University of Twente

Dropout prediction in MOOCs

Using sentiment analysis of users' comments to predict engagement

Nataliia Dmoshinskaia

Supervisors:
Dr. Bas J. Kollöffel
Prof. Dr. Bernard P. Veldkamp

University of Twente

2016

Acknowledgments

I would like to express my deep gratitude to my supervisors Dr. Bas J. Kollöffel for his helping me to see the bigger picture of the research area and constant support, and Prof. Dr. Bernard P. Veldkamp for his expert look at my data and interest in the results. Their valuable advice, constructive feedback and positive attitude did contribute a lot to the process of my thesis writing and the final result.

I would like to thank the Programme Coordinator Jan Nelissen and The Scholarship Committee of the University of Twente for believing in me and awarding me with the opportunity to study in this programme. Not only has this year enriched my knowledge of the Educational Science and improved my research skills as I learned from the leading experts in various fields but it has also made me a more broad-minded educationalist.

I would also like to thank my family and friends both in Russia and in the Netherlands for their constant support and high hopes for me which helped me to go through the ups-and-downs of the learning process. And my special thank is to Ilya Musabirov for his priceless advice on programming and patience to deal with my emotional reaction to this activity.

Table of Contents

Summary

Massive open online courses (MOOCs) have recently become very popular and many universities and independent companies provide them. However, along with popularity these types of courses are characterized with a quite high dropout rate. Finding out if there are any predictors of dropping or staying behaviour can be of practical usage for teachers and course developers. One of the new directions of studying learners' behaviour is using their comments and posts as a source of information. Comments are even more informative and interesting to analyse in those MOOCs which stimulate participants to take an active part in forums by giving extra points for that. Thus this research aims at investigating the link between sentiment tonality of learners' comments and dropout from forums in MOOCs with digital badges. The current study contribution can be described as following. First, the problem of drop out in MOOCs is not solved and any research adds to the knowledge of this area, thus helping to find the solution. Second, using not only quantitative data such as a number of competed quizzes and watched videos but also qualitative data of learners' comments facilitates better understanding of a phenomenon of online learning. Finally, employing a very new method of sentiment analysis in the educational context enables getting a new type of data and broadens the applicability of the method.

Two MOOCs were analysed in the study with the total amount of participants being 321: 169 for one course and 152 for the other and total number of commented objects (videos, articles, exercises etc.) accounting for 88: 53 for the first course and 35 for the other. The first MOOC offered two tracks – technological and strategical so the comparison was made between the tracks inside the course as well as between the courses. Both MOOCs took place in 2015 and the data were collected automatically during the course.

Participants' comments were assessed with the help of R packages so the sentiment score of the comments were obtained. At the course level the following variables were used: a sentiment score per each object (part) of the course, a share of negative comments (the score below zero) and a number of comments per object (dropout rate). No particular pattern in changes of sentiment score and a share of negative comments during the course was identified. Moreover, relationships between lower sentiment score and dropout, which were discovered in other studies, was not proved by the present data sets. These trends are true for the whole population as well as for separate groups of students – those who completed more than 70 per cent of topics, those who completed less than 30 per cent and the middle group.

At the student level the following variables were used for the analysis: an average score, a minimal score, a maximal score, a share of negative comments and a share of positive comments. No significant link was identified neither between average score and completion rate, nor between minimal score and completion rate. These results were received for both courses and for both tracks within xAPI course.

To identify the most influential variables more complex models were used such as conditional inference trees which split the population according the value of a parameter with the most statistical significance. Some correlation between the variables and the completion rate was found for different groups. For all courses and tracks one of the lowest completion rate was registered for the group of participants with the highest minimal score. In other words, those who do not express any negative emotions are likely to drop the course very soon. On the other hand, those who express negative emotions a lot are also in the group of risk in terms of completion. The highest completion rate is associated with the group of learners who have some negative comments but not too many. It means that some critical cognitive involvement in the course is vital for succeeding. These findings may influence the focus of MOOC teachers and course developers. Critical comments can be more appreciated as they signal more involvement and, thus, potentially more completion.

Key words: MOOC, dropout, sentiment analysis

Chapter 1. Introduction

Massive open online courses (MOOCs) had their hype in 2012 and they still are a very new and discussed topic in a much broader audience than just the educational society. This is, first of all, explained by the increasing need for lifelong learning to adapt to a rapidly changing job and study requirements. Nowadays changes in any professional sphere are inevitable so any education or training obtained once is not enough anymore. There is a shift from a “knowledge” employee to a “learning” employee which means that companies are not any longer interested in hiring people who know everything in the target area but rather those who are ready to learn more in this area. This also means that workers should take pro-active position and be responsible for obtaining the right skills and knowledge on time. These trends make MOOCs an ideal solution as they provide open and free education from the best universities or industry experts, which makes them an ideal tool for professional self-development and bringing courses to various student audiences. The course materials are available at any time so people can choose not only when to study but also how, in which order and at which pace.

This opportunity to enormously enlarge learners’ number and reach geographically remote areas appealed to universities a lot. So MOOCs were thought to solve many educational problems by providing free access to the-state-of-the-art edge courses, thus reducing the cost of university level education and enhancing education opportunities for learners in developing countries. This encouraged discussions on the potential of such courses to become a stimulus to re-think and re-structure existing educational models (Reich, 2015). However, the reality turned out to be less idealistic.

Creation of a big amount of courses attracted a huge audience which was very diverse in every possible criterion. This might partly explain a huge dropout rate, however, more studies should be done in this area. The number of quitting students, which was and still is quite striking – up to 95%, has obviously attracted a lot of attention of educators (Perna et al, 2014). These numbers even caused some speculations about uselessness of MOOCs. Nevertheless, such courses are still very popular among various groups of learners and this trend is going up in developing countries and among younger people (Macleod, Haywood, Woodgate & Alkhatnai, 2015). This means that MOOCs are already an important and integral part of a learning process and efforts should be put into improving them rather than discuss their drawbacks. There has not been found any solution for the dropout problem yet but there is a strong belief that the key is in more detailed studying of this phenomenon.

MOOCs differ by their structure and teaching principles behind them, which may influence the dropout rate as well. This means that separate studies may be required for different types. Regardless the type all MOOCs have forums for users to comment the course, ask questions and learn from each other. These forums can be a valuable source of information about participants’ attitude towards the course and, ultimately, about their likelihood to drop out. Taking part in forums is always voluntarily for the users, however, there are courses which encourage this activity by giving digital badges for that. Courses with digital badges are designed the way that requires learners to demonstrate completing some skills or levels to proceed further. This may include earning points for participating in a discussion, which means that learners are more likely to be actively involved in the forum activity and produce a more realistic picture by their comments.

To deal with a big amount of online data collected during the course learning analytics is used. One of its tool is sentiment analysis. This technique has been successfully employed to assess users’ emotions and opinions about different goods and services from household appliances to entertainment. This method enables to calculate the sentiment level of the comment which then be added to get the general sentiment level for the topic (Liu, 2011). Comparing this level for all topics of the course can create a sentiment map of the course.

Applying sentiment analysis in the educational online context is a new step in MOOCs’ development and can provide insight into better understanding of learners’ attitude towards the course, learning satisfaction and possible sentiment predictors of the dropout. Knowing them teachers can plan their interference timely to keep more students on track. Moreover, comparing sentiment levels of different courses can reveal common patterns in emotions and attitudes of learners, which can be of practical importance for instructional designers. All together this can give valuable ideas for the direction of course development.

The present research aims at contributing to studying of dropout and engagement in MOOCs analyzing a particular example of courses with digital badges and with the help of sentiment analysis of learners' comments. The study investigates the link between sentiment tonality of learners' comments and dropout from forums. Two MOOCs are studied with one of them providing two tracks. Thus the comparisons within the course and between the courses are made.

Chapter 2 provides a reader with a theoretical basis of the research. Key issues such as MOOCs, their characteristics and dropout rate are dealt with in more details. The chapter finishes with stating of the aim and the research questions.

Chapter 3 presents the methods used for the research together with the description of the procedure, participants and data sets. More elaborate explanation of sentiment analysis technique is also given.

Chapter 4 introduces the results, most of which are visualized to make it easier to see the trends, and basis for making inferences.

Chapter 5 includes conclusion and discussion parts and sums up the paper.

Chapter 2. Theoretical framework

MOOCs' characteristics

Massive open online courses have become an indispensable part of the educational environment. They provide an opportunity for people from different geographical and time zones attend the same course online at their own pace and in convenient time. Such courses can be provided by universities, other educational institutions or specific providers – big platforms which unite many different developers e.g. Coursera or edX. Courses can vary in duration and structure, which depends on the goals and the topic. The average course consists of four to eight weeks with self-study materials for each week, a small quiz at the end of each week and a bigger assignment or test at the end of the course.

Educational designers distinguish two types of MOOCs – xMOOCs (extended MOOCs) and cMOOCs (connectivist MOOCs). The former is provided by a university and usually is an extension of a regular course. Such courses might be open to students of different majors but they normally aim at a particular target group and may have some face-to-face elements, e.g. final exam. Because of being a part of a university curriculum they usually employ teaching methodology, course structure and kinds of tasks which also characterize traditional education. This type of online courses is becoming more and more popular among universities worldwide.

cMOOCs differ from the other type by the approach to teaching and learning that is advocated by their developers. They are based on the idea of connectivism that learning happens within a network, where learners use digital platforms such as blogs, wikis or social media platforms to make connections with content, learning communities and other learners to create and construct knowledge (Siemens, 2014). In such courses users are encouraged to contribute actively, using these digital platforms. This contribution is never obligatory but always welcomed and can be supported by the usage of digital badges or other types of motivation. This approach influences the structure of the course as well as some formal assessment parameters. cMOOCs are independent of any traditional courses and are taken completely online. This is the type of MOOCs that is going to be studied in this research so it is noteworthy to study the general characteristics of the audience as well.

This division is also reflected in the audience. xMOOCs usually attract more students either because this course is a part of their programme or because it is delivered by a good university. As these courses are more formal in structure learners tend to pursue not only the knowledge but also a certificate of completion. cMOOCs are more likely to have older audience of professionals who care more about learning new things or skills and less about formal prove of the completion. Usually, however, the researches study MOOCs' audience at a general level. Its general characteristics cover age groups, gender, level of education, geographical location, purpose. So the majority of learners are adults older than 25, with a university degree, employed and living in various locations worldwide (with a tendency to be in America or Europe). Gender is very content-dependent factor as well as purpose, however, the main is learning new things relevant but not closely related to the current work or study area. The major changes in the characteristics show more active involvement of younger people (even younger than 18 years old) from outside Europe and America (Christensen et al., 2013; Liyanagunawardena, Lundqvist, & Williams, 2015; Macleod, Haywood, Woodgate, & Alkhatnai, 2015).

Activities in MOOCs

Learning in MOOCs differ from any traditional form of education. Learners are much more independent and should exercise much more self-regulation to complete the course. Usually this involves watching short video lectures, reading suggested articles or blogs, doing practical exercises and weekly quizzes or tests. This part is quite autonomous and learners do it at their own pace. And indeed the research prove that autonomy, diversity, openness and are characteristics of MOOCs that make them so attractive for many people (Mackness, Mak & Williams, 2010). However, the same research shows that there is one more characteristic which is equally important and this is connectedness or interactivity. Participants tend to communicate and engage themselves in online groups to seek support and discuss the learning process and the course with their fellow-students.

Online learners are affected by their social learning environments, which can support and/or hinder both their learning and motivation (Boling, Hough, Krinsky, Saleem, & Stevens, 2012). This means that communication and collaboration become more important. Moreover, it is correct to say that online learning is driven by online participation, which also has definite positive effects on learning satisfaction and retention. So increasing online participation will consequently enhance online learning satisfaction and decrease dropout (Hrastinski, 2009).

In online context communication is only possible via forums and discussions and this way they become a tool for social learning in MOOCs which means users can learn from each other. This means that students' active participation in interactions becomes one the major online learning activities (Cercone, 2008). Such participation can serve as a good involvement indicator especially in MOOCs that encourage that with the help of digital badges. This means that completing different learning activities users 'earn' badges which allow them to proceed to the next week material. Participating in discussion (usually with a pre-set question from the instructors) is one of the activities to get badges. It is still possible to obtain the access to all the materials without participating in the discussion but most learners are more likely to get involved in it in comparison with the usual course. Thus, forums should be paid a special attention to as they are so important for succeeding in the course.

