions that they may be interested in participating. In future research, we aim at further investigating generalizability issues by implementing alternative feature extraction methods and also expanding to other subject matters.

Acknowledgments

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.

References

[1] M. M. Terras and J. Ramsay, "Massive open online courses (MOOCs): Insights and challenges from a psychological perspective," British Journal of Educational Technology, vol. 46, no. 3, pp. 472–487, 2015.

[2] P. Hill, "Online Educational Delivery Models: A Descriptive View," EDUCAUSE Review, vol. 47, no. 6, pp. 84–86, 2012

[3] K. Jordan, "Massive open online course completion rates revisited: Assessment, length and attrition", The International Review of Research in Open and Distributed Learning, vol. 16, no. 3, 2015.

[4] U.-M. O'Reilly and K. Veeramachaneni, "Technology for mining the big data of MOOCs.," Research & Practice in Assessment, vol. 9, pp. 29–37, 2014.

[5] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses," Learning Analytics & Knowledge, 2013, pp. 170–179.

[6] G. Veletsianos, A. Collier, and E. Schneider, "Digging deeper into learners' experiences in MOOC s: Participation in social networks outside of MOOC s, notetaking and contexts surrounding content consumption," British Journal of Educational Technology, vol. 46, no. 3, pp. 570–587, 2015.

[7] G. Kennedy, C. Coffrin, P. De Barba, and L. Corrin, "Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance," Learning Analytics & Knowledge, 2015, pp. 136–140.

[8] A. Ntourmas, N. Avouris, S. Daskalaki, and Y. A. Dimitriadis, "Comparative study of MOOC forums: Does course subject matter?," Panhellenic and International Conference, ICT in Education, pp. 1–8, 2018.

[9] T. R. Liyanagunawardena, P. Parslow, and S. A. Williams, "Dropout: MOOC Participants' Perspective," Proceedings of the European MOOC Stakeholder Summit 2014, p. 95.

[10] M. Yudelson, R. Hosseini, A. Vihavainen, and P. Brusilovsky, "Investigating automated student modeling in a Java MOOC," Educational Data Mining 2014, pp. 261–264, 2014.

[11] S. A. Barab and T. Duffy, "From practice fields to communities of practice," Theoretical foundations of learning environments, vol. 1, no. 1, pp. 25–55, 2000.

[12] M. K. Smith et al., "Why peer discussion improves student performance on in-class concept questions," Science, vol. 323, no. 5910, pp. 122–124, 2009.

[13] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, "Studying learning in the worldwide classroom research into edX's first MOOC.," Research & Practice in Assessment, vol. 8, pp. 13–25, 2013.

[14] V. L. Peters and J. Hewitt, "An investigation of student practices in asynchronous computer conferencing courses," Computers & Education, vol. 54, no. 4, pp. 951–961, 2010.

[15] A. F. Wise, F. Marbouti, Y.-T. Hsiao, and S. Hausknecht, "A Survey of Factors Contributing to Learners''Listening' Behaviors in Asynchronous Online Discussions," Journal of Educational Computing Research, vol. 47, no. 4, pp. 461–480, 2012.

[16] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong, "Learning about social learning in MOOCs: From statistical analysis to generative model," IEEE Transactions on Learning Technologies, vol. 7, no. 4, pp. 346–359, 2014.

[17] A. Sharif and B. Magrill, "Discussion forums in MOOCs," International Journal of Learning, Teaching and Educational Research, vol. 12, no. 1, 2015.

[18] J. A. Baxter and J. Haycock, "Roles and student identities in online large course forums: Implications for practice," The International Review of Research in Open and Distributed Learning, vol. 15, no. 1, 2014.

[19] B. K. Pursel, L. Zhang, K. W. Jablokow, G. Choi, and D. Velegol, "Understanding MOOC students: motivations and behaviours indicative of MOOC completion," Journal of Computer Assisted Learning, vol. 32, no. 3, pp. 202–217, 2016.

[20] A. Ntourmas, N. Avouris, S. Daskalaki, and Y. Dimitriadis, "Teaching assistants' interventions in online courses: a comparative study of two massive open online courses," Pan-Hellenic Conference on Informatics, 2018, pp. 288–293.

[21] L. Kidzinsk, K. Sharma, M. S. Boroujeni, and P. Dillenbourg, "On Generalizability of MOOC Models.," Educational Data Mining, International Educational Data Mining Society, pp. 406- 411, 2016.

[22] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth, "Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach," Learning Analytics & Knowledge, 2015, pp. 146–150.

[23] A. Ezen-Can and K. E. Boyer, "Unsupervised classification of student dialogue acts with querylikelihood clustering," Educational Data Mining, pp. 20–27, 2013.

[24] W. Liu, Ł. Kidziński, and P. Dillenbourg, "Semiautomatic annotation of mooc forum posts," in State of-the-Art and Future Directions of Smart Learning, Springer, 2016, pp. 399–408.

[25] T. Atapattu and K. Falkner, "A framework for topic generation and labeling from MOOC discussions," presented at the Proceedings of the Third (2016) ACM Conference on Learning@ Scale, 2016, pp. 201–204.

[26] A. M. F. Yousef, M. A. Chatti, M. Wosnitza, and U. Schroeder, "A cluster analysis of MOOC stakeholder perspectives," International Journal of Educational Technology in Higher Education, vol.12, no. 1, pp. 74–90, 2015.

[27] S. Moon, S. Potdar, and L. Martin, "Identifying student leaders from MOOC discussion forums through language influence," EMNLP 2014 Workshop, Analysis of Large Scale Social Interaction in MOOCs, 2014, pp. 15–20.

[28] S. Chaturvedi, D. Goldwasser, and H. Daumé III, "Predicting instructor's intervention in MOOC forums," Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, vol. 1, pp. 1501–1511.

[29] M. Kumar, M.-Y. Kan, B. C. Tan, and K. Ragupathi, "Learning Instructor Intervention from MOOC Forums: Early Results and Issues.," Education Data Mining, 2015, pp. 218-225.

[30] A. Ramesh, D. Goldwasser, B. Huang, H. Daume, and L. Getoor, "Understanding MOOC discussion forums using seeded LDA," presented at the Proceedings of the ninth workshop on innovative use of NLP for building educational applications, 2014, pp. 28–33.

[31] Y. Cui and A. F. Wise, "Identifying Content-Related Threads in MOOC Discussion Forums.", Learning@ Scale, 2015, pp. 299–303.

[32] A. F. Wise, Y. Cui, and J. Vytasek, "Bringing order to chaos in MOOC discussion forums with content related thread identification," Learning Analytics & Knowledge, 2016, pp. 188–197.

[33] M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha, "Beyond kappa: A review of interrater agreement measures," Canadian journal of statistics, vol. 27, no. 1, pp. 3–23, 1999.

[34] G. Petasis, V. Karkaletsis, D. Farmakiotou, I. Androutsopoulos, and C. D. Spyropoulos, "A Greek morphological lexicon and its exploitation by natural language processing applications," Panhellenic Conference on Informatics, 2001, pp. 401–419.

[35] C. Boulis and M. Ostendorf, "Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams," International Workshop in Feature Selection in Data Mining, 2005, pp. 9–16.

[36] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," European conference on machine learning, 1998, pp. 137–142.

[37] T. Zhang and F. J. Oles, "Text categorization based on regularized linear classification methods," Information retrieval, vol. 4, no. 1, pp. 5–31, 2001.

[38] D. Yang, M. Piergallini, I. Howley, and C. Rose, "Forum thread recommendation for massive open online courses," Educational Data Mining 2014, 2014.

[39] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," biometrics, pp. 159–174, 1977.