sustainability

MDPI

Article

Evaluation of an Algorithm for Automatic Grading of Forum Messages in MOOC Discussion Forums

Raquel L. Pérez-Nicolás, Carlos Alario-Hoyos *, Iria Estévez-Ayres, Pedro Manuel Moreno-Marcos *, Pedro J. Muñoz-Merino and Carlos Delgado Kloos

Department of Telematics Engineering, Universidad Carlos III de Madrid, E-28911 Leganés, Spain; raperezn@pa.uc3m.es (R.L.P.-N.); ayres@it.uc3m.es (I.E.-A.); pedmume@it.uc3m.es (P.J.M.-M.); cdk@it.uc3m.es (C.D.K.)

* Correspondence: calario@it.uc3m.es (C.A.-H.); pemoreno@it.uc3m.es (P.M.M.-M.); Tel.: +34-91-624-8779 (C.A.-H.)

Abstract: Discussion forums are a valuable source of information in educational platforms such as Massive Open Online Courses (MOOCs), as users can exchange opinions or even help other students in an asynchronous way, contributing to the sustainability of MOOCs even with low interaction from the instructor. Therefore, the use of the forum messages to get insights about students' performance in a course is interesting. This article presents an automatic grading approach that can be used to assess learners through their interactions in the forum. The approach is based on the combination of three dimensions: (1) the quality of the content of the interactions, (2) the impact of the interactions, and (3) the user's activity in the forum. The evaluation of the approach compares the assessment by experts with the automatic assessment obtaining a high accuracy of 0.8068 and Normalized Root Mean Square Error (NRMSE) of 0.1799, which outperforms previous existing approaches. Future research work can try to improve the automatic grading by the training of the indicators of the approach depending on the MOOCs or the combination with text mining techniques.

Keywords: MOOC; discussion forum; automatic grading; evaluation; interactions; quality; impact

Citation: Pérez-Nicolás, R.L.; Alario-Hoyos, C.; Estévez-Ayres, I.; Moreno-Marcos, P.M.; Muñoz-Merino, P.J.; Delgado Kloos, C. Evaluation of an Algorithm for Automatic Grading of Forum Messages in MOOC Discussion Forums. *Sustainability* **2021**, *13*, 9364. https://doi.org/10.3390/su13169364

Academic Editors: Waleed Mugahed Al-Rahmi and Qusay Al-Maatouk

Received: 29 July 2021
Accepted: 17 August 2021
Published: 20 August 2021

1. Introduction

Massive Open Online Courses (MOOCs) are one of the new forms of digital education. This concept was born in 2008, but it was not until 2012 that it really became popular [1,2]. This new format of online education allows users from all over the world to access courses taught by leading universities. Global platforms such as edX, Coursera, FutureLearn, or MiríadaX offer thousands of courses free of charge to millions of students [3]. These contents can include, for example, individual or peer-to-peer assessments, video lectures or live video sessions, reading materials, and discussion forums [4].

MOOC discussion forums provide a way of communication among students and between students and instructors. In these forums, participants can ask questions, give their impressions of course content, or even report problems with the platform. Of course, the versatility and relevance of the forums also provide a great source of data, which can offer a lot of information about the users [5]. Thus, analyzing the content of course messages could be a good practice and could even be used as part of the summative assessment of the course. Discussion forums are not only specific for MOOCs and they exist in other educational platforms, but MOOCs emphasized their use with many users taking advantage of them.

One of the main problems with forums in MOOCs is usually the large number of messages posted. Most of the times, instructors are not able to read and review all the messages posted in the forum [6]. For this reason, it is also common for instructors to use assistants to manage the forum and answer questions or help users during the course [7]. In other cases, it is the community of learners itself the one that resolves the doubts of

other colleagues, thus contributing to the sustainability of the course through the creation of a community around the forum. Moreover, in most cases, the messages posted by learners are appropriate and the interaction enriches the user experience. Nevertheless, some studies have shown that some users may take advantage of the open access to MOOC contents and forum to post messages that are totally irrelevant to the course and to the rest of the students [8].

Therefore, a good way to reward those users who use the forum correctly adding value by helping fellow students or posting messages of interest, and penalize those who do not make good use of the forum, could be to include forum participation as part of the course assessment, but not only the quantity of the participation but the quality. This would encourage interaction between users and might ensure good practices in the forum. But as previously stated, manual review of all messages generated in a MOOC forum can be a tedious and unfeasible task [9]. Hence, learners' automatic assessment based on MOOC forums could be considered as an alternative.

The existing related literature contains several studies in which the quality of interactions in MOOC discussion forums is evaluated [10–14]. However, these articles focus only on the quantitative aspects of these interactions (such as the number of messages posted), but do not typically consider the content of these messages. For this reason, the aim of this work is to develop and evaluate an automatic grading model in which user participation in discussion forums is evaluated analyzing aspects of their content, their impact, and also of the user's activity. This algorithm consists of evaluating different metrics belonging to each of the three dimensions mentioned before, and assigning weights to these metrics according to their importance or relevance in each specific course. In addition to the design of the algorithm, an evaluation of the algorithm should be performed in a specific scenario, with certain weights, and with the participation of experts to verify the correct grading of user interactions. In order to accomplish this, the following research question (RQ) is proposed:

RQ: What is the accuracy and Normalized Root Mean Square Error (NRMSE) that can be obtained to assess students' grades in forum messages with an approach with fixed metrics and selected weights by experts by default?

This paper is structured as follows. Section 2 discusses published articles related to MOOC discussion forums, and research studies focused on user interactions in other open forums. Section 3 describes the dimensions and the functioning of the automatic grading algorithm. Section 4 details the materials and methods used in the work presented, including the description of the data used, the application of the automatic grading algorithm designed in edX, and the specific scenario in which it has been evaluated. Section 5 presents the results obtained for the different test performed. Section 6 discusses the obtained results, the implications, and the possible limitations of the study. Finally, Section 7 includes the conclusions of the research and possible future work directions

2. Related Work

Studies conducted on the data obtained from MOOC discussion forums are numerous [15,16]. However, only few of the related articles focuses on assessing the quality of interactions in the forums. In contrast, this type of analysis is more frequent and popular in other Question and Answers (Q&A) forums. Studies focused on Q&A sites such as Quora or StackOverflow are particularly noteworthy, as the structure and functioning of these forums is very similar to that of MOOCs discussion forums [17–19]. Moreover, these two forums are the most popular Q&A sites, being the former a general purpose forum and the latter focused on computing.

2.1. Open Q&A Forums Analysis

There are several articles that study the messages posted on Quora with very diverse objectives [20–22]. However, there are some common features that are used to perform

an analysis of the quality of the messages. For example, Maity et al. [23] tried to predict whether a question would be answered or would remain open (answerability) by performing a linguistic analysis at message level (length, sentiment analysis, n-gram counts, topic diversity, etc.) and at user level (linguistic style), and also by analyzing user activity. In this way, it is understood that the most answerable questions are those that are more complete and have higher quality content. Their model achieved a maximum prediction accuracy of 76.26% and they concluded that linguistic analysis metrics were the most discriminative features.