Another reason for focusing on forums is that they enable teachers, course-developers and providers to get some real time feedback from learners. This feedback is qualitative, on the spot and quite substantial in amount, which is both a benefit and a hurdle. Being able to analyse and use this information will definitely provide deeper understanding of the learning process and, consequently, of the ways to improve it. The current research concentrates on forums as sources of information for analysis.

Dropout in MOOCs

The notion of dropout is very disputable in the context of online courses. The traditional view on dropout as the number of students who have not completed the course may not be so appropriate here. In traditional courses dropout rate is the difference between the number of students who registered for the course and those who completed it. This principle was firstly applied to MOOCs as well. Calculated this way dropout rate is strikingly high being between 88 and 95 % (Perna et al., 2014). This huge numbers attracted a lot of attention to this issue and there were even doubts about efficiency of MOOCs in general. However, later research proved that this is not the right approach in online context.

The learning process and the motivation for studying in a MOOC differ dramatically from face-to-face courses. In this context, learners are free to decide what, when, where, and how they will learn. As materials are usually freely available they can browse, pick and choose, and follow their own agenda, which was not possible before even with online education. Volunteer nature of MOOCs makes it possible for people to join it for some particular skills or information and not complete the whole course (Liyanagunawardena, Parslow, & Williams, 2014). This is specifically true for cMOOCs – courses that are taken exclusively online and not as a part of a regular university course. This quite often means that learners have different purposes of joining the course from getting a very specific skill trained to spending time more intellectually than watching TV (Yang, Sinha, Adamson, & Rose, 2013). Such situation makes completing the course to be one of the goals but not the only one. However, Park and Choi (2009) showed that learners who are satisfied with the course and find it relevant to their needs are less likely to quit.

The ways to address dropout are still very different for different researchers. As Onah, Sinclair and Boyatt (2014) showed in their research many participants who may be classified as droppers as, for example, they do not complete the necessary tasks to get a certificate, are still participating in the course but in their own manner. This may include a slower pace or selective engagement in the learning activities. Such disagreement about the notion of dropout makes the investigation a bit more difficult but also stimulates more research as a shared vision can be reached only through more study and a discussion based on the findings.

Recent studies suggest that dropout rate should be calculated for ‘active users’ – those users who have performed some learning activities during the course (Onah, Sinclair, & Boyatt, 2014, Liyanagunawardena, Lundqvist, & Williams, 2015). These activities can vary in different studies. The most commonly used approach is viewing logging into the course website as a minimal activity. According to the Report of Edinburgh University team (MOOCs@Edingurgh, 2013) with one course as an example, just above the half (53%) of the enrolled users were actually engaged in any learning activities during the whole life span of the course. And a bit less than a third (29%) of those who launched during the first week was still active during the fifth. Even though this number shows not very big attrition it is still significantly bigger than 12%, which would be the completion rate if compared to the number of enrolled.

Another way to tackle the dropout issue is to evaluate the number of learners that submitted at least one quiz. The numbers here are quite comparable with the engagement metrics, for example, in the Duke MOOC, about 26% of enrolments attempted at least one quiz in the first week (Belanger & Thornton, 2013). Some researches go even further and suggest taking into account those students who successfully complete the assignments of first two weeks (Coffrin, Corrin, de Barba, & Kennedy, 2014). They found out that students’ performance during this period is a good predictor of the final grade as these students have required prior knowledge and invest enough time and effort. Applying this approach, the retention rate for two studied courses were 42.1% and 27.4% compared with the 5% to 3% respectively calculated in a traditional way. All these examples illustrate how important it is to set the right standard of reference for analyzing dropout and that this standard can no longer be a traditional number of registered users.

Calculating dropout rate referring to ‘active users’ allows to exclude people who registered but because of different reasons never actually attended the course. However, it is not that obvious what to consider a basic activity for an active user. As the enrollment procedure is extremely simple and a lot of courses are free there is a big amount of people who enroll just to get more information about the course. Ideally these users should not be included in the dropout rate as they never had an intention to do the course. This makes defining an ‘active user’ as the one who logs in at least once a less reliable measure. People looking for more or very specific information would probably log in just once pushing up the dropout rate but not having real reasons for quitting.

To use quiz submitting as the main metrics for identifying ‘active users’ may also have some disadvantages because it does depend on the course structure and number of quizzes. This is especially true for less formal courses whose certificates of completion may not be very useful for learners. In this situation they may just participate in some activities relevant for their needs and ignore others, including submitting quizzes.

Various researches have shown that students taking part in online discussion boards are more willing to stay through the course. In their study Xing, Chen, Stein and Marcinkowski (2016) found out that the participation in forums and dropout rate demonstrate opposite trends during the course – the weeks with the highest forum activity have the lowest drop out and vice versa. Another study demonstrated that students who began their active participation in the first week of the course were 35% less likely to drop out on each time point than the population average (Rosé et al., 2014). This makes using participation in discussion a justified criterion for detecting ‘active users’. So based on the previous research evidence the present study views dropout as a stop and lack of further participation in the forum discussions.

A growing trend in MOOC research advocates the approach that not only simple metrics should be taken into account while assessing learning in MOOCs but also more qualitative data such as users’ feelings and opinions (Reich, 2015). Going into this direction, studies discover various factors which can influence the dropout. Researches in this area, which are described below, reveal the connection between dropout and several variables such as learning satisfaction, self-regulation abilities and sentiment level.

For instance, a significant relationship was found between learning satisfaction and continuing learning intention with some particular aspects of learning satisfaction such as course content satisfaction and administrative services encounter being more influential than others (Wu, Hsieh & Lu, 2015). Analysing factors that discriminate retained and dropped learners Lee, Choi and Kim (2013) discovered that academic locus of control and metacognitive self-regulation for learning play the key role in the decision to quit the course. Thus, participants with internal locus of control and more developed self-regulation skills are more prone to complete the course. These researches usually use more traditional methods of collecting data such as surveys and questionnaires, which are distributed either on paper or online. Though being reliable and informative these methods do not employ the distinguish feature of online courses i.e. a substantial amount of collected information.

Learning analytics

The fact that a lot of data are collected automatically leads to the trend that using and analysing data about the online activity of MOOCs participants is becoming more and more popular. This enhances the development of learning analytics and educational data mining, applying traditionally business approaches to the educational context. Learning analytics techniques and educational data mining deal with the low level trace data regarding users' interactions with a course and with other users and draw conclusions on their basis. Moreover, they enable to predict higher level student's behavior such as dropout from this type of low-level structured data. The potential benefits of using learning analytics and educational data mining are reinforced by the automatic nature of these methods which allows them to be applied to the large scale MOOC context.

One of the examples of learning analytics is studying the relationship between the sentiment level and dropout. Though this is the newest approach several researches have been already done in this area. Using machine learning approach Chaplot, Rhim, and Kim, (2015) showed that sentiment analysis of forum posts is an important instrument to predict student attrition in MOOCs. Another research was conducted by Wen, Yang and Rose (2014), who used lexical analysis of words in the comments of learners of three MOOCs. The areas of the courses were specifically chosen to be very different from each other – programming, teaching and literature. According to this study higher sentiment ratios (more positive tone of the comments) are associated with fewer dropouts at the course level, i.e. the weeks with higher average sentiment level demonstrated smaller number of dropped out students. These trends were registered for all three courses. However, findings at the student level were slightly different. Intuitive assumptions that negative sentiment level predicts learner's dropping out were not completely proved. This correlation seemed to be more course dependent as for different MOOCs different trends in sentiments correlate with the dropout rate. This situation proves that this area requires further studying to generalize these findings for other fields.

Most published researches and both mentioned above ones study dropout from the course. However, the presented discussion demonstrated the necessity to address only active users and difficulties to identify the core activities. The approach taken by the current research studies the dropout from forums as this group of learners is the most active and thus is the most important to keep on track. Studying dropout patterns in this particular area can deepen the understanding of this phenomenon and this is the aim of the current study. Another specification is the characteristic of MOOCs being studied – they are courses with digital badges and no earlier researches deal with this particular type.

This means that in the present work the relationship between the sentiment level and dropout for 'active users' will be studied with forum participation being a criterion for dividing users for active and not. This is done with the help of sentiment analysis. Though technical details of the method will be given in Chapter 3, the general description is worth mentioning here. The method is so new that public awareness of it even in academic circles is not very high, however, understanding how it works contributes to formulating the research questions.

Sentiment analysis

Sentiment analysis is still a hot area in computer sciences, which means that it only makes its way to be recognised and applied in other scientific domains. Sentiment analysis means extracting information on opinions and attitudes from the various types of texts by analyzing the general tone of the extract and considering it positive, negative or neutral (Liu, 2011). Neutral is sometimes left out as less informative, however, this depends on the algorithm. Sentiment analysis uses linguistic and textual assessment, such as Natural Language Processing, to analyze word use, word order, and word combinations and, thus, to classify sentiments into the mentioned categories. This can be done with the help of the lexical method or machine learning (Bhadane, Dalal, & Doshi, 2015).

The main area of implementation of sentiment analysis is mining people's blogs, twits and posts to get information about general attitude towards various subjects from a particular product to a political party or candidate. Data obtained through sentiment analysis are believed to provide information from the source that was previously unavailable directly – public opinion and feeling. Such information could show success or failure of any particular company, service, product, party or policy as mining opinions helps companies to get their clients' feedback very quickly and act accordingly. The same principle can be applied to the educational context. Assessing and taking into account learners' opinions and attitudes towards the course can be a very useful feedback for course developers, moreover, the feedback received just on the spot which increases its validity compared with the surveys conducted before or after the course. However, as the technique of sentiment analysis is still being developed, the reliability standards reflect this process of mastering. Up to date the relevance of 70 per cent between a computerized and human assessment is considered to be good (Kennedy, 2012).

Though becoming popular this approach is not unconditionally approved. One of the critiques of sentiment analysis is 'monitisation' of feelings as companies try to use people's opinions and feelings expressed in social networks to assess their products or services, to adjust their policy and, eventually, to get more profit. They are supposed to make decisions responsibly and be influenced by some ethical norms, which is often referred to as 'moral economy' (Kennedy, 2012). Moreover, there are some concerns about privacy of the comments, tweets and posts used for the analysis. The up-to-date approach to this problem is to use only public information and not try to obtain the information that is marked as private or protected by firewalls, even if it might be technically possible. However, the difficulty to separate public and private information on the Internet and especially in social networks still leaves some room for concern. On the bright side there is the fact that sentiment analysis is usually used to deal with big data sets. It is not about an opinion of an individual but rather about accumulative perception and feelings of a big group, which gets a particular person out of focus of the analysis. This is especially true for the educational environment. Though the standards for such researches are still being developed the perspectives of using this technique in various areas are quite attractive.

The idea of applying sentiment analysis for MOOCs is still very fresh, however, the amount of digital information available in the form of comments makes it very appealing. Several studies have been already conducted in this field attempting to find correlation between sentiment level and different other variables. For example, Tucker and Divinsky (2014) investigated the link between the sentiment level and student performance. They discovered a slight positive correlation between student sentiments and quiz performance and a stronger negative correlation between the sentiment and homework assignments. These findings are just an example that demonstrates that intuitive assumptions about sentiment level may not always be true and require further and thorough investigation.

We are unaware of any specific researches on the relationship between sentiment level and dropout rate specifically for courses with digital badges, i.e. for those which stimulate learners to participate in discussions. Therefore, further exploring of these trends in such MOOCs and possible sentiment predictors of quitting the course can contribute to the course development. Furthermore, identifying common sentiment patterns in different MOOCs can be of practical importance for instructional designers. This leads to formulation of the following aim.

The aim of this research is to investigate the link between sentiment tonality of learners' comments and dropout from forums in MOOCs with digital badges.

Research questions and a hypothesis

Based on the theoretical analysis of existing researches in this field and practical needs the following research questions were formulated.

1. Is there a link between sentiment characteristics of comments and quitting the forum?
2. Is there a common pattern in sentiment dynamics in different MOOCs?

Regarding these research questions and taking into account the findings of previous studies the following expectations were **hypothesized**:

Sentiment level of users' comments has a relationship with the dropout and engagement and can be used to predict them.

To investigate this phenomenon the link between such variables as sentiment level per each topic, share of negative comments and dropout rate was assessed at the course level. And variables such as average score, maximal or minimal score and dropout were studied at the learner's level. Some previous researches allow to assume linear relationships between the mentioned above parameters, however, others suggest more complex nature of the link. Therefore, both linear regression analysis and more complex models were used in the current study.

