Learning analytics towards identifying the processes of higher order thinking in online discussions of MOOCs

Master thesis

Student:

Ingrid Smalbergher

Supervisors:

DR. BAS KOLLÖFFEL

PROF. DR. IR. BERNARD VELDKAMP

UNIVERSITY OF TWENTE.

Acknowledgements

First, I would like to thank my two supervisors Dr. Bas Kolloffel and Prof. Dr. Ir. Bernard Veldkamp for the opportunity they gave me to work on this project for my Master Thesis. I would also like to thank you for investing your time in helping me and providing me with so much support while doing this research and writing my thesis. In addition, I want to express my appreciation for the fact that you allowed this project to be my own work by supporting me in the decisions I made.

Bernard, your inputs were of true value to me, they helped me to find my way in moments where I felt a bit lost with the specific direction of the project by helping me to understand what the most important things that I should have in mind were.

Bas, first I would like to thank you for your always calm and supportive energy. This helped me especially in moments where I felt overwhelmed by stress to get back on track and to remain constructive. Thank you also for the patience you had in working with me, for your guidance and openness to my ideas. Furthermore, I want to thank you outside the context of this thesis for the chance I had to discover my passion for Learning Analytics within the course 'HRD and technology in a live context' and my interest in the evaluation of MOOCs through the course 'Trending Topic- MOOCs as tool for continuous learning'.

I would also like to thank Sytske Wiegersma for her help and support during my project.

Next, I would like to thank my friends Adhitya, Jolien, Tom, Angela, Remco, Rheza, Gaia, Wouter, Ruxandra and Andreea for making this experience of studying abroad wonderful. I appreciate all your support, our great days of studying together, cooking, playing, laughing or simply talking. Alexandra, I very much appreciate our friendship, the big support I had from you throughout the master and for this thesis. Kristel you've got a special place in here, not (only) for being my person in my everyday life since I moved here, but for how much help I had from you during this final project and especially in the last days of writing this paper. Thank you! Also, to my friends from Romania (or other countries), thank you for being in my life even from distance, thank you for the late or morning online calls, and for the feeling that nothing changed in our relationships every time we get in touch.

Last but not least, I would like to thank my family for giving me the opportunity to study abroad. I am grateful for your continuous encouragements throughout this study and your full support in my development for so many years now. Thank you Cezara, Gabriela, and Danut for your unconditional love and for being my family to the full extent of this word.

Summary

More and more organizations all over the world focus on learning innovations, providing digital education, and researching in best ways to support individuals' learning and development. Even so, the current evaluation systems in Massive Open Online Courses' (MOOCs) lack in knowledge about students' learning effectiveness. This requires the development of an evaluation system that can automatically track learners' progress, and constantly inform of what the students need for improvement. The first step towards that, is to explore how higher order thinking processes in online discussions can be analyzed and to what extent the process can be automatized.

Ways to assess students' thinking process demonstrated in the online discussions were investigated. Then, for a deeper understanding of the quality of student's learning, a new framework was constructed which helped in designed a coding schema for higher order thinking processes identification. The coding schema was then used to classify the data manually and further used for the automatization process. By teaching a machine learning how to search for higher order thinking indicators, a first attempt of the automatization of higher order thinking processes was made.

The results show that a Supervised Multiclass Classification Model can recognize the indicators of higher order thinking processes and classify the comments of students from the online discussions of a MOOC in three levels of thinking in proportion of 67%, and can make a distinction between lower and higher order thinking in proportion of 85% by using a coding schema designed specifically for the identification of higher order thinking in online discussions.

Table of Contents

Introduction

In today's fast changing world, the society is moving towards an increased value of the creation and management of knowledge as this ensures a competitive advantage for both individuals and organizations (Girard & Girard, 2015). Therefore, organizations' competitive advantage depends on providing individuals with the right opportunity to acquire knowledge quickly, easily and effectively (Koller, Harvey, & Magnotta, 2006). Massive Open Online Courses (MOOCs) gained the attention of learners, academic institutions (Admiraal, Huisman, & Pilli, 2015) and organizations all over the world in response to the need of quick, easy, and effective learning opportunities (Dodson, Kitburi, & Berge, 2015). MOOCs are courses designed with the intention to offer large-scale online education to anyone interested to participate (Sutton, 2013).