Patil et al. [17] relied on users' activity (Activity Features—AF), the quality of their messages (Quality of Answer Features—QAF) and their linguistic (Linguistic Features—LF), and temporal characteristics to detect experts on Quora. Therefore, they studied several characteristics such as the number of followers, the number of published messages, the length of those messages, their entropy, or their readability (using the Simple Measure of Gobbledygook—SMOG formula [24]), among others. The results of their study showed an accuracy of up to 97%, so that the classifiers used could reliably identify experts on different topics.

Finally, Roy [25] used multilayer convolutional neural networks to detect low-quality messages on Quora. In this case, messages were ranked according to their quality using attributes very similar to those of the two previous studies. Roy managed to improve the results obtained in other research, achieving an accuracy of 98% in the best case.

As for the articles that analyze the StackOverflow forum, it is possible to find papers with very varied objectives. Regarding those articles in which the quality of posts is studied, Ponzanelli et al. [18,26] analyzed the own metrics of StackOverflow (length, replies, votes, tags count, etc.) and others related to the readability and popularity of the messages; Duijn et al. [27] and Arora et al. [28] analyzed the type of question according to its content; and Roy et al. [29] analyzed up to 26 different textual and non-textual features.

However, there are other articles in the literature that uses StackOverflow data to: detect the most voted messages based on textual, temporal, semantic, and behavioral features as in the case of Neshati [30]; predict the success of a response based on the quality of the presentation, the speed of the response, or the reputation of the author, as in the case of Calefato et al. [31]; or analyze its reputation system and the contributions of users, as in the case of Movshovitz-Attias et al. [19]. In the latter article, grades were assigned to the different possible actions to calculate users' reputation, and it was concluded that users classified as experts perform more interactions in the forum than other users.

All the aforementioned articles use similar features to infer or calculate the quality of the messages in the forums of both Quora and StackOverflow, and all of them obtain quite satisfactory results in terms of the accuracy of their models. Since the functioning of these two open forums is very similar to that of MOOC forums, the features used here could also be used in the context of online education in general, and MOOCs in particular.

2.2. MOOC Discussion Forums Analysis

It is very common for users to use the course forum in a MOOC to ask questions or express their opinion about the course. So, the messages posted in the discussion forums throughout the course are a great source of information from which it is even possible to infer the student's performance. However, processing all the messages generated in a course manually is virtually unfeasible. For this reason, a large amount of research analyzing different aspects of MOOC discussion forums has been published in recent years [7,13,32].

It is possible to find in the literature several articles that do not analyze the information that can be obtained from the messages as such, but rather the type of content of these messages. Some studies have been oriented towards the classification of messages into different types according to their content. Ntourmas et al. [7] classified the content of messages into three categories: course-related problems (CR), course logistics-related problems (LR), and community-related discussions that do not require action by the course

moderator (NAR). In their study, they used different elements such as video transcripts to calculate the similarity of the posted messages with the different possible topics, and thus classify them into the different categories. However, in this first approach they only achieved a maximum accuracy of 69%.

In this same line of research, the articles of Stump et al. [33] and Cui et al. [34] are also noteworthy. In the latter, they used different linguistic features to identify six types of messages or structures: general discussion messages, presentation messages, Q&A messages, messages about course assignments, messages about technical problems, or Open Learning Initiative (OLI) Textbook questions. Their analysis showed that only 28% of the messages posted by students were content-related and, in terms of classification, they achieved a 86% accuracy with the base model.

For its part, Stump et al. [33] classified the messages into eight types according to the topic. They also identified the user role when posting. In this way, the user who posts a message can be classified as a help/information seeker, a help/information giver, or other in case the user does not match any of the two previous types.

There are also articles in the literature that use the information contained in forum messages to detect patterns or predict events. Imran et al. [35] pointed out that one of the biggest challenges of online courses in general, and MOOCs in particular, is the dropout rate. Therefore, they used deep neural network architectures to detect users at higher risk of dropping out. They used different user attributes, including the number of messages posted in the forum, to train their prediction algorithm. They achieved an accuracy of more than 99% in the prediction of dropouts.

Furthermore, Ramesh et al. [36] reviewed previous articles and used a seeded Latent Dirichlet Allocation-LDA model to analyze forum interactions. They predicted the survival or dropout of a user from features such as the number of messages posted or viewed, the number of votes data, or by performing a sentiment analysis of the messages.

In the same line of research, Wen et al. [37] used the result of sentiment analysis of messages posted in the discussion forum of three MOOCs in Coursera to predict the risk of user dropout. This study shows that there is a correlation between the overall tone of forum messages and the number of dropouts. However, they concluded that at user level, sentiment analysis by itself is not a reliable predictor. They also published another paper in which they evaluated the level of user engagement using the number of posts and performing a linguistic analysis of the messages [38].

Other studies focused on analyzing the different types of users based on their contribution to the forum can be found beyond message classification, user engagement and dropout prediction. For example, Wong et al. [14] identified the most active users based on interactions in the forum, focusing on views, replies and thread duration. They also investigated the influence of these users on conversations and concluded that the more active users tend to make more positive contributions.

Other articles such as He et al. [11] or García-Molina et al. [32] addressed the relationship between participation in the forum and performance in the course. He et al. [11] performed both non-parametric tests and multiple linear regressions to test whether there is a correlation between course performance and forum participation. To do so, they collected data from the different tests and activities of a Chinese college course, as well as the interactions in the forum and the final grades obtained. They concluded that there is a positive correlation between these two aspects, since the users who posted more in the forum obtained better grades.

In the case of García-Molina et al. [32], they designed an algorithm to grade messages according to their qualitative and quantitative aspects. Based on the grades assigned to the messages, the overall interaction of the users in the forum was evaluated and its correlation with the final grade obtained in the course was studied. Although a certain correlation was seen, they concluded that participation in the forum could not be used as a unique predictor of users' grade.

Finally, it is worth mentioning the work of Coetzee et al. [13] who proposed a user reputation system for MOOCs. The authors use only quantitative measures of forum participation such as the number of messages posted, the number of responses obtained, or the number of votes both given and received, among others, to calculate user reputation. In addition, each of these variables has a different contribution in the algorithm. This article obtained good results and a strong correlation was seen between a user's reputation in the forum and his or her outcome in the course. However, the research was conducted on a single course, so its scalability is yet to be determined.

The papers presented are varied and use the information obtained in the forums to carry out different research projects. Although some of them, such as those by García-Molina et al. [32] or Coetzee et al. [13] try to evaluate user interactions in some way, these articles only consider quantitative aspects of the messages (replies, votes, views, follows, etc.), and do not perform any type of analysis of their content. Moreover, in some of the research, acceptable results have been obtained, but there is still room for improvement. However, some of the aspects discussed in this subsection, together with the research on interaction quality reviewed in the previous subsection, may be a good starting point for the development of an automatic grading model not only based on quantitative aspects, but also on qualitative and impact features.

3. Automatic Grading Algorithm

The approach defines three dimensions: (1) the quality of the posted messages in terms of their content, (2) the impact of those posted messages, (3) the user's activity in the forum. The first dimension, related to the quality of the messages, aims to analyze several aspects of the content in order to evaluate whether these messages are adequate to the course, appropriate and well-written. When talking about the impact of the messages, this dimension aims to study how much discussion these messages generate in the forum. It is considered that the messages that generate more impact or discussion are those that contribute more to the enrichment of the course and the users. Finally, the evaluation of user activity helps to detect the most active users. It is understood that the most active users in the forum seek to interact with other users and to improve their learning experience. Therefore, this aspect should also be taken into account in the final grading of the users.