Chapter 3. Methods

Participants

Participants of the study are learners of two MOOCs: 'Introduction to xAPI' and 'How to create an outstanding MOOC', which originally took place in April-June and September-October 2015 respectively. However, the materials were open for the public for several months after the course and some people took the courses later. The total number for both courses is 321. They are adults, approximate age range, basing on previous courses, is 25-45 years old, the gender composition is expected to be 50/50. However, there is no more exact information about the respondents as the data about them is absolutely anonymous. Data collection did not require participants to perform any extra or special task as the comments, which are analysed, were posted by participants voluntarily and as a part of the course activities.

This research can be considered an internet research as it uses the Internet to collect data and studies people's activity on the Internet so it applies principles introduced in Ethical Decision-Making and Internet Research (version 2.0) Recommendations from the AoIR Ethics Working Committee (Markham & Buchanan, 2012). The privacy of participants is protected by storing and analysing data with the users' ID and not real names. User's ID is a five-digit unique number given to every learner during the registration and kept for the whole course. There is no possibility for the researcher to obtain contact or personal details of the participants and no connection can be made from the research results to a real person.

Prior to working with data the approval of the Ethical Committee of the University of Twente was obtained. Basing on this in the research report the analysed and generalized data will be presented without any citation of original comments. This does not interfere with the research goals as the data were always viewed at the general level without any connection to a real person.

Research design

This is a quantitative correlational research with the aim of discovering the relationship between sentiment level and dropout rate. Qualitative data of users' comments are quantified with the help of sentiment analyses and this way the sentiment level of every object (part) of the course is created. Sentiment characteristics of each part are later compared with the dropout rate to find any relationships. The measurement of sentiment is done with the help of R packages. The current research is cross-sectional as the data are compared inside the course between two tracks and between two courses.

Data preparation and analysis

The present research employs the method of sentiment analysis to the users' comments posted in the course discussion. The analysis can be conducted with various methods and at various levels. Methods can be generally divided into three groups – manual, dictionary-based and corpus-based. The manual method involves coding lexicon for a specific area, which is definitely very time- and energy-consuming, so it makes this method rarely used. The second approach uses dictionaries, which give positive and negative connotation to words, as a source of reference. Finally, corpus-based methods employ a wide corpus of similar documents to assess the tonality of the piece being analysed. As for the levels, it can be the whole document, sentences or even phrases or parts of the sentences, so called aspect-analysis (Feldman, 2013). The choice depends on the type of the document and the purpose of the analysis. For example, a product review is more likely to have aspect-analysis as a user can have different opinions about different product characteristics, while posts in political forum would rather employ the sentence level as a more informative unit.

In the current research sentiment analysis is conducted at the post level, with most of the posts consisting of one or two sentences. In other words, in this research sentiment analysis is conducted at the sentence level with the dictionary-based approach. This way every comment is assessed separately which means it gets its own sentiment score.

Before any analysis, the text is processed to minimize extra information and noise. There are different approaches to this data ‘cleaning’ in lexical method, which is being used. In the current research the baseline approach, stemming and stop words are applied. Baseline means using the dictionary list of positively and negatively tagged words. Stemming reduces the words to their stems cutting down different ending and suffixes. This facilitates the analysis as the meaning is kept and all the words from the same stem are more easily recognized and labeled. Stop words are functional words such as articles, prepositions, particles, etc. which do not have an independent meaning and do not add any sentiment to the sentence. They are deleted from the sentence to simplify the analysis. After this processing a natural language sentence “I liked this exercise very much” is transformed into the chunk “like exercise very much”. Such chunk is first analysed and then given a sentiment score by adding the number of all words with positive meaning and subtracting the number of all words with negative meaning.

Raw data presents a *json* file with the following information: a comment itself, an ID of the object to which the comment was made, an ID of the participant and a date. As the courses were available for several months with the supposed duration of several weeks, learners took them at their own pace and in convenient time. This makes information about the date of commenting not relevant for identifying dropout. Instead, the ordinal number of the last commented object (topic) was used. No data of users' log in were available.

Qualitative data of learners’ comments were then quantified by a special method of sentiment analysis to assign a positive or negative tone to each comment. This was done with the help of R packages: each word in the sentence is compared with an extended pre-downloaded list of positive and negative words based on the dictionary meanings. Each positive word and negative word weighs +1 and -1 respectively. The sum of all words in the sentence gives the sentiment score of the sentence. This way the sentiment score of a sentence can be a negative figure, a positive figure or zero if the number of positive words in the sentence equals the number of negative.

Data cleaning and sentiment processing were done with a help of several pre-installed R packages. The first one is *dplyr* which provides a flexible grammar of data manipulation. The packages for data processing include *NLP* which is responsible for natural language processing, *quanteda* which allows quantitative analysis of textual data and *tm* which is charge of text mining. Finally, *ggplot* is used to visualize the input data frame in a graphic form.

The medians of sentiment score for each part were accumulated to create an ‘emotional map’ of the course. Such maps can indicate parts with lower and higher levels of sentiment. Two MOOCs were analysed and one of them has two tracks. So comparison was made between the tracks within one course and between courses as well to investigate if common patterns in sentiment dynamics exist.

Apart from the average sentiment level two more parameters were studied at the course level. They are a number of comments per topic – to demonstrate the dropout, a share of negative comments per topic – to find out the topics which cause the biggest number of negative comments. The relationships between average and minimal sentiment levels and a dropout rate were studied with the help of descriptive statistics. Linear regression was applied to investigate the correlation.

Other variables were taken at the individual level. They are an average sentiment level per person, minimal and maximal sentiment score per person and a share of negative and positive comments per person. The relationships between them and the completion rate of a particular individual were studied with the help of descriptive statistics. Linear regression was applied to see if any correlation exists.

To identify the most influential variables a Conditional Tree model was applied, in particular a method called Unbiased Recursive Partitioning (Conditional Trees) (Hothorn, Hornik & Zeileis, 2006). Unbiased is a part of the name and indicates some improvement to older methods to the ability to build tree-based models. As the name suggests, this approach operates by recursively extracting smaller and smaller segments of the data set, with the goal of finding a split sequence (a tree), which ensures an optimal prediction of the outcome variable. To avoid some unwanted statistical properties of the model, at each step the algorithm chooses among the possible splits the one which is associated with maximal statistical criterion value. This ensures statistical significance of each split and reduces the bias towards the variables with many distinct values. Thus, this method allows to identify the variables which influence the dependent variable the most – this is the completion rate in the present research – and the values of the independent variables where the split occurs. This is done with the help of the R package *partykit*.

Data sets

MOOCs which are analyzed in the current research are provided by the company HT2. The developers are strong advocates of social learning approach. Their goals presented on the website look like the following.

Connect

First, we connect learners with inspiring content to fire their imaginations. We use content from any source and organise it into bitesize playlists that learners can browse in any order.

Discuss

Next, we encourage learners to add back to the community using discussion questions. We gamify the process to encourage quality contributions. Learners can respond in text or using our in-browser video capture technology.

Curate

Finally, learners are encouraged to curate the best comments and conversations, voting, bookmarking and recommending to crowdsource the most popular ideas and opinions (www.curatr3.com/features/).

The mission shows how important participation in discussions is for the courses provided by this company. Both MOOCs in this research are courses with digital badges so learners earn extra points for participating in the discussion. Both courses consist of a number of parts which include articles, blogs, short video lectures and practical tasks. These parts will be referred to as objects in the analysis part. Every object has a separate forum on which the discussion is initiated by teachers. Participants are invited to answer a question or share their opinion about a statement related to the theme of this object. However, the participation is absolutely voluntarily. Another distinguish feature of the courses is that they were open to public for a long time unlike most courses that are available for several weeks and then are closed. This led to a different way to measure the time of dropping out.

First MOOC is called “Introduction to xAPI”. It targets company people who deal with learning analytics as xAPI is a new form of LMS (Learning Management System). That means that the audience is supposed to be quite narrow and specific. Though already being specialist-orientated the course still provides two tracks – technological and strategical. The former reveals more technical mechanisms behind the system and the latter introduces general implications and purposes of the system. The technical track consists of 21 objects and the strategical track covers 32 objects. Learners were free to choose one track or to follow the both. The total number of the participants was 169 with 102 following the strategic track and 88 doing the technical one. Some data were visualized for the course in general and some for each track separately. That allowed comparing within the course as well as outside the course.

The other course is “How to Create an Outstanding MOOC” with one track and 35 objects in it. The course is targeted to those who want to create MOOCs themselves and thus should attract a more diverse audience. There were 152 participants in the course discussions.

Chapter 4. Results

Results for the "Introduction to xAPI" course

To start the research, the most straightforward parameter of the course, which is dropout, was assessed. To do it the number of comments per object was counted for both tracks separately and the results are presented in Fig. 1 and Fig. 2 for the technical and the strategic tracks respectively, where *n* is the number of comments and *object_id_new* is an ordinal number of the object, which gives the chronological order of topics in the course.

These data show that there is a dropout of forums as the number of comments per object decreases dramatically for both tracks. In other words, there is a dropout from forums even in MOOCs with digital badges. It is worth mentioning that the tracks have different decreasing rates. For the technical track (Fig. 1) the biggest fall happens after the first third of the course and then the number fluctuates around 12 comments per object. This can illustrate the idea that learners are enthusiastic to try this track but most of them still find it too difficult so they stop and probably focus on the strategical track. Those who survive this first period stay till the very end.

Situation with the strategic track (Fig. 2) is a bit more traditional. The number of comments goes down steadily during the whole period, which means that it is unlikely caused by some particular topic or difficulty level. However, a similar to the technical track pattern can be registered – first 10 objects and the rest of the course have different decline. This aligns with general belief that first several topics can be done to get the impression of the course and not to study it seriously.