Besides individuals with personal interests in the offered subjects, MOOCs are also accessed by employees who need to enhance their knowledge in specific areas, prepare for the in-house courses, or develop role-specific skills (Dodson et al., 2015). With MOOCs, organizations have the possibility to provide their employees with learning opportunities (Beigi, Wang, & Shirmohammadi, 2015; Koller et al., 2006), through different types of instructions like video, text, and hyperlinks (Pursel, Zhang, Jablokow, Choi, & Velegol, 2016), and without time and place restrictions (Dodson et al., 2015). In order to adapt to each individual's learning needs, however, MOOCs need to deliver development opportunities that allow customization (personalization) of the learning process and content (Admiraal et al., 2015; Chapman, Goodman, Jawitz, & Deacon, 2016). This customization can be achieved by monitoring MOOCs effectiveness through formative assessments of the learning effectiveness of the students participating in it (Admiraal et al., 2015) and this in turn can be achieved by evaluating their learning processes (Chapman et al., 2016; Greller & Drachsler, 2012).

However, in traditional courses, the teacher could evaluate the learning effectiveness based on assessment forms like tests, but also based on his interaction with the students which provides him with insights about student's learning process. Based on the latter he could give immediate feedback to each participant and adapt his instructions when needed. However, due to the big number of participants in a MOOC and the less personal setting it is impossible for teachers to do the same (Admiraal et al., 2015; Capuano & Caballé, 2015; King, Goodson, & Rohani, 1998).

Current assessments of MOOCs, cannot actually help in customizing the courses according to each student's needs (Capuano & Caballé, 2015), and that could be explained by the idea that these assessments do not focus on students' learning effectiveness by formatively evaluating their learning processes in contrast with the traditional courses (Chapman et al., 2016). Furthermore, in some cases it is even questionable whether some assessments are "reliable" at all (Admiraal et al., 2015; Beigi et al., 2015; Capuano & Caballé, 2015). For example, some developers measure the effectiveness of MOOCs through completion rates (or dropout rates) (Beigi et al., 2015). However, some participants may have no intention to complete the course or earn a certificate when they start but instead only want to learn new skills or establish a basis of knowledge on a subject (Beigi et al., 2015). Thus, counting dropout rates without knowing and taking into account the personal objectives of students might lead to a distorted image of the MOOC's effectiveness (Pursel et al., 2016) and may not provide useful information to the course developers and students to improve the learning processes.

Other existing assessments of learners' progress are based on automated grading machines or peer assessments. Automated grading machines check if the participants responded well to the questions (Capuano & Caballé, 2015). This allows for massive assessment, but it is not suitable to provide a deep evaluation of the student's learning process (Capuano & Caballé, 2015). Peer assessments on the other hand, may help the learner to further develop (Admiraal et al., 2015), as they do provide remarks and feedback based on the interpretation of the learning process. However, with peer assessment different factors such as students' subjectivity, language barriers, culture and educational differences, their prior knowledge and understanding of the subject can influence the validity and accuracy of the peer assessment (Capuano & Caballé, 2015). In conclusion, because the current ways of assessment lack in accurate information about student's learning processes, they cannot allow for customization of the courses. The reason could be that in comparison with traditional learning environments, the nature of the learning environment in MOOCs hinders the assessment of students' learning processes (Chapman et al., 2016).

A solution could be to analyze students' interaction with and within a MOOC as this can give insights into their learning processes and this, in turn, can give a good indication of students' learning effectiveness (Chapman et al., 2016). However, due to the lack of face-to-face interactions in MOOCs the social interactions between students are made through online discussions (Awuor & Oboko, 2012). This makes online discussions a very important component

in the online learning environment as they reflect the social aspect of learning (Awuor & Oboko, 2012; Pursel et al., 2016). Tausczik and Pennebaker (2010) argued that the words individuals use in their social interactions reflect their internal thoughts. By responding to the online discussions' questions individuals can show to the community what they learned, share their experiences, and ask questions. Analyzing these discussions, therefore, can provide an understanding of how students think and learn and the dynamics of the MOOC (Khoshneshin, 2011; Pursel et al., 2016).