Within each of the previous dimensions, different aspects are evaluated. In addition, a different weight has to be assigned to each of these metrics. In this way, it is established which are the most relevant and significant metrics in the model for the calculation of the final grade. Although the following section will detail the application of the algorithm and the assignment of weights in a specific scenario, the general operation of the algorithm is described below.

The metrics in the content quality category aim to evaluate aspects of the content of the messages exclusively, such as their length, readability, entropy or the number of mentions to other users within the body of the message. Then, the impact of the messages was evaluated by counting the number of votes received, the number of responses generated, the number of followers and the number of views obtained. Finally, student activity within the forum was also evaluated to obtain the final grade for user interaction in the discussion forum. For this purpose, the number of messages posted and the average grade of those messages, the number of votes given, the number of messages viewed, the number of messages followed and the number of searches performed in the forum were taken into account. In order to obtain the average grade of the messages posted, the dimensions of quality of content and impact must be taken into account, since these are the ones that grade the messages. In this way, there are two dimensions (content quality and impact) that aim to evaluate the messages themselves, and a third dimension (user activity) that makes use of these two previous dimensions and establishes the user's final grade.

Regarding the evaluation process itself, the approach first filters the messages according to their type and obtains the general information of the message. This includes the identifier of the user who publishes it, the content of the message, the number of votes,

the number of replies (if applicable) or the number of views, among other aspects. Then, the body of the messages is processed to obtain information regarding content quality metrics, such as length, readability, sentiment analysis or entropy. The message grade is then calculated based on the content quality and impact metrics and the weights assigned to each of these metrics. Finally, the events related to user activity in the forum are processed and the final grade is calculated taking into account the weights of this last dimension and the average grade of the messages posted by the user.

In addition, the approach has also been implemented in a sustainable and flexible way, so that it can be adapted to different courses. One of the objectives of this research was to be able to give flexibility and scalability to the automatic grading model designed. To this end, it was decided to create a configuration file that would allow the moderators of future courses to adjust the model to their needs.

In the configuration file, the course administrator has the possibility to adjust the values of certain variables within a recommended range. The variables whose values can be adjusted are those that are more representative and involve a greater modification of the grading algorithm. If any of the available fields are left unchanged, the default values indicated will be taken.

4. Materials and Methods

4.1. Data Description

The data used for the evaluation of the proposed automatic grading algorithm belongs to different editions of a trilogy of MOOCs offered on edX. In particular, these MOOCs are the "Introduction to Java Programming" courses created by the Telematics Engineering Department of the Universidad Carlos III de Madrid (UC3M). The events from the 2015 to 2019 editions of these MOOCs and from both their English and Spanish versions have been used for this research.

Combining the data of the five editions considered, more than 117,000,000 pseudo-anonymized events generated by the 572,082 participants are available. These events contain all the users' interactions on the platform, from their enrollment, to their interactions with the videos or the completion of the assessment activities.

Furthermore, the structure of the edX discussion forum allows different actions to be performed. It is a forum based on three levels of messages, where the first level message that starts a new conversation is called Thread. The second level messages that reply to a Thread are called Responses. Finally, a Comment is a third level message that replies to a Response. Over these 3-level messages a variety of interactions such as upvotes, follows, or searches can be performed. However, not all actions can be performed on all levels of messages.

From the total number of events in the MOOCs, those corresponding to the interactions in the course discussion forum were filtered retaining the following events: creation of Threads, Responses and Comments; viewing of Threads; searches in the forum; upvotes or unvotes of Threads and Responses; follows or unfollows of Threads; endorsement of Threads and Responses (confirmation that they add value to a discussion); pin or unpin of Threads (to make them appear at the top of the list of posts or to undo the action); and abuse flagging of messages (to flag inappropriate posts). Finally, the resulting dataset consists of 287,160 events unique to the forum, of which 57,454 correspond to message postings. Figure 1 shows the distribution of messages posted in each of the editions analyzed. As can be seen, the trend has been downward in the five editions studied. The 2015 edition is the one with the highest number of interactions and messages published, doubling the publications of the second most active edition and multiplying by 8 the interactions of the 2019 edition. This may be due to the fact that the first edition was a synchronous edition in which all students took the course at the same time, while in the following editions, the modality was changed to self-pace so that students could follow the lessons at their own pace. In addition, it can be observed another particularity of user behavior in the discussion forum of these courses. In each of the years studied, the number of Threads

published significantly exceeds the number of Responses and Comments. Therefore, users tend to open new conversations and prefer to start a new thread rather than reply to other messages or review conversations already started. Thus, most of the Threads opened are usually left unanswered.

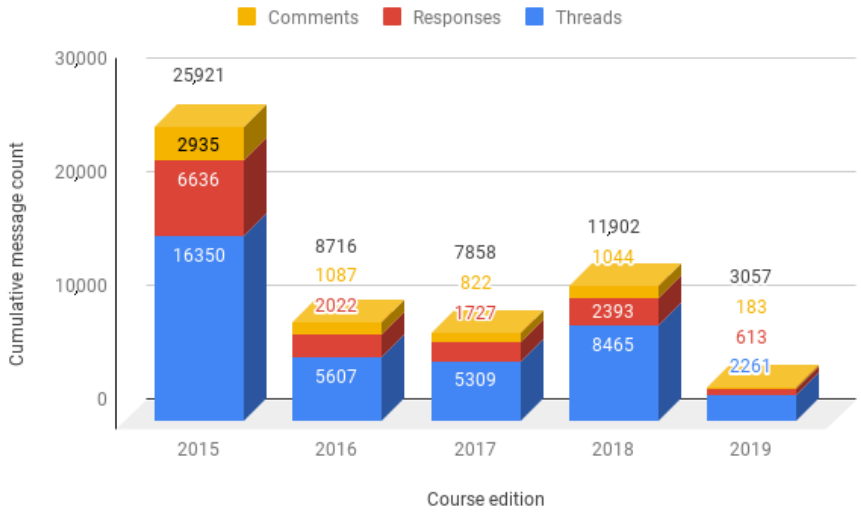

Figure 1. Cumulative course message count by year.

4.2. Application in edX

As detailed in the previous subsection, the data used for the evaluation of this automatic grading algorithm belong to courses offered on the edX platform. The particularity of the edX discussion forum directly affects the designed algorithm. For example, some metrics of the first dimension of content quality are not applicable to the three types of messages existing in edX. Hence, some of the metrics have been marked as non-mandatory. In addition, something similar occurs with the second dimension of message impact. In fact, in the case of the edX forum, third level messages (Comments) can neither receive votes, nor replies, nor can they be followed or viewed. Therefore, the impact of Comment type messages is not evaluated in this algorithm.

Table 1 indicates how each of these metrics has been computed considering the metrics used for the analysis of the different dimensions and the application of the algorithm to the specific case described in Section 4.1.

Table 1. Metrics computation.

Dimension	Metric	Computation Method
Content quality	Length	Word count in the message body.
	Sentiment analysis	TextBlob library [39] which counts the negations and modifier words and use Naive Bayes classifier to determine the polarity value between $[-1.0, 1.0]$.
	Entropy	$entropy = -\sum_{i=1}^{k} P(x_i) * logP(x_i)$ where k is the number of distinct words in the messages and $P(x_i) = \frac{frequency-of-word-i}{total-number-of-words}$ (SciPy library [40]).
	Readability	Flesch Reading Ease formula [41] from the library Textstat [42].
	Mentions	Count of the number of "@" in the message body.
	Endorsement	edX endorsement events count.
	Pin	edX pin events count.
	Ontology terms	Count of the number of terms of the defined course ontology present in the message body.
	Abuse flag	edX abuse flagging events count.