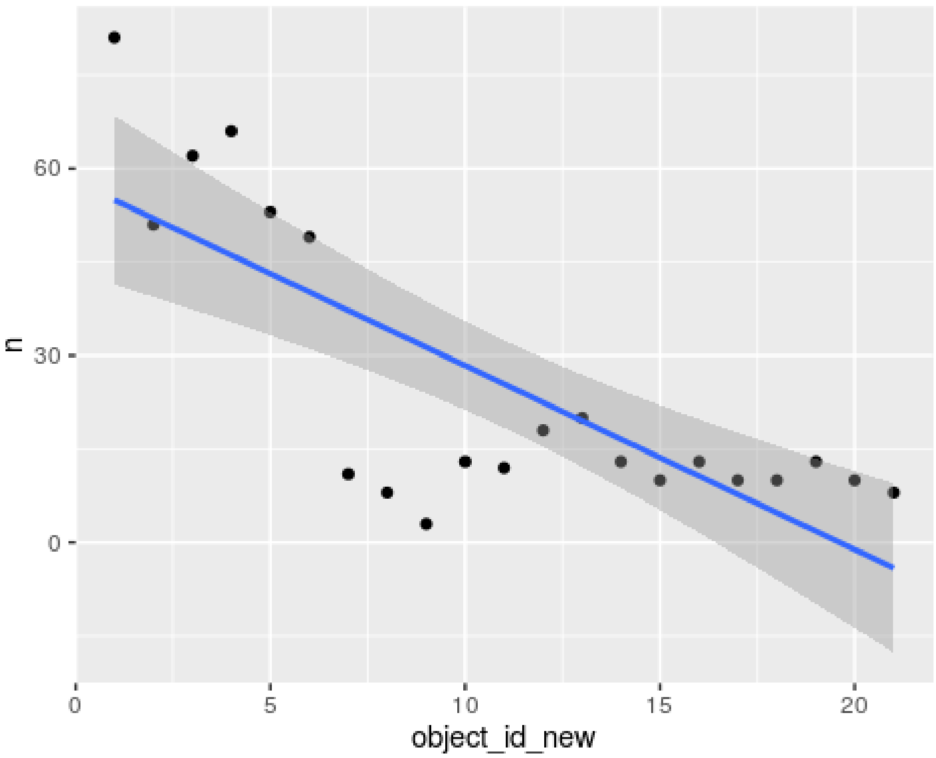

Fig 1. Number of comments per object in the technical track

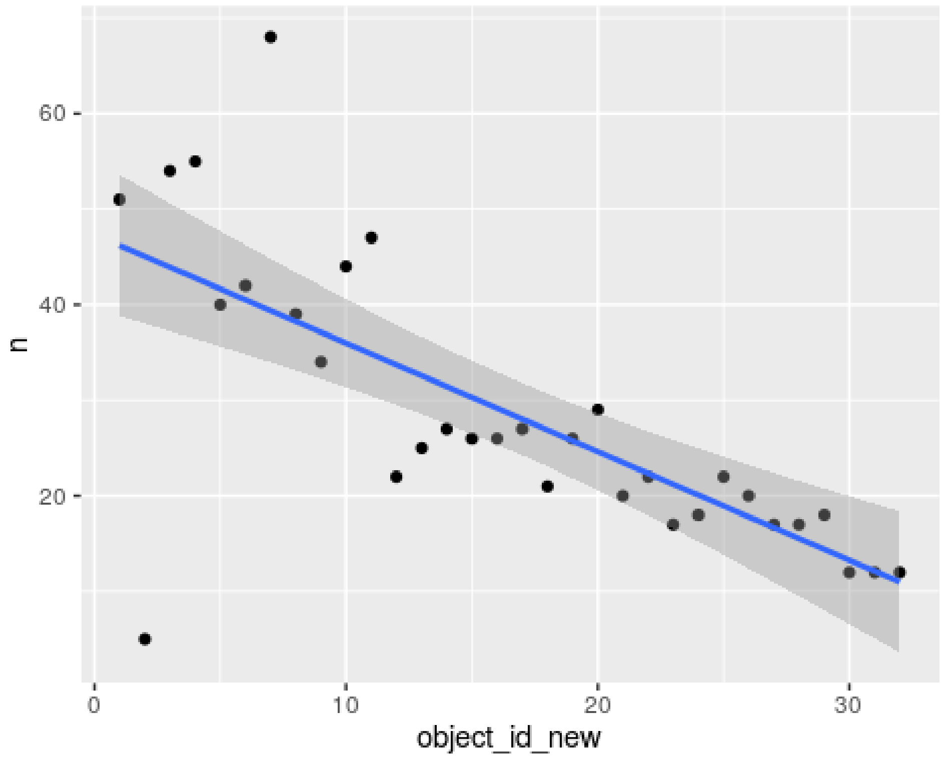

Fig 2. Number of comments per object in the strategical track

Having discovered that dropout exists made it logical to proceed with the further study. The next part was sentiment analysis conducted at the both course or track levels. To assess the sentiment level of a sentence R packages (tm, quanteda and NLP) were used and a specially written code compared each word in the sentence with an extended list of positive and negative words based on the dictionary meanings.

The sentiment level of an object is a median of sentiment levels of all comments given to this object. Median is chosen over an average to decrease the influence of outlyers. A sentiment "map" of a course or track is a graph with the medians of all objects. Fig. 3 shows this type of a graph of median sentiment score per each object (part of the course). The first important observation is that the whole graph is in positive area. So no medians below zero were registered for the whole course. Another finding is that there is no obvious dynamic with the course going on as the most common median is 1. The drops and rises in the pattern are more likely to be caused by the content of the particular part (object) than the general tendency.

It is worth mentioning that there is no link between the type of the object and the corresponding sentiment level. Among nine objects with the lowest sentiment level (median score = 0) there are three videos, three reading pieces (two articles and one blog), two tasks and one solution to the task. This range demonstrates all types of learner's activities within the course and every type is represented almost equally. The situation for the parts with the highest sentiment level (median score >1.5) is similar. Among eight objects there are two videos, two reading pieces (an article and a blog), three tasks and one solution. In this group all types of activities are also presented, though with a small dominance of practical parts – tasks together with solutions take fifty per cent. The amount of data does not allow making statistically significant inferences, however, the idea of very positive reaction to practical parts should be taken into account by course developers.

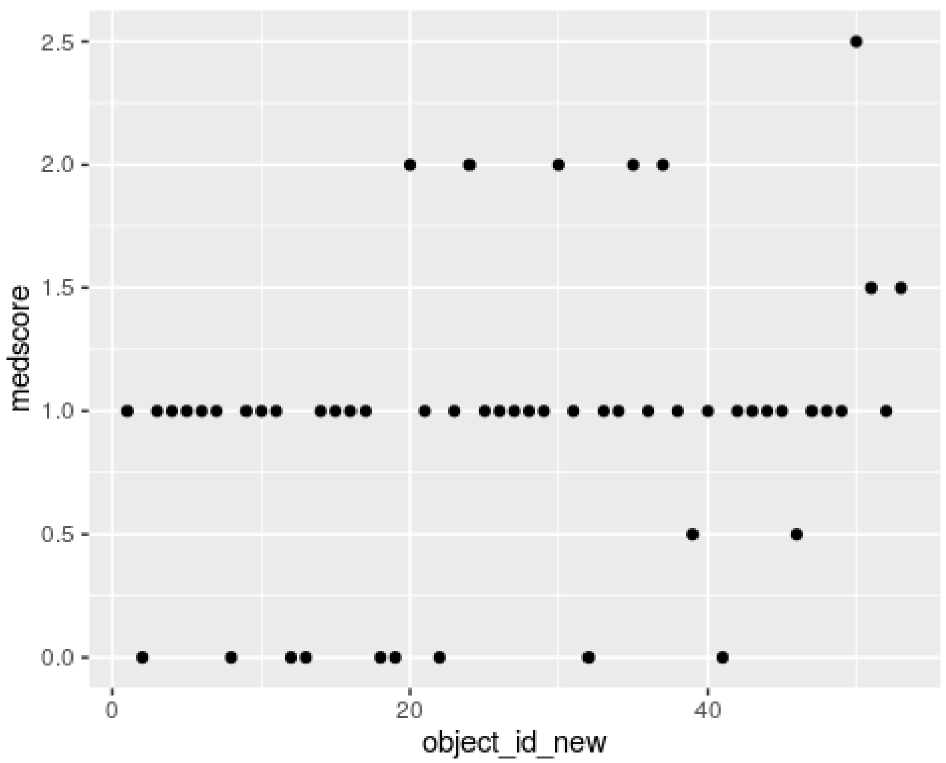

Fig. 3. Medians of sentiment score for the whole course

A similar picture with the absence of clear sentiment dynamic is true for both tracks taken separately. Fig. 4 presents the technical track and Fig. 5 the strategical track. These data show that there is no evidence of existence of a pattern in changing sentiment with time. Moreover, the fact that the median stays almost the same suggests that both people who comment positively and people who comment negatively drop the course.

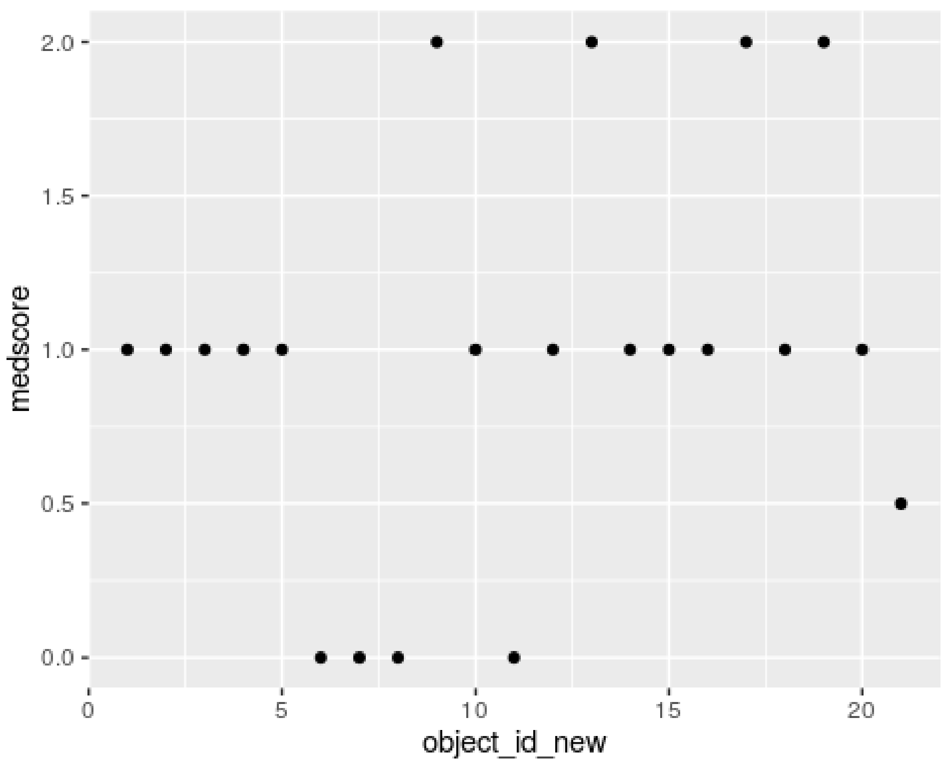

Fig. 4 Medians of sentiment score for the technical track

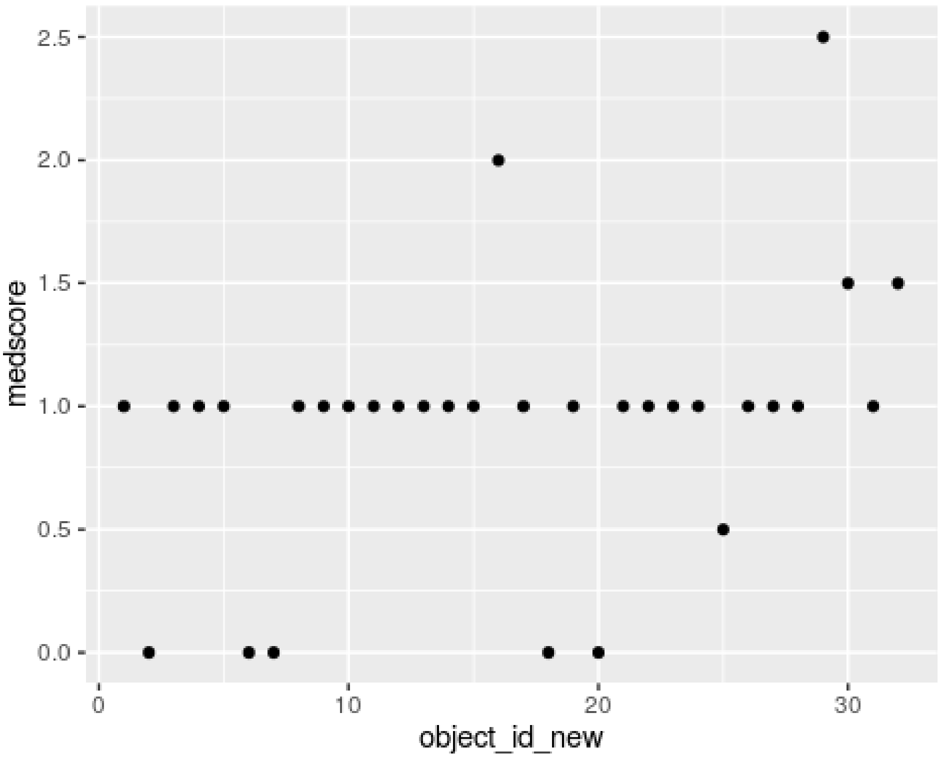

Fig. 5 Medians of sentiment score for the strategical track

Another metric, which was used to evaluate the general trends in the tracks, is the share of negative comments per object. This is a part of comments with the sentiment level below zero. The results are presented in Fig. 6 and Fig. 7 where *object_id_new* is an ordinal number of the object. A linear regression line and a 95% confidence interval is added to the graph. There is a slight difference in patterns for tracks. While the share of negative comments increases slightly for the strategic track, the share for the technological track goes down slightly with time. The growth could be seen as more critical attitude towards the course by those learners who stay. And the decrease can be explained by the fact that those learners who found the track too difficult had quit it by this moment. This explanation can also be attributed to a small difference in the absolute numbers between the tracks. While for the strategic track some negative comments are posted throughout the whole course, for the technical track there are almost none of them at the end. However, the patterns are not very clear and do not allow making strong inferences.

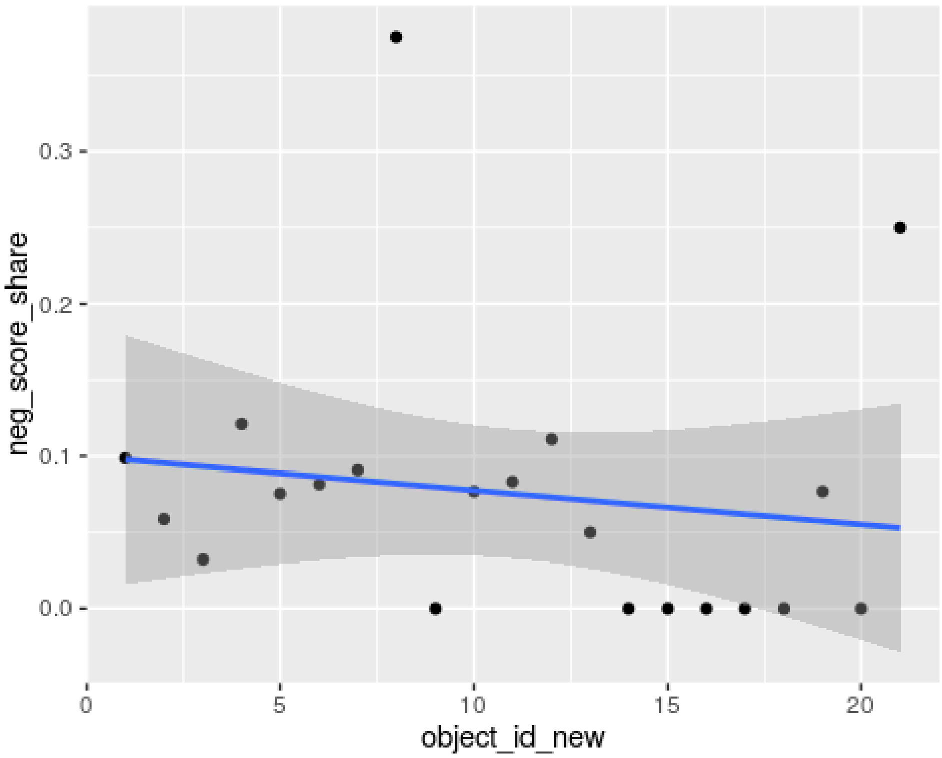

Fig. 6 Share of negative comments per object in the technical track

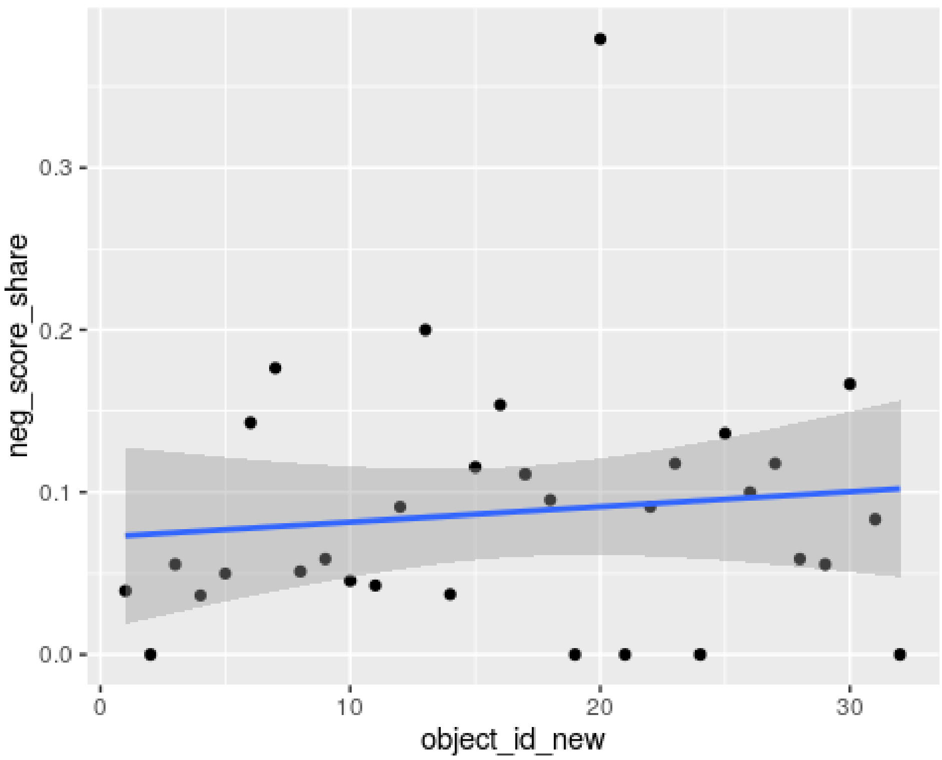

Fig. 7 Share of negative comments per object in the strategic track

To answer the first research question the correlation analysis of the relationship between the sentiment level (score) and drop out from forums was conducted. As the distribution of sentiment medians is skewed to the positive area (right side), Spearman's correlation coefficient was calculated for both tracks separately. Results for both tracks have too high p-value so they are not statistically significant: for the technical track $r = .09$, $p = .39$, for the strategical track $r = .17$, $p = .07$. That means that with the existing data the correlation between the sentiment level (score) and drop out from forums was not found.

The next step was to check if this trend is true for different groups of learners. So they were divided according to their completion rates. Completion rate is measured by the ordinal number of the object they commented the last, in other words, the topic they stopped commenting at. This method was used instead of usual date of the last post as the courses stay available for public for a long period and time marks were not informative any more. The percentage of learners commented 70% or more topics and 30% or less topics are presented in the Table 1. 70% cutoff point was taken as in the academic sphere it is commonly used as a cutoff point for the completion of any course. 30% represents the lowest third so this way participants are grouped always equally in terms of statistics, which is obviously different in real life.

Table 1. Percentage of stayed and dropped students for both tracks.

Track	Percentage of completing learners (commented 70% of topic or more)	Percentage of dropped students (commented 30% of topics or less)
Technical	17%	50%
Strategical	31%	39%

After grouping the ANOVA test was conducted to find out if there is any difference in sentiment scores for these three groups – stayed, dropped and in-between. The result for both tracks is the same – there is no statistically significant difference between these groups: $F(2, 85) = 2.16$, $p = .12$ for the technical track and $F(2, 99) = .71$, $p = .49$ for the strategical track.

As the data distribution is not normal ANOVA test might not be the most exact so a Kruskal-Wallis test, which is non-parametric, was conducted as well for the mentioned above categories of learners (Field, Miles & Field, 2012). The results are similar and show no significant difference between the groups in their completion rate depending on the average sentiment score: $H(2) = 5.94$, $p = .051$ for the technical track and $H(2) = 3.04$, $p = .22$ for the strategical track.

The next step was to investigate the relationships between a dropout and some other variables at the student level. An average sentiment score does not seem to influence a completion rate in neither technical (Fig. 8) nor strategical (Fig. 9) tracks. Controlling for the outlier in the technical track makes the confidence interval more narrow but does not change the general picture. Participants with the same completion rate may have very different average sentiment scores. For example, sentiment score for a completion rate of .25 varies from .5 to 3 in the technical track. It is also true to say that people with the same sentiment score have very different completion rates. For instance, for both tracks some of those whose sentiment score is 2 did less the 10% of the course and some completed it fully. This may be an indicator of a more complex nature of these relationships.

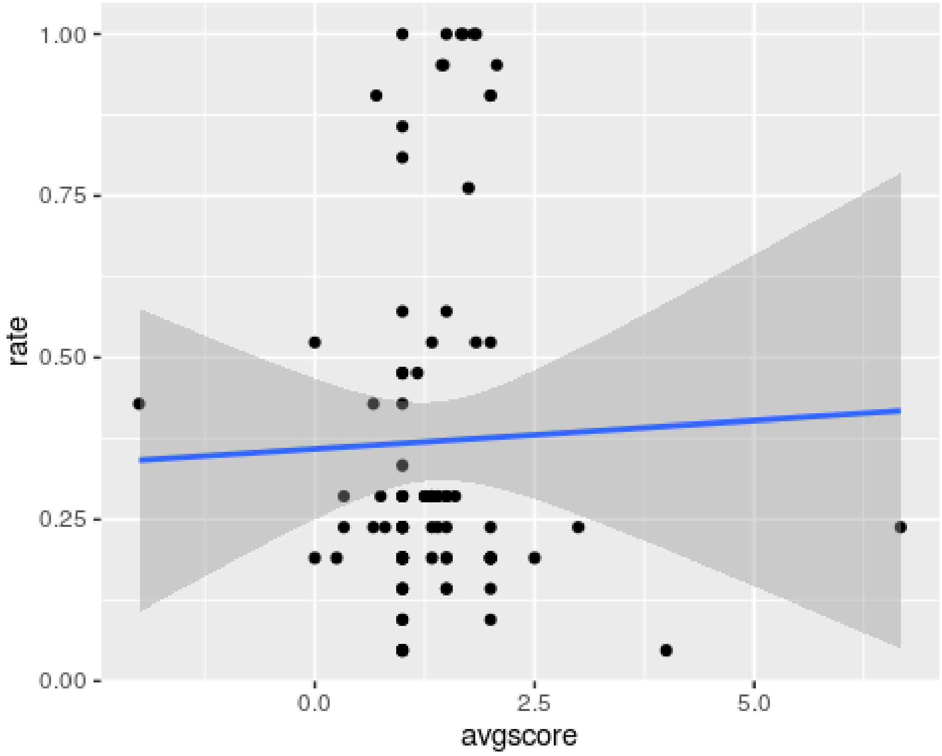

Fig. 8 Average sentiment and completion rate for learners of the technical track

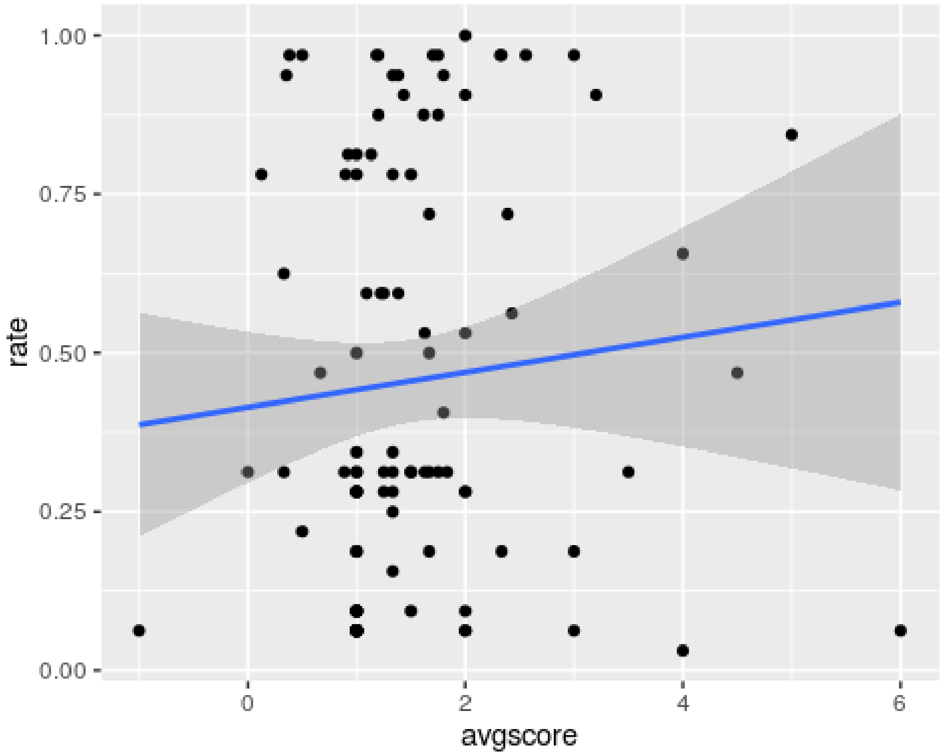

Fig. 9 Average sentiment and completion rate for learners of the strategic track

Another studied variable that was a minimal sentiment score. Fig. 10 and Fig. 11 show the relationships between it and the completion rate. Both tracks demonstrate a very weak relationship and a linear regression is hardly an answer to the question about the link between them. This also suggests that the nature of the relationships is more complex. Therefore, a more complicated method to investigate that was used.