Table 1. *Cont.*

Dimension	Metric	Computation Method
Impact	Votes	edX vote events count for votes received.
	Replies	edX *Response* creation or *Comment* creation events count.
	Follows	edX follow events count.
	Views	edX view events count for views received.
User Activity	Posted messages	edX *Thread* creation, *Response* creation or *Comment* creation events count.
	Average posts grade	Average grade of posted messages obtained from the dimensions of quality of content and impact.
	Given votes	edX vote events count for given votes.
	Viewed messages	edX view events count for messages viewed.
	Followed threads	edX follow events count for messages followed.
	Forum searches	edX search events count.

Finally, regarding the configuration file, Table 2 shows the metrics available for modification, together with a brief description of the variables, the range of suggested values and the defined default value.

As can be seen, the Table 2 does not include all the metrics evaluated by the model. To make the configuration file more manageable and intuitive, but equally functional, this file only includes the 17 most influential and important variables of the model. Some variables such as *max_content_thread* or *max_impact_response* allow the moderator to adjust the weight of the first dimension of content quality and the second dimension of impact in the calculation of the score of the messages (*Threads* and *Responses* individually). Other variables such as *msg_length_average*, *msg_votes_average* or *user_views_average* are intended to determine the average acceptable value for message length, votes received, or number of messages viewed, respectively. The grades for these metrics where the average value is defined are computed considering whether their value is above or below the defined average. In addition, the extra grade variables (*max_content_extra*, *max_impact_extra* and *user_extra*) give the moderators the possibility to decide whether users can get extra points for any of the interactions. Finally, the *max_grade* variable defines the maximum possible grade that users can achieve. That is, it serves to determine whether the grade is to be calculated over 10 points or over 100 points, for example.

The metrics mentioned in Table 1 have not been included in this configuration file in order to keep its use simple yet effective. The variables included in the configuration file and described in Table 2 give the moderator the opportunity to adjust the contribution of the three dimensions studied and provide data on the expected performance of the course in order to assign grades to users appropriately. In this way, the operation of the algorithm can be conveniently adapted to different topics and performances although the specific weight of the particular variables evaluated in the different dimensions of the algorithm cannot be directly modified.

Table 2. Configuration file metrics.

Metric	Description	Suggested Range of Values	Default Values
max_content_thread	Maximum possible contribution/weight of *Threads* content quality metrics in the final grade.	0–10	5
max_content_response	Maximum possible contribution/weight of *Responses* content quality metrics in the final grade.	0–10	5
max_content_extra	Maximum possible contribution/weight of message content quality extra metrics in the final grade.	0–5	2
max_impact_thread	Maximum possible contribution/weight of *Threads* impact metrics in the final grade.	0–10	5
max_impact_response	Maximum possible contribution/weight of *Responses* impact metrics in the final grade.	0–10	5
max_impact_extra	Maximum possible contribution/weight of message impact extra metrics in the final grade.	0–5	2
msg_length_average	Average length of messages in the forum. Messages with a length greater than this value will get the full grade.	100–200	150
msg_polarity_threshold	Polarity (sentiment analysis) threshold of messages in the forum. Messages with a polarity greater than this value will get a proportional grade.	0.2–0.5	0.3
msg_votes_average	Average number of votes of the messages in the forum. Messages with a number of votes greater than this value will get the full grade.	1–10	2
msg_replies_average	Average number of replies of the messages in the forum. Messages with a number of replies greater than this value will get the full grade.	1–10	1
msg_views_average	Average number of views of the messages in the forum. Messages with a number of views greater than this value will get the full grade.	5–20	15
user_views_average	Average number of messages viewed by a user. If a user has viewed a number of messages greater than this value, he/she will get the full grade.	5–20	10
user_votes_average	Average number of votes given by a user. If a user has voted a number of messages greater than this value, he/she will get the full grade.	1–10	5
user_searches_average	Average number of searches in the forum performed by a user. If a user has performed a number of searches greater than this value, he/she will get the full grade.	1–10	5
user_msg_threshold	Threshold for the number of messages posted by a user. If the number of messages posted is above the threshold, the user is rewarded with extra points.	5–20	15
user_extra	Extra points if the user is a top contributor (top 1%).	0–5	3
max_grade	Maximum possible final grade.	Sum of max_content and max_impact	10

4.3. Scenario

The automatic grading algorithm was used to evaluate user interactions in the different editions of the "Introduction to Java Programming" course described in Section 4.1. For the specific case of the MOOC data available, different maximum grades or weights have been assigned for each of the metrics evaluated in each dimension. For example, Table 3 shows the content quality metrics used and the weights assigned to each of them for the evaluation of this dimension. As can be seen, the sum of the weights of the mandatory metrics add up to ten points, while the optional metrics can add up to one extra point to the grade of the messages in this category.

Table 3. Content quality metrics.

Metric	Mandatory	Maximum Grade		
		Thread	Response	Comment
Length	Yes	3	4	5
Sentiment analysis	Yes	3	3	2
Entropy	Yes	2	1	1
Readability	Yes	2	2	2
Mentions	No	0.2	0.4	1
Endorsement	No	0.4	0.6	N/A
Pin	No	0.1	N/A	N/A
Ontology terms	No	0.3	N/A	N/A
Abuse flag	No	if 5 or more, then 0 points		
Maximum total grade		10 points (+1 possible extra point)		

The second dimension is related to the impact or discussion generated by the messages. Table 4 shows the metrics used in this dimension for each type of message and the weights assigned to each of them. In this case, the sum of the metrics is ten points and in this category there are no extra points. In addition, as discussed already before, this dimension of impact is not applicable to *Comments*. In edX these third level messages cannot be replied to, voted on, followed, or viewed, so this second dimension is only evaluated on *Thread* and *Response* type messages.

Table 4. Impact metrics.

Metric	Mandatory	Maximum Grade		
		Thread	Response	Comment
Votes	Yes	3	6	N/A
Replies	Yes	4	4	N/A
Follows	Yes	1	N/A	N/A
Views	Yes	2	N/A	N/A
Maximum total grade		10 points		

Finally, the third dimension (user activity) is used to calculate the final grade of the users. Table 5 indicates the weights assigned to each metric and, as in the previous cases, the sum is equal to ten points. As can be seen, one of the metrics is the average grade of the messages posted. As mentioned in the previous subsection, to obtain this data, the results obtained from the evaluation of the dimensions of content quality and impact should be used. In this scenario, in order to calculate the average grade of the messages posted, the arithmetic mean of both the content quality dimension and the impact dimension has been calculated. Thus, both dimensions contribute equally to the final grade of the messages posted.

Table 5. User activity metrics.

Metric	Mandatory	Maximum Grade
Posted messages	Yes	3
Average posts grade	Yes	4
Given votes	Yes	1
Viewed messages	Yes	1
Followed *Threads*	Yes	0.5
Forum searches	Yes	0.5
Maximum total grade		10 points

It is also worth mentioning that the weights or maximum grade assigned to each of the metrics used in the different dimensions evaluated have been decided taking into account the characteristics of the analyzed courses. In addition, the grades in each particular case have been generally assigned using four ranges and considering the general performance of the forum. In this way, the evaluation is carried out in a fairer and more appropriate way to the data available.