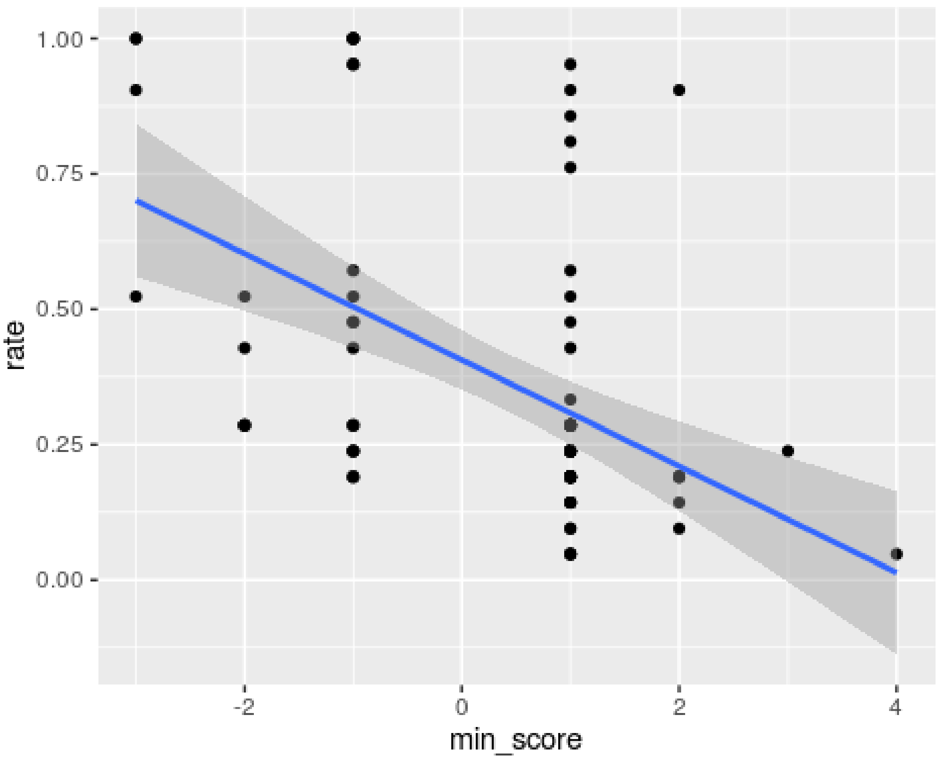

Fig. 10 Minimal score and completion rate for learners of the technical track

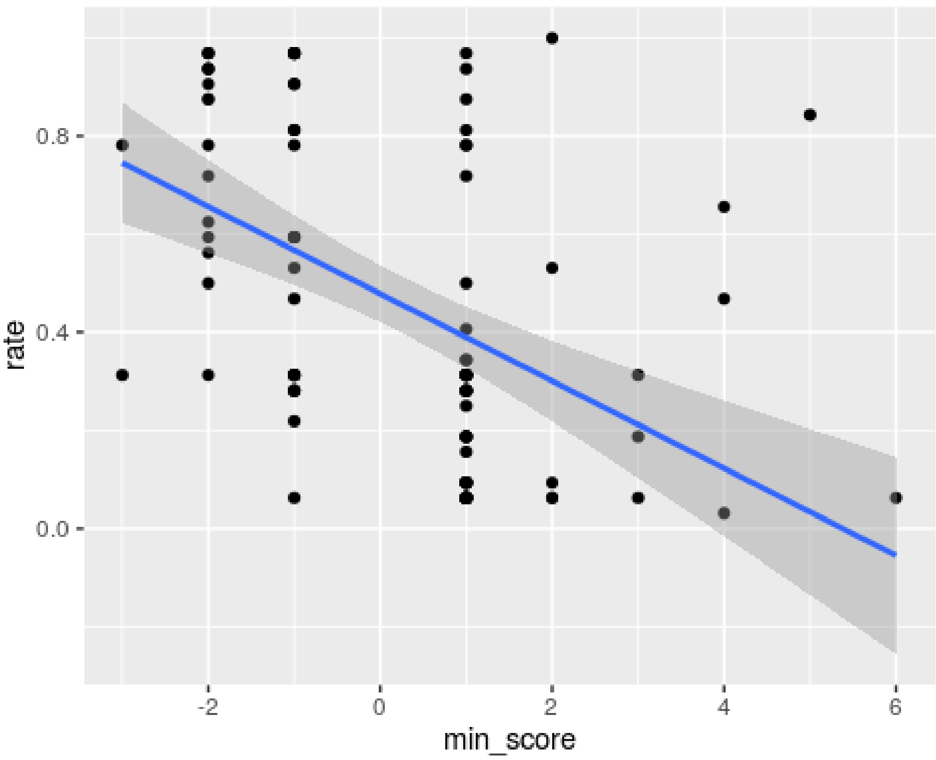

Fig. 11 Minimal score and completion rate for learners of the strategic track

To discover which variables, if any, influence the completion rate the conditional inference tree method was used, which presents a non-parametric class of regression trees applying tree-structured regression models into the theory of conditional inference procedures. This method allows identifying influential parameters and presents the results in the form of groups split by the biggest statistically significant difference. The nodes present the split point and the parameter value. This analysis was done with the conditional inference trees in the R package *partykit*. This package is suitable for all kinds of regression problems (Hothorn, Hornik & Zeileis).

For the analysis the variables referring to participants were employed. In addition to the those used for regression analysis such as an average score, a minimal score and a share of negative comments, a maximal score and a share of positive comments were included. These last two variables are a logical continuation of their negative counterparts and aim at increasing the validity of the models. The dependent variable is the completion rate, which varies from 0 to 1. The results are shown in Fig. 12 and Fig. 13.

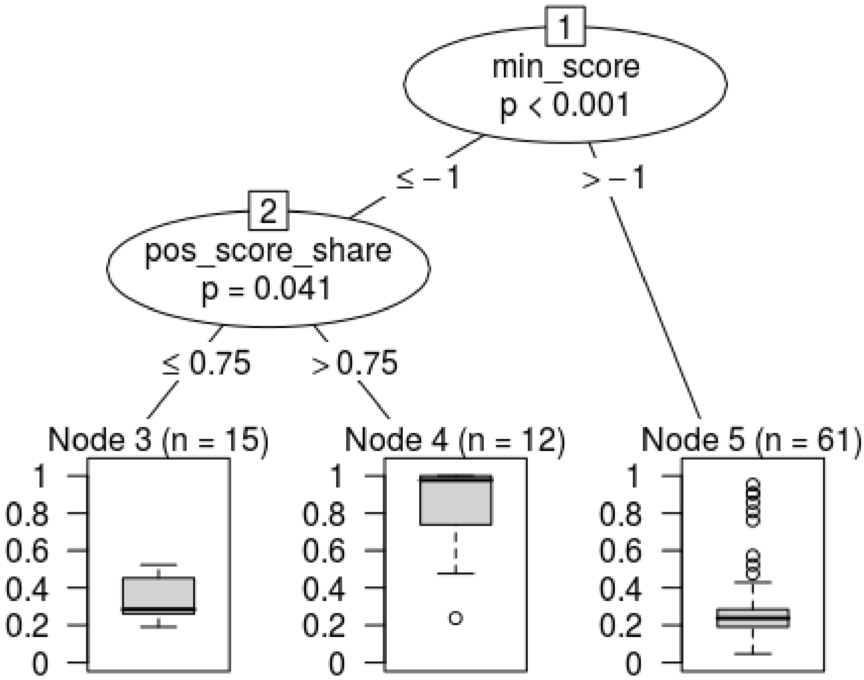

Fig. 12 Conditional inference tree for the technical track

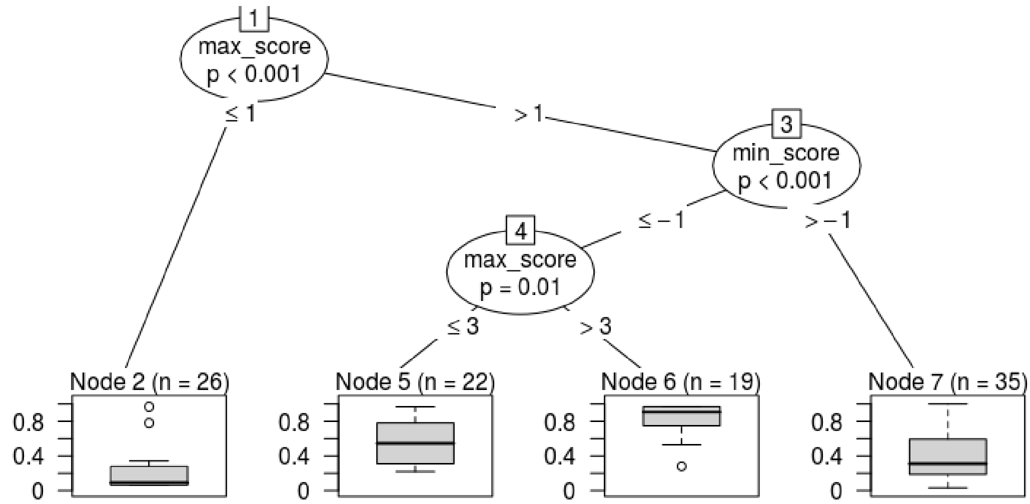

Fig. 13 Conditional inference tree for the strategic track

It is worth mentioning that an average sentiment score does not seem to play a significant role as it is not shown in any split. More local characteristics such as minimal or maximal scores are demonstrated to be more important for the analysis. Though the models look a bit different some common patterns can be identified.

The first and the most surprising finding is that for both tracks the groups with no negative comments (minimal score >-1) have a very low or the lowest completion rate around .3. They are node 5 for the technical track and node 7 for the strategical. In other words, those participants who post only positive comments are prone to quit very soon as they do not invest an effort into the studying process.

The second observation, which is quite intuitive, is that the group with a big amount of negative comments is also very likely to drop. For the technical track that is shown through a low positive comments share (node 3) and for the strategical track it is demonstrated by a low maximal score (node 2). For both tracks the completion rate for these groups is also around .3.

For both tracks the highest completion rate is registered for the groups with some but not too many negative comments on a generally positive background. This is node 4 for the technical track and nodes 5 and 6 for the strategic one. In other words, these results allow suggesting that some critical attitude is necessary for the full completion and no negative comments predict a quicker dropout. The latter is unexpected and, thus, even a more important finding as it warns not to overestimate too positive sentiment level.

These findings somewhat align with the research about the importance of critical engagement in online courses (Garrison & Cleveland-Innes, 2005; Zhu, 2006). Just interaction by itself does not automatically mean that the learners are involved in the process cognitively as this can be social interaction or surface engagement. Studies show that closed-end questions do not stimulate neither constructive discussions nor knowledge construction (Ke & Xie, 2009). This means that the course structure and discussion questions should move participants to the higher-order thinking levels. Cognitive engagement can be retrieved and analysed on the base of discussion messages, however, this requires further research.

For the technical track R^2 =.45 and for the strategical track R^2 =.43 which, though being not very high, can still demonstrate some trends. Such low figures can be partly explained by a quite low from a statistical point of view number of participants.

Results for the "How to create an outstanding MOOC" course

The same procedure of the analysis was done to the data set of the other course – "How to Create an Outstanding MOOC". The results are quite similar in many ways.

Quite predictably there is a dropout from the course. The number of comments per object as shown in the Fig. 14 decreases steadily with time, which is similar to the strategical track situation.

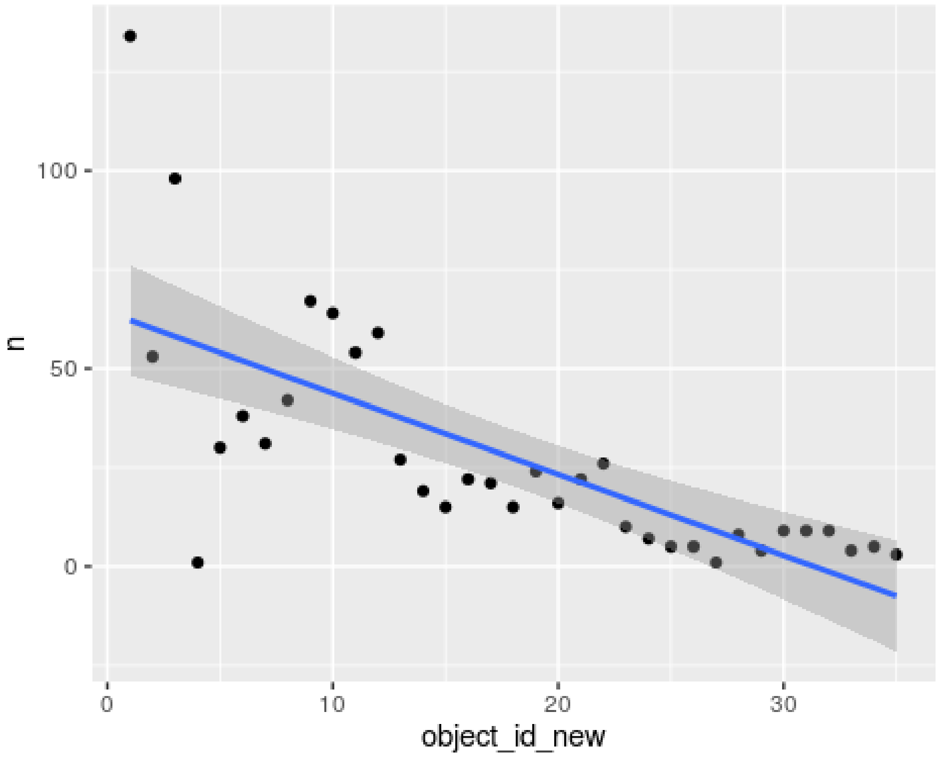

Fig. 14 Number of comments per object

Fig. 15 presents the medians of sentiment level for the whole course. As in the first MOOC there is no clear pattern and occasional drops happen through the whole period.

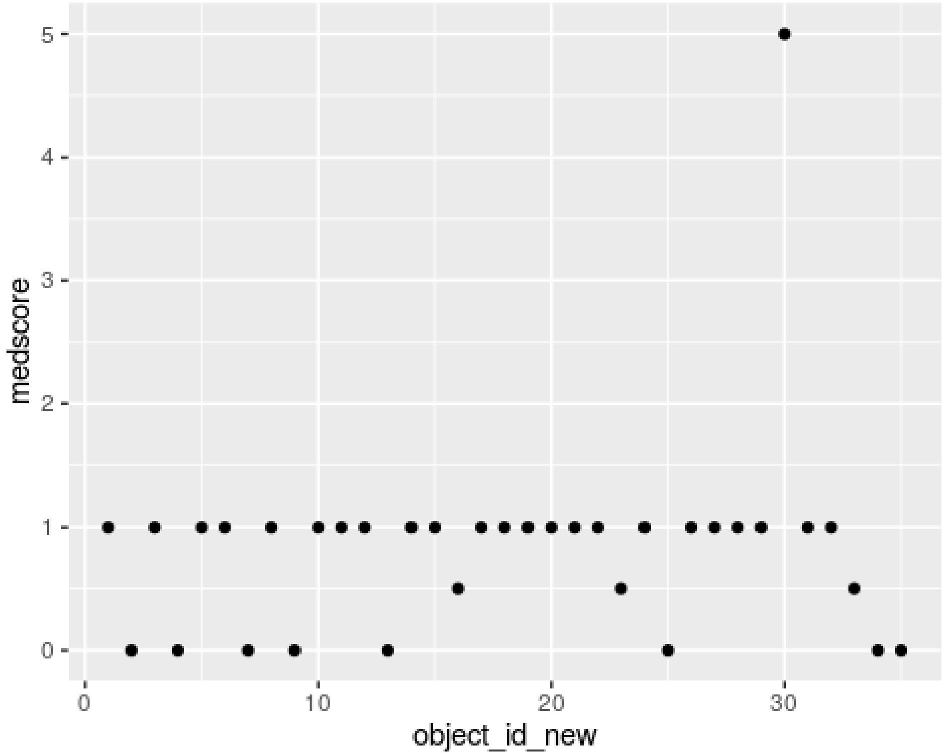

Fig. 15 Medians for the course

The share of negative comments per object is given in the Fig. 16 with the regression line and 95% conference interval. As in the previous cases it does not seem to have any clear pattern so no statistical inferences are possible to be made.

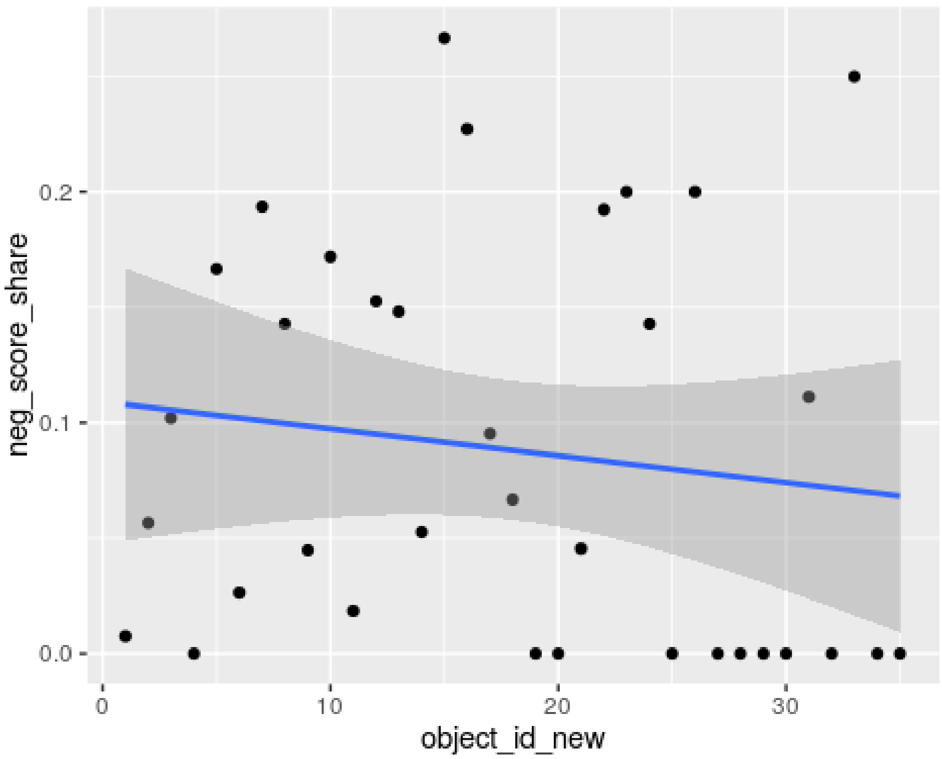

Fig. 16 Share of negative comments per object

As well as for the first course the relationship between the sentiment level and the drop out was check with the help of Spearman's correlation analysis. And as in the previous cases p-value is too big to make this relationship statistically significant: $r = -.13, p = .14$, though the fact of a negative correlation might have been quite interesting to study.

To check if this trend stays for different groups of learners, they were divided according to their completion rate with the same basis as for the other course. The percentage of learners who stayed and dropped the course are given in the Table 2. An interesting difference with the other course is that the upper and the lower groups are much smaller than the 'between' group. However, the investigation of this is beyond the scope of this research.

Table 2. Percentage of stayed and dropped students.

	Percentage of completing learners (commented 70% of topic or more)	Percentage of dropped students (commented 30% of topics or less)
Outstanding MOOC	16%	17%

ANOVA test was also conducted to explore if there is any difference in sentiment scores for these three groups – stayed, dropped and in-between. The result shows that there is no evidence of a statistically significant difference between these groups: $F(2, 131) = 1.08, p = .34$.

As with the previous course, a Kruskal-Wallis test was conducted to serve skewed data analysis better with the following results obtained: $H(2) = 2.67, p = .26$. This shows that no statistically significant difference between the groups were discovered, which is similar to the results of both tracks in the first MOOC.

Checking for the relationship between the completion rate and the average score (Fig. 17) and the completion rate and the minimal score (Fig. 18) gives the results similar to the first course. The average rate seems to be very loosely related to the completion rate. The minimal score demonstrates a bit more of relationship but the linear regression might not be the best-fitting model in this case either. This leads to the necessity to apply a more complex model.

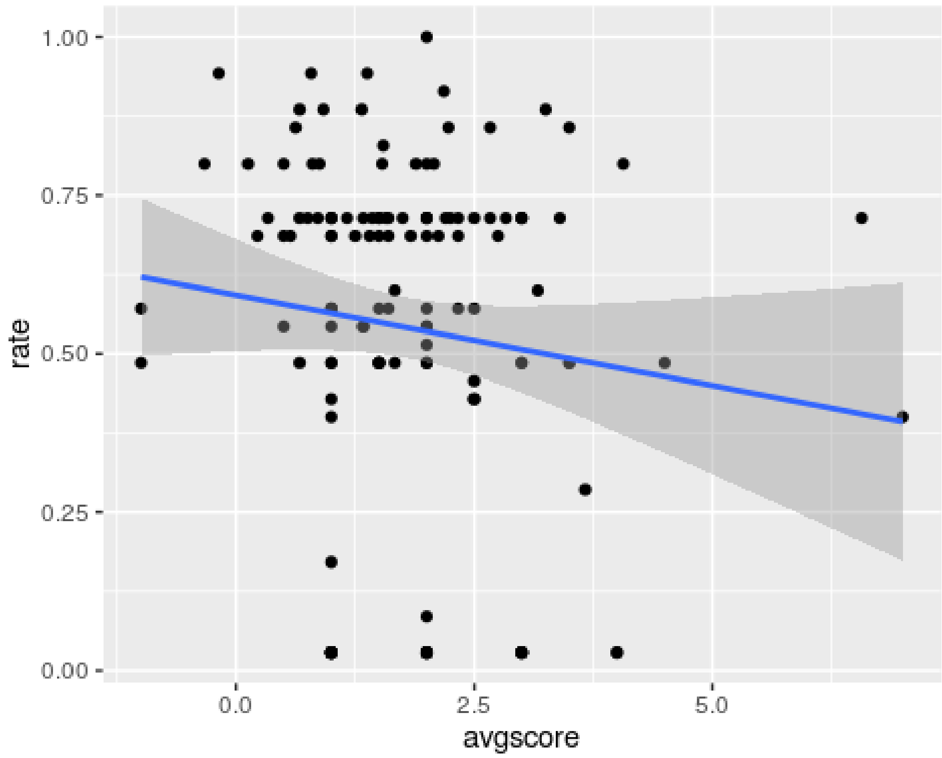

Fig. 17 Average score and completion rate.

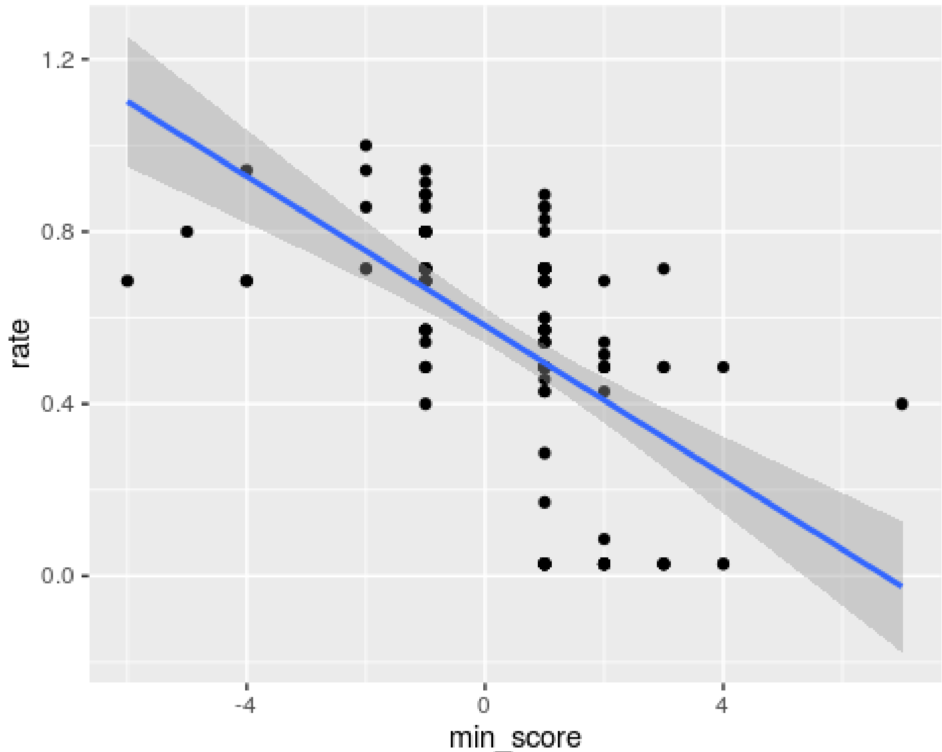

Fig. 18 Minimal score and completion rate.

The rationale and the method behind creating the conditional inference tree for this course is the same as for the first course. The tree is shown in Fig. 19.