5. Results

Once the metrics to be evaluated and the maximum grades for each of them had been chosen, the model was evaluated. The objective was to check whether the automatic grading model designed was sufficiently accurate to be used in future courses.

The first step of the evaluation consisted of manually labeling a random subset of the messages from the available dataset. Three experts on the topic of the MOOCs (programming) assisted in labeling 580 random messages (1% of the total number of messages in the entire dataset, each message was labeled by the three experts) at four different levels according to their quality, with level 1 being the lowest quality and level 4 the highest quality. In addition, each expert had the possibility to use a second label (secondary label) in case he/she hesitated between two possible levels. Different criteria were used to determine the final level of the messages in case of discrepancies in the manual labels. When the main label of the three experts was the same, the final level of the message was the one determined by the experts. The choice of the final level was also straightforward when two experts agreed on the main label and the third agreed on its secondary label. Finally, when there was total discrepancy in the labels of the messages, the three experts discussed which was the most appropriate level for those messages.

Considering the methodology used to establish the final labels, it was also considered appropriate to study the agreement among the experts in assigning the labels. Since the labels can take four different values, they can be considered an ordinal variable. In addition, since there are three observers (the three evaluating experts), it was decided to use Kendall's W coefficient to obtain the degree of agreement of the main labels assigned in the manual labeling. Finally, after performing the test, a W value of 0.6136 was obtained. This result indicates a substantial agreement among the experts, although it also shows that the evaluation of the messages has an important subjective component. Figure 2 shows the distribution of the messages in the different levels according to the manual labeling performed by the experts.

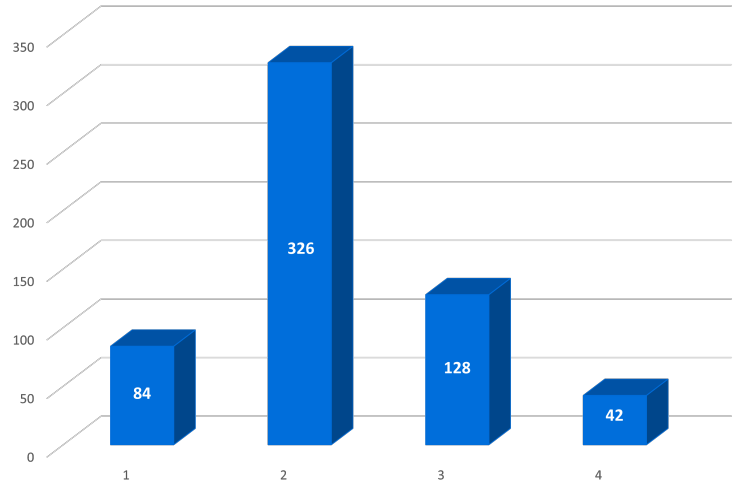

Figure 2. Quality levels distribution by experts.

As can be seen in the Figure 2, almost 80% of the messages were classified with intermediate labels (classes 2 and 3), especially highlighting rank 2. Therefore, most of the

messages were labeled with an average or medium-low grade, and extremely unsatisfactory messages or excellent ones are minority cases.

Then, after the expert evaluation, the manually assigned labels were compared with the result obtained with the automatic grading model. For this purpose, the numerical grades obtained were transformed into four different ranges. In this way, the results of the manual and automatic evaluation could be compared more efficiently. Figure 3 shows the distribution of the messages in the different ranges according to the automatic model.

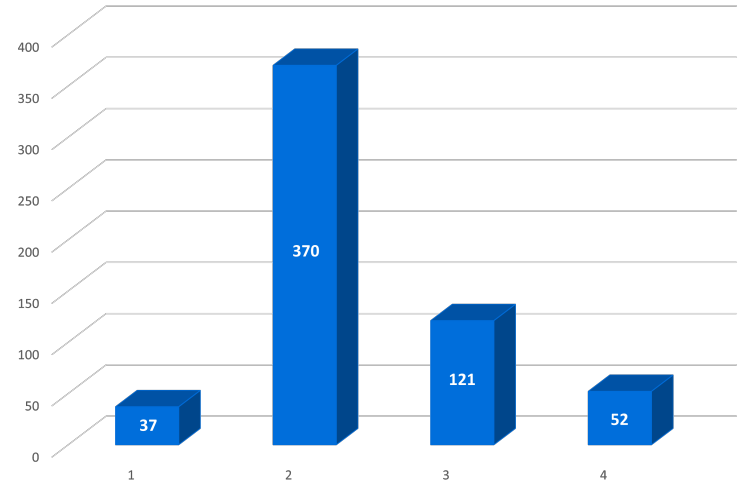

Figure 3. Quality levels distribution by automatic model.

As can be appreciated in the Figure 3, the distribution of labels in this case is very similar to the distribution obtained with manual labeling. The most noticeable differences are the decrease in the number of messages classified with label 1 and the increase in the number of messages classified with label 2. However, as discussed later, in this case this variation may be due to the evaluation of messages containing code fragments, since the model is not able to evaluate them completely well.

Finally, to definitively evaluate the performance of the model, different metrics were studied comparing the result obtained with manual labeling with experts and that obtained with the application of the model. First, the confusion matrix of the data was obtained.

Observing the Figure 4, it can be seen that the values of the main diagonal are quite high, so that the model (predicted label) correctly classifies most of the messages with respect to the experts (true label). The degree of accuracy for labels 2 and 3 stand out, since they are the ones with the highest values. This could be an expected result, as these are the most popular categories both in manual labeling and in the application of the model. Furthermore, the deviation in the classification of messages can be considered low. In each of the four categories, the deviation values are significantly lower than the hit values, and in most cases the deviation is centered on contiguous ranks (one label above or one label below). Then, other metrics were analyzed in addition to the confusion matrix of the model, and their results are shown in the Table 6.

Table 6. Evaluation metrics.

Metric	Value
Accuracy	0.8068
Precision	0.8228
Recall	0.6918
F1-score	0.7394
NRMSE	0.1799

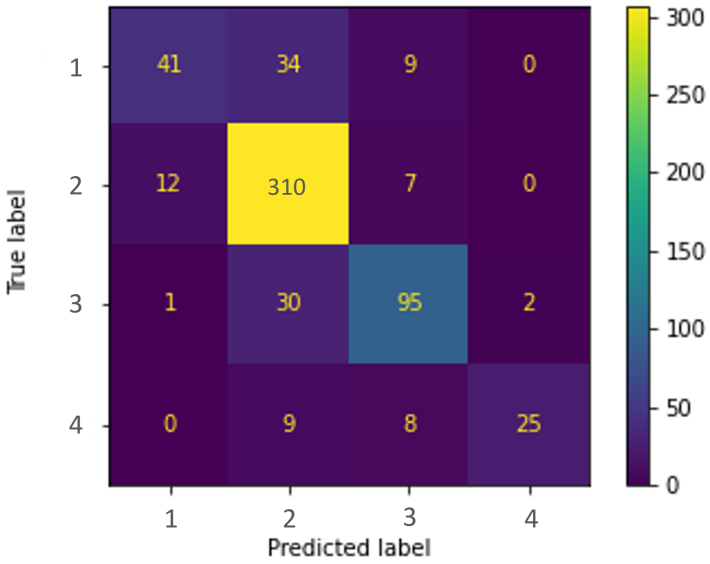

Figure 4. Confusion matrix.

When comparing the results of manual labeling and automatic grading, an accuracy of 80.68% is obtained. However, it should be noted that this result includes some messages that contain only code. Due to the nature of the course, some of the published messages contain pieces of code or answers to some exercise. These types of messages are not recommended, so the experts manually labeled these messages with the lowest rank. However, these messages are usually long and tend to have a lot of impact and interactions (many replies and numerous votes), so the designed automatic grading model usually grades this type of messages with high grades. Therefore, the accuracy obtained in this first evaluation is somewhat lower than what could be obtained if this type of code messages were filtered, but it is still a good result.

Regarding the rest of the metrics, the values obtained are equally good. For example, a precision of more than 82% has been obtained. This indicates that out of every 100 messages classified in a particular rank by the model, 82 belong to that rank when compared to the experts' classification. Therefore, the precision of the model is not a perfect value but a more than acceptable one. As for recall, this is the lowest value of the first four metrics studied. In this case a result of 69.18% has been obtained, so the sensitivity of the model can be somewhat improved, although it cannot be considered a bad result. Then, the F1-score, which is the combination of precision and recall, was also calculated. In this case a value of 73.94% was obtained, which is still a good value. Finally, when calculating the Normalized Root Mean Square Error (NRMSE) a value of 0.1799 was obtained. This value is appropriately low and close to 0.0, thus demonstrating that the model fits well the data used. With all these values, it can be concluded that the designed model is more precise than accurate, and that its overall performance is quite good.

6. Discussion

6.1. Implications

It has been possible to effectively evaluate user interactions in the forum with the automatic grading model presented in the previous section. This work has been influenced by other research conducted in open online Questions and Answers (Q&A) forums such as StackOverflow or Quora [17,19,31]. In addition, the articles by García-Molina et al. [32] and Coetzee et al. [13] also served as a starting point for the development of the algorithm. Hence, the presented model takes elements from the mentioned papers but adds an original and innovative value. In contrast to all the algorithms and evaluation methods in MOOC forums that can be found in the literature, the designed model considers three dimensions.

This characteristic makes it a robust and effective model that takes into account all aspects of interactions in discussion forums.

In addition, to answer the research question (RQ) posed at the beginning of the paper, an exhaustive evaluation of the model has been carried out analyzing the most representative and appropriate metrics for this specific case. As described in Section 5, a ground truth consisting of manual labeling by three experts was used to solve this question. The numerical grades resulting from the application of the model were transformed to the four possible quality ranges and compared with the ground truth to check the effectiveness of the model. In this way, it has been proven that a high accuracy can be obtained with the designed algorithm. Specifically, in this case an accuracy of more than 80% has been obtained despite the limitations of the code messages. Furthermore, other metrics were also studied since the classes were unbalanced, and equally good results were obtained. The precision, F1-score, and NRMSE values were particularly favorable, and the recall value remained a correct value even though it was the least successful result. It should also be noted that the evaluation was carried out by classifying the messages into four ranges, instead of the more usual two ranges. This increase in the number of categories generally implies a decrease in the precision and accuracy of the models, since the difference between the categories is more subtle and it is easier to deviate from the real value. However, as could be observed in the results obtained and in the confusion matrix presented, the model designed is fairly reliable in its classification and its deviations are usually centered on the categories adjacent to the ground truth given by the experts. Therefore, the obtained results are promising and show that the model is ready to be used in upcoming real courses.

Moreover, these results also improve the performance of some of the literature articles reviewed [7,23]. In other cases, the related studies do not provide data of the evaluation of their models but simply make the proposal of their methods and algorithms. Hence, the value proposition of this article is not only the model designed, but the evaluation performed and the positive results obtained.

6.2. Limitations

Although the result obtained in the model evaluation indicates that the algorithm is able to grade users efficiently, this study has had some limitations. First, as mentioned before, for this article, only data from a single trilogy of MOOCs was available. Although data from several editions of the "Introduction to Java Programming" courses were used, at the end of the day they are courses with the same dynamic and a very specific subject matter. Therefore, the data used for testing and evaluation are somewhat influenced by the subject matter of the particular course. This has been especially noticeable in the case of the impact measures. The algorithm designed is based on quality and impact measures, but in the different versions of this course, it has been observed that interactions of the type "vote", "follows", or "mentions", for example, have been very scarce. This has influenced the assignment of the weights of the different metrics, among other things.

Having data from several editions of the same MOOC has also affected the evaluation of the proposed configuration file. The use of this configuration file undoubtedly provides scalability to the model presented. However, it was not possible to test the effectiveness of the grading model using the configuration file on other courses, although the extrapolation process would be straightforward. The process for using the designed configuration file and for applying the model to evaluate the interactions of new MOOCs should begin with a brief analysis of the characteristics of the course to be evaluated. The moderator should first study the general behavior of the users in the course and assess which factors he/she wants to give more weight to and which actions he/she wants to encourage or reward. Then, the moderator should proceed to modify the configuration file to adjust the values of the configurable variables after identifying the most significant and influential ones. In addition, it is also possible to leave some (or all) of the variables unchanged to use their default value. After adjusting the variables, the model could then be applied to the course data to obtain the grades of the user interactions. Finally, although it has been shown

that the model is capable of correctly evaluating interactions, moderators are advised to have a quick review of the grades obtained in order to detect extremely bad or unusually good results and to analyze them more closely. In this summarized form, the performance described in Section 4.3. could be extrapolated to other courses.

Moreover, the course on which the design and evaluation process has been focused is a course to have a first contact with programming. For this reason, many of the messages published in the course forum include code snippets, either to ask a question or to give the solution to an exercise. As expected, the model is somewhat more imprecise with this type of messages. For instance, in the case of content quality metrics, readability and sentiment analysis usually comes out quite low, while the length of these types of messages is usually large. In terms of impact evaluation, these messages tend to generate more discussion than the rest and tend to have a greater impact on the forum. For all these reasons, the grade that the model assigns to this type of messages usually differs from what a teacher would do manually. Therefore, this also influences the final accuracy of the model obtained.

There is also another limitation related to the performance of the particular course used for the evaluation of the model. As already mentioned in Section 4.1, the delivery mode of the course has been changing in its different editions. The first edition of the MOOC was synchronous, so all users accessed the content and assessments at the same time. This favored participation in the forum and the answerability of messages. However, the delivery mode of the following editions was self-pace. This asynchronous model can be detrimental to user interaction and can result in many messages remaining unanswered.

Finally, it should be kept in mind that both the English and Spanish versions may include foreign students with different mother tongues than the one of the course. Hence, as is to be expected, some of the messages written by these users may contain typos or may have a slightly peculiar structure. This undoubtedly affects above all the grades obtained in the evaluation of the quality of the content of the messages, since, in this case, aspects such as the readability of the message are taken into account. All these details have affected the development of the automatic grading model proposed and have marked part of the work carried out.