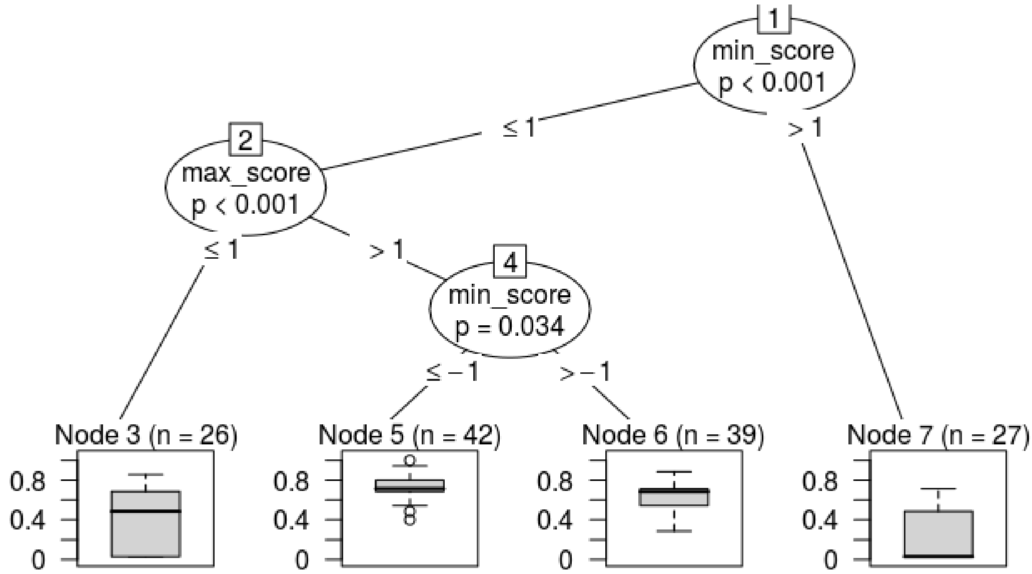

Fig. 19 Conditional inference tree for the course "How to create an outstanding MOOC".

The assumptions that were made for the models in the first course are still relevant. Participants with very positive minimal score have a very low completion rate (node 7), which supports the idea of the importance of critical engagement in the course. Learners with lower maximal score also demonstrate lower completion (node 3), which resembles the splits on the previous models for people with low maximal score or positive comments share. The highest completion rate is still demonstrated by groups with some amount of negative comments.

As for the first course R^2 was calculated for this model: R^2=.49. This figure also shows some, but not very high, explanation power of the model. The possible reason for it being so low is again a small in statistical terms number of participants.

Chapter 5. Conclusion and discussion

The first results prove that MOOCs with digital badges which stimulate participants to take an active part in discussions still experience dropout from forums. The number of comments per object decreased for both courses and for both tracks in the xAPI MOOC. This illustrates that the problem of dropout is topical for this type of MOOCs as well and more research is needed to study this phenomenon.

At the course level several different analyses were conducted. Unlike in the previous researches (Tucker & Divinsky, 2014, Chaplot, Rhim & Kim, 2015) changes in sentiment levels in both courses do not demonstrate any particular pattern. The graph of average sentiment seems to fluctuate around 1 and occasional falls and rises do not depend on the type of the object – video, article, exercise – or its order in the course. The share of negative comments per object does not seem to have any clear dynamic either. This absence of a detected pattern did not allow to formulate any general recommendations for course developers.

Discovered in other studies (Wen, Yang & Rose, 2014, Chaplot, Rhim & Kim, 2015) relationships between a lower sentiment score and dropout was not proved by the present data sets. Moreover, in terms of this lack of relationships no difference was found between different groups of students – those who completed more than 70 per cent of topics, those who completed less than 30 per cent and the middle group. This might be explained by the fact that studied MOOCS are with digital badges and participation in discussions are taken more seriously. Learners' behaviour in other types of courses can be different as discussions can be viewed as an additional part of the course which can be skipped completely. Furthermore, due to the nature of the provided data there was no possibility to assess learners' engagement by using login data. This means that no relationships between quitting the forum and stopping visiting the course website could be studied. This can be viewed as one of the limitations of the research and influences the ability to generalize the results.

At the student level the following variables were used for the analysis: an average score, a minimal score and a share of negative comments. No significant link was identified neither between an average score and a completion rate, nor between a minimal score and a completion rate. These results were received for both courses and for both tracks within xAPI course.

This lack of linear correlation might be accounted for a comparably small number of participants as for each course the figure was a bit more than 150 while many courses on different platforms attract hundreds or even thousands learners. However, more likely that this straightforward relationship does not exist at all and more complicated phenomenon with local effects different for different groups of learners takes place.

To get a better understanding of these more local effects and to identify if any variables influence completion rate significantly the method of a conditional inference tree was used. For the analysis the variables referring to participants were employed such as an average score, a minimal score, a share of negative comments, a maximal score and a share of positive comments. For both tracks of xAPI MOOC the results were quite similar. First important split was between the group with a negative minimal score and the group with a positive minimal score. Those with a positive minimal score had a lower completion rate. Another important split was the amount of negative comments demonstrated through the maximal score or the share of positive comments. The same trends could be seen for the other MOOC, which means that the main conclusions are supported by all three models.

For all courses and tracks the lowest completion rate was registered for the group of participants with the highest minimal score. In other words, those who do not express any negative emotions are likely to drop the course very soon. One can probably interpret that as the lack of real involvement in the course. On the other hand, those who express negative emotions a lot are also in the group of risk in terms of completion. This is a more intuitive result as those who are not happy with the course should be also more prone to quit. The highest completion rate in associated with the group of learners who have some negative comments but not too many. It means that some critical cognitive involvement in the course is vital for succeeding, which is supported by the research about the importance of critical engagement in online courses (Garrison & Cleveland-Innes, 2005; Zhu, 2006). These findings may influence the focus of MOOC teachers and course developers. Critical comments can be more appreciated as they signal more involvement and thus potentially more completion. This means that instead of looking for negative sentiments as dropout predictors, which is a very intuitively obvious assumption, course developers should pay more attention to the part with the most positive sentiment. However, mentioned above limitations to generalize the results require more studying of this area.

Prior to the directions of the research it is worth mentioning its ethical ground. Online learning, especially within MOOCs, stands on a completely different technical basis than a traditional studying process. A lot of data are collected automatically and its usage and analysis can enhance understanding of this type of learning, which is still quite new, and course development. This technological difference requires different research approaches. And though a researcher should still be very careful when human objects are involved there is no possibility to use the same techniques and standards as for a face-to-face experiment. Some other scientific fields have been going through this type of shift in standards and re-evaluating the research procedures. For instance, there are many studies in the area of Social Sciences which use data from Twitter or Facebook posts and profiles to investigate different relationships. In this process no direct consent from users is acquired based on the assumption that these data were made public and open by their creators themselves. Standards for the Internet research are being set now in different scientific fields and it seems reasonable to have them different from their more traditional counterparts. Educational Science should not stay behind in this process and not miss the opportunity to use a completely new type of data in a research process.

In terms of the present research more investigation can be done in the area of learners' cognitive involvement and deeper lexical analysis of their comments to find the indicators of such engagement. Trying to deconstruct sentences and identify particular aspects such as a sentence structure, reasoning or logic linking devices etc. can be a next step in this direction. Another interesting aspect of the analysis is neutral statements as they may include more cognitive than emotional part. Specifying what is neutral for sentiment analysis and what information can be extracted from such sentences also seems a potential direction of the research.

As for the present results practitioners can redirect their attention and view a group of learners with only positive comments as the group with a high probability of dropping out. The main focus should be given to those who do include some negative comments though generally staying in the positive area. This behaviour may signal that such participants put some effort in the learning process and have the highest chances to complete the course.

References

Belanger, Y., & Thornton, J. (2013). Bioelectricity: A quantitative approach Duke University's first MOOC.

Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment Analysis: Measuring Opinions. *Procedia Computer Science*, *45*, 808–814.

Boling, E. C., Hough, M., Krinsky, H., Saleem, H., & Stevens, M. (2012). Cutting the distance in distance education: Perspectives on what promotes positive, online learning experiences. *The Internet and Higher Education*, *15*(2), 118-126.

Cercone, K. (2008). Characteristics of adult learners with implications for online learning design. *AACE Journal*, *16*(2), 137–159.

Chaplot, D. S., Rhim, E., & Kim, J. (2015). Predicting student attrition in MOOCs using sentiment analysis and neural networks. In *Proceedings of AIED 2015 Fourth Workshop on Intelligent Support for Learning in Groups*.

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E. J. (2013). The MOOC phenomenon: who takes massive open online courses and why? *Available at SSRN 2350964*.

Coffrin, C., Corrin, L., de Barba, P., & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in MOOCs. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 83–92).

Field, A., Miles, J., Field, Z. (2012) Discovering Statistics Using R. SAGE Publication Ltd.

Feldman, R. (2013). Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science. *Communications of the ACM*, *56*(4), 82–89.

Garrison, D. R., & Cleveland-Innes, M. (2005). Facilitating cognitive presence in online learning: Interaction is not enough. The American Journal of Distance Education, 19(3), 133-148.

Hothorn, T., Hornik, K., & Zeileis, A. ctree: Conditional Inference Trees.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, 15(3), 651-674.

Hrastinski, S. (2009). A theory of online learning as online participation. *Computers & Education*, *52*(1), 78-82.

Ke, F., & Xie, K. (2009). Toward deep learning for adult students in online courses. *The Internet and Higher Education*, *12*(3), 136–145.

Kennedy, H. (2012). Perspectives on Sentiment Analysis. *Journal of Broadcasting and Electronic Media*, *56*(4), 435–450.

Lee, Y., Choi, J., & Kim, T. (2013). Discriminating factors between completers of and dropouts from online learning courses. *British Journal of Educational Technology*, *44*(2), 328-337.

Liu, B. (2011). Opinion Mining and Sentiment Analysis. In B. Liu, *Web Data Mining* (pp. 459–526). Berlin, Heidelberg: Springer Berlin Heidelberg.

Liyanagunawardena, T. R., Lundqvist, K. Ø., & Williams, S. A. (2015). Who are with us: MOOC learners on a FutureLearn course: Who are with us: MOOC Learners. *British Journal of Educational Technology*, *46*(3), 557–569.

Liyanagunawardena, T. R., Parslow, P., & Williams, S. (2014). Dropout: MOOC participants' perspective. *Presented at the EMOOCs 2014, the Second MOOC European Stakeholders Summit, Lausanne, Switzerland,* 95–100.

Mackness, J., Mak, S., & Williams, R. (2010). The ideals and reality of participating in a MOOC.

Macleod, H., Haywood, J., Woodgate, A., & Alkhatnai, M. (2015). Emerging patterns in MOOCs: Learners, course designs and directions. *TechTrends*, *59*(1), 56–63.

Markham, A. & Buchanan, E. (2012). Ethical decision-making and internet research. Retrieved from http://pure.au.dk/portal/files/55543125/aoirethics2.pdf

MOOC @ Edinburgh 2013 – Report #1 (2013). *MOOC @ Edinburgh 2013 – Report #1*. University of Edinburgh, Edinburgh, Scotland.

Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 Proceedings*, 5825–5834.

Park, J.-H., & Choi, H. J. (2009). Factors influencing adult learners' decision to drop out or persist in online learning. *Journal of Educational Technology & Society*, *12*(4), 207–217.

Perna, L. W., Ruby, A., Boruch, R. F., Wang, N., Scull, J., Ahmad, S., & Evans, C. (2014). Moving Through MOOCs Understanding the Progression of Users in Massive Open Online Courses. *Educational Researcher*, 0013189X14562423.

Reich, J. (2015). Rebooting MOOC research. *Science*, *347*(6217), 34–35.

Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., & Sherer, J. (2014). Social factors that contribute to attrition in moocs. In *Proceedings of the first ACM conference on Learning@ scale conference* , 197–198.

Siemens, G. (2014). Connectivism: A learning theory for the digital age.

Tucker, B. C., & Divinsky, A. (2014). Mining student-generated textual data in moocs and quantifying their effects on student performance and learning outcomes. In *2014 ASEE Annual Conference, Indianapolis, Indiana, Indianapolis, Indiana*.

Wen, M., Yang, D., & Rose, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In *Educational Data Mining 2014*.

Wu, Y. C., Hsieh, L. F., & Lu, J. J. (2015). What's The Relationship between Learning Satisfaction and Continuing Learning Intention?. *Procedia-Social and Behavioral Sciences, 191*, 2849-2854.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, *58*, 119–129.

Yang, D., Sinha, T., Adamson, D., & Rose, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop* (Vol. 11, p. 14).

Zhu, E. (2006). Interaction and cognitive engagement: An analysis of four asynchronous online discussions. Instructional Science, 34(6), 451-480.