7. Conclusions and Future Work

In this study, a new automatic grading model that seeks to evaluate user interactions in MOOC discussion forums has been designed. Unlike other existing algorithms, the model presented in this paper evaluates the user based on the quality and impact of the messages he/she has posted and his/her activity within the forum, which is a more comprehensive and sustainable approach. Once the assessment metrics were chosen, the evaluation of the model was performed on a particular MOOC, assigning a weight or maximum grade to each of the metrics. Manual labeling of a random subset of messages by experts was then used to test the effectiveness of the model and to obtain its accuracy. Finally, the results obtained show that the model presented successfully evaluates messages and users.

Although this work has been applied for just some types of MOOCs, one of the aims of the work was to create an effective and flexible evaluation model that could be applied to any type of MOOC. For this purpose, a configuration file was also developed that offers the possibility to customize some aspects of the algorithm. In this way, the instructor of any course can adapt the model to his or her needs and still have it work, regardless of the content or operation of the particular course. In the future, from the training of data, we could have some insights about the best combination of weights of indicators and probabilistic models depending on the types of MOOCs.

Furthermore, although the main objectives of the study have been achieved, some issues would need to be addressed in the future. First, as mentioned in other sections of the paper, the accuracy and performance of the model need to be tested in other types of courses. The evaluation of the algorithm presented has been performed on data from different editions of the same course, so it is considered convenient to study the effectiveness of the configuration file, and of the model in general, on other types of data.

Another interesting possible future research direction is to train different probabilistic models (e.g., neural networks, random forest, decision trees, linear regression, etc.) with some of the data of the MOOC in order to set up the best weights for the different indicators of the model. Next, the obtained model could be evaluated. This way, the weights are obtained from previous data instead of predefined by default.

As for the configuration file presented, it would be convenient to improve its format. At the moment, this first version of the configuration file is a Comma Separated Values (CSV) file that includes the different fields and values. However, the possibility of making the visualization of this file more aesthetic, professional, and intuitive to the instructor who may use it could be studied.

Finally, as it has already been mentioned on numerous occasions, one of the main obstacles encountered during the evaluation process has been the code messages. A possible solution could be the improvement of message type detection. The classification of messages into different categories according to their content could be added as part of the evaluation. For programming courses such as those studied in this work, within these categories there could be one to encompass messages containing code. In this way they could be detected and treated in more appropriate ways.

As can be seen, many of these future lines of work are closely related to the limitations encountered during the development of the project. However, despite these possible future improvements, a complete and functional work has been achieved, since the designed model meets the specifications and the expected results for a first development version.

Author Contributions: Conceptualization, R.L.P.-N., C.A.-H. and I.E.-A.; methodology, R.L.P.-N., C.A.-H., I.E.-A., P.M.M.-M. and P.J.M.-M.; software, R.L.P.-N.; validation, C.A.-H., I.E.-A., P.M.M.-M. and P.J.M.-M.; formal analysis, R.L.P.-N., C.A.-H. and I.E.-A.; investigation, R.L.P.-N., C.A.-H., I.E.-A., P.M.M.-M. and P.J.M.-M.; resources, C.A.-H., I.E.-A. and C.D.K.; data curation, R.L.P.-N.; writing—original draft preparation, R.L.P.-N.; writing—review and editing, C.A.-H., I.E.-A., P.M.M.-M. and P.J.M.-M. and C.D.K.; supervision, C.A.-H., I.E.-A.; project administration, C.D.K.; funding acquisition, C.A.-H., P.J.M.-M. and C.D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the FEDER/Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación, through the Smartlet and H2O Learn Projects under Grants TIN2017-85179-C3-1-R and PID2020-112584RB-C31, and in part by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307 and under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M21), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation), a project which is co-funded by the European Structural Funds (FSE and FEDER). Partial support has also been received from the European Commission through Erasmus+ Capacity Building in the Field of Higher Education projects, more specifically through projects InnovaT (598758-EPP-1-2018-1-AT-EPPKA2-CBHE-JP), and PROF-XXI (609767-EPP-1-2019-1-ES-EPPKA2-CBHE-JP). This publication reflects the views only of the authors and funders cannot be held responsible for any use which may be made of the information contained therein.

Institutional Review Board Statement: The study was approved by the Data Protection Office of Universidad Carlos III de Madrid.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used in this study are available at https://doi.org/10.5281/zenodo.5115573 (accessed on 16 August 2021). This dataset includes all the pseudonymised events of the discussion forum of the different editions of the MOOC "Introduction to Java Programming" offered in edX by the Telematics Engineering Department of the Universidad Carlos III de Madrid.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Boyatt, R.; Joy, M.; Rocks, C.; Sinclair, J. What (Use) is a MOOC? In *The 2nd International Workshop on Learning Technology for Education in Cloud*; Uden, L., Tao, Y.H., Yang, H.C., Ting I.H., Eds.; Springer Proceedings in Complexity; Springer: Dordrecht, The Netherlands, 2014; pp. 133–145. [CrossRef]
2. Christensen, G.; Steinmetz, A.; Alcorn, B.; Bennett, A.; Woods, D.; Emanuel, E. The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why? 2013. Available online: https://ssrn.com/abstract=2350964 (accessed on 14 July 2021). [CrossRef]
3. Daniel, J. Making Sense of MOOCs: Musings in a Maze of Myth, Paradox and Possibility. *J. Interact. Media Educ.* **2012**, *2012*. [CrossRef]
4. Grainger, B. *Massive Open Online Course (MOOC) Report 2013*; University of London: London, UK, 2013.
5. Moreno-Marcos, P. M.; Alario-Hoyos, C.; Muñoz-Merino, P. J.; Estevez-Ayres, I.; Delgado-Kloos, C. A learning Analytics Methodology for Understanding Social Interactions in MOOCs. *IEEE Trans. Learn. Technol.* **2018**, *12*, 442–455. [CrossRef]
6. Buder, J.; Schwind, C.; Rudat, A.; Bodemer, D. Selective Reading of Large Online Forum Discussions: The Impact of Rating Visualizations on Navigation and Learning. *Comput. Hum. Behav.* **2015**, *44*, 191–201. [CrossRef]
7. Ntourmas, A.; Daskalaki, S.; Dimitriadis, Y.; Avouris, N. Classifying MOOC Forum Posts Using Corpora Semantic Similarities: A Study on Transferability Across Different Courses. *Neural Comput. Appl.* **2021**, 1–15. [CrossRef]
8. Alario-Hoyos, C.; Pérez-Sanagustín, M.; Delgado-Kloos, C.; Munoz-Organero, M. Delving into Participants' Profiles and Use of Social Tools in MOOCs. *IEEE Trans. Learn. Technol.* **2014**, *7*, 260–266. [CrossRef]
9. Arguello, J.; Shaffer, K. Predicting Speech Acts in MOOC Forum Posts. In Proceedings of the International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015.
10. Gillani, N.; Eynon, R. Communication Patterns in Massively Open Online Courses. *Internet High. Educ.* **2014**, *23*, 18–26. [CrossRef]
11. He, C.; Ma, P.; Zhou, L.; Wu, J. Is Participating in MOOC Forums Important for Students? A Data-Driven Study from the Perspective of the Supernetwork. *J. Data Inf. Sci.* **2018**, *3*, 62–77. [CrossRef]
12. Kizilcec, R.F.; Piech, C.; Schneider, E. Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In Proceedings of the 3rd International Conference on Learning Analytics and Knowledge, Leuven, Belgium, 8–12 April 2013; pp. 170–179. [CrossRef]
13. Coetzee, D.; Fox, A.; Hearst, M. A.; Hartmann, B. Should your MOOC Forum Use a Reputation System? In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, Baltimore, MD, USA, 15–19 February 2014; pp. 1176–1187. [CrossRef]
14. Wong, J.S.; Pursel, B.; Divinsky, A.; Jansen, B.J. An Analysis of MOOC Discussion Forum Interactions from the Most Active Users. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Washington, DC, USA, 31 March–3 April 2015; pp. 452–457._58. [CrossRef]
15. Alonso-Mencía, M.E.; Alario-Hoyos, C.; Estévez-Ayres, I.; Kloos, C.D. Analysing Self-Regulated Learning Strategies of MOOC Learners Through Self-Reported Data. *Australas. J. Educ. Technol.* **2021**, 56–70. [CrossRef]
16. Wong, J.S.; Pursel, B.; Divinsky, A.; Jansen, B.J. Analyzing MOOC Discussion Forum Messages to Identify Cognitive Learning Information Exchanges. *Proc. Assoc. Inf. Sci. Technol.* **2015**, *52*, 1–10. [CrossRef]
17. Patil, S.; Lee, K. Detecting Experts on Quora: By their Activity, Quality of Answers, Linguistic Characteristics and Temporal Behaviors. *Soc. Netw. Anal. Min.* **2016**, *6*, 5, doi10.1007/s13278-015-0313-x. [CrossRef]
18. Ponzanelli, L.; Mocci, A.; Bacchelli, A.; Lanza, M.; Fullerton, D. Improving Low Quality Stack Overflow Post Detection. In Proceedings of the 2014 IEEE International Conference on Software Maintenance and Evolution, Victoria, BC, Canada, 29 September–3 October 2014; pp. 541–544. [CrossRef]
19. Movshovitz-Attias, D.; Movshovitz-Attias, Y.; Steenkiste, P.; Faloutsos, C. Analysis of the Reputation System and User Contributions on a Question Answering Website: Stackoverflow. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), Niagara, ON, Canada, 25–28 August 2013; pp. 886–893. [CrossRef]
20. Mathew, B.; Dutt, R.; Maity, S.K.; Goyal, P.; Mukherjee, A. Deep Dive into Anonymity: Large Scale Analysis of Quora Questions. In Proceedings of the International Conference on Social Informatics, Doha, Qatar, 18–21 November 2019; pp. 35–49._3. [CrossRef]
21. Maity, S.; Sahni, J.S.S.; Mukherjee, A. Analysis and Prediction of Question Topic Popularity in Community Q&A Sites: A Case Study of Quora. In Proceedings of the International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015; Volume 9.
22. Wang, G.; Gill, K.; Mohanlal, M.; Zheng, H.; Zhao, B.Y. Wisdom in the Social Crowd: An Analysis of Quora. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 1341–1352. [CrossRef]
23. Maity, S. K.; Kharb, A.; Mukherjee, A. Analyzing the Linguistic Structure of Question Texts to Characterize Answerability in Quora. *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 816–828. [CrossRef]
24. Mc Laughlin, G. SMOG Grading-a New Readability Formula. *J. Read.* **1969**, *12*, 639–646.
25. Roy, P. K. Multilayer Convolutional Neural Network to Filter Low Quality Content from Quora. *Neural Process. Lett.* **2020**, *52*, 805–821. [CrossRef]
26. Ponzanelli, L.; Mocci, A.; Bacchelli, A.; Lanza, M. Understanding and Classifying the Quality of Technical Forum Questions. In Proceedings of the 14th International Conference on Quality Software, Allen, TX, USA, 2–3 October 2014; pp. 343–352. [CrossRef]

27. Duijn, M.; Kucera, A.; Bacchelli, A. Quality Questions Need Quality Code: Classifying Code Fragments on Stack Overflow. In Proceedings of the IEEE/ACM 12th Working Conference on Mining Software Repositories, Florence, Italy, 16–17 May 2015; pp. 410–413. [CrossRef]
28. Arora, P.; Ganguly, D.; Jones, G. J. The Good, the Bad and Their Kins: Identifying Questions with Negative Scores in Stackoverflow. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 25–28 August 2015; pp. 1232–1239. [CrossRef]
29. Roy, P. K.; Ahmad, Z.; Singh, J. P.; Alryalat, M. A. A.; Rana, N. P.; Dwivedi, Y. K. Finding and Ranking High-Quality Answers in Community Question Answering Sites. *Glob. J. Flex. Syst. Manag.* **2018**, *19*, 53–68. [CrossRef]
30. Neshati, M. On Early Detection of High Voted Q&A on Stack Overflow. *Inf. Process. Manag.* **2017**, *53*, 780–798. [CrossRef]
31. Calefato, F.; Lanubile, F.; Marasciulo, M.C.; Novielli, N. Mining Successful Answers in Stack Overflow. In Proceedings of the IEEE/ACM 12th Working Conference on Mining Software Repositories, Florence, Italy, 16–17 May 2015; pp. 430–433. [CrossRef]
32. García-Molina, S.; Alario-Hoyos, C.; Moreno-Marcos, P.M.; Muñoz-Merino, P.J.; Estévez-Ayres, I.; Delgado Kloos, C. An Algorithm and a Tool for the Automatic Grading of MOOC Learners from Their Contributions in the Discussion Forum. *Appl. Sci.* **2021**, *11*, 95. [CrossRef]
33. Stump, G.S.; DeBoer, J.; Whittinghill, J.; Breslow, L. Development of a Framework to Classify MOOC Discussion Forum Posts: Methodology and Challenges. In Proceedings of the NIPS Workshop on Data Driven Education, Lake Tahoe, NV, USA, 9–10 December 2013; pp. 1–20.
34. Cui, Y.; Wise, A.F. Identifying Content-Related Threads in MOOC Discussion Forums. In Proceedings of the 2nd ACM Conference on Learning @ Scale, Vancouver, BC, Canada, 14–15 March 2015; pp. 299–303. [CrossRef]
35. Imran, A.S.; Dalipi, F.; Kastrati, Z. Predicting Student Dropout in a MOOC: An Evaluation of a Deep Neural Network Model. In Proceedings of the 5th International Conference on Computing and Artificial Intelligence, Bali, Indonesia, 19–22 April 2019; pp. 190–195. [CrossRef]
36. Ramesh, A.; Goldwasser, D.; Huang, B.; Daumé, H., III; Getoor, L. Understanding MOOC Discussion Forums Using Seeded LDA. In Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications, Baltimore, MD, USA, 26 June 2014; pp. 28–33.
37. Wen, M.; Yang, D.; Rose, C. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In Proceedings of the 7th International Conference on Educational Data Mining, London, UK, 4–7 July 2014; pp. 130–137.
38. Wen, M.; Yang, D.; Rosé, C. Linguistic Reflections of Student Engagement in Massive Open Online Courses. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8.
39. PyPI textblob 0.15.3. Available online: https://pypi.org/project/textblob/ (accessed on 14 July 2021).
40. PyPI scipy 1.7.0. Available online: https://pypi.org/project/scipy/ (accessed on 14 July 2021).
41. Flesch, R. A New Readability Yardstick. *J. Appl. Psychol.* **1948**, *32*, 221. [CrossRef] [PubMed]
42. PyPI textstat 0.7.1. Available online: https://pypi.org/project/textstat/ (accessed on 14 July 2021).