An important advantage of these online discussions is the fact that written recordings of the thinking processes can be later accessed by students for reflection and by the developers for further analysis (Meyer, 2004; Pursel et al., 2016). Therefore, analyzing students' interactions within a course can give insights into the quality of their thinking processes. Because thinking process can be an indicator for the level of their understanding and the quality of learning, by analyzing it, it can provide information about students learning processes (Moseley, Elliott, Gregson, & Higgins, 2005). This can therefore be an effective way to assess students' learning effectiveness and finally MOOC effectiveness. However, manually analyzing student's thinking processes that are present in the online discussions and providing each student with timely, personalized feedback, is not feasible in the setting of MOOCs due to the number of participants. Therefore, automatizing the assessment of learning effectiveness can help to customize the course content according to each individual's learning needs and provide them with feedback in a timely manner.

Research in how to analyze and interpret the quality of the thinking processes of students in online discussions of MOOCs is still scarce as this is quite a young field. Moreover, there is a lack of frameworks designed for the analysis of the quality of thinking processes in online discussions (Meyer, 2004). Therefore, this paper, aims to provide a framework that can be used to analyze the quality of thinking. Based on this framework, it will then be evaluated to what extent automatization of analyzing the thinking processes of students within online discussions is possible.

Theoretical Framework

As it was stated that we want to evaluate learning effectiveness through the learning processes, first, Mayer's model of cognitive processes will be presented further. Next, since the learning processes can be identified through the quality of thinking, this paper will then define the quality of thinking and present frameworks that can help to analyze the quality of thinking through thinking processes. After that, by using the information from these theories an understanding on how to identify (automatically) higher order thinking in online discussions of students will follow.

Literature Review

Learning processes.

Mayer (1996) proposed a model which presents three cognitive processes that result in meaningful learning. These processes entail 'Selecting', 'Organizing', and, 'Integrating' (SOI). "Selecting" is the first process and it is defined by him as: students focusing attention on relevant pieces of information from the course (Mayer, 1996). In the next two levels, students engage in higher quality of thinking by activating cognitive processes for understanding and by integrating past experiences into their learning process (Mayer, 1996). The second cognitive process in Mayer's (1996) framework is "Organizing" and it is defined as forming a coherent structure from the construction of internal connections between the selected information. Whereas the last cognitive process in Mayer's (1996) framework is "Integrating" and it is defined as relating the new knowledge to the existing information. In conclusion, this model explains that by being able to go through all three processes, learning will be meaningful, which means that students will be able to use the learned information within their own contexts (Mayer, 1996).

Learning processes alone, however, cannot be easily identified within MOOCs, because of the learning environment (as stated in the introduction). Therefore, this paper will evaluate the learning processes through the thinking processes of students, by assessing the quality of their thinking expressed in online discussions.

Thinking processes and Quality of thinking.

The disciplines philosophy and psychology made significant contributions to the conceptualization of thinking (Lewis & Smith, 1993). Each discipline, however, has a different perspective towards what defines the quality of thinking (Lewis & Smith, 1993). In short, from a

philosophical perspective the quality of thinking is represented by the ideal way of thinking, one using logical reasoning in order to decide on a course of action or what to believe (Lewis & Smith, 1993). Whereas from a psychological perspective, the quality of thinking is represented by a meaning making process that helps people understand their own experiences (Lewis & Smith, 1993).

As a result of the differences between the two perspectives, different types of thinking have been identified to give an indication on the quality of thinking. This in turn, caused confusion regarding the terminology to be used in defining the quality of thinking (Lewis & Smith, 1993). First, it seems that the most chosen terms are "higher order thinking" (King et al., 1998; Lewis & Smith, 1993) and "critical thinking" (Lai, 2011). At the same time, some researchers often interchange the terms critical thinking and higher order thinking , while others are referring to critical thinking as a form of higher order thinking (King et al., 1998). For example, critical thinking is defined as "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or conceptual considerations upon which that judgment is based" (Facione, 1990a, p. 3). While Ennis (1985), stated that critical thinking represents a significant part of the higher order thinking, in particular the practical side of it. He defined critical thinking as: "reflective and reasonable thinking that is focused on deciding what to believe or do" (Ennis, 1985, p. 45). The later definition is then introducing a new term that can be used as an indication for the quality of thinking - reflective thinking. Reflective thinking is, according to Dewey (1993) "a meaning-making process that moves a learner from one experience into the next with deeper understanding of its relationships with and connections to other experiences and ideas." (as cited in Carol, 2002, p 845). However, since Dewey (1933) stated that the process of thinking is as a consecutive sequence of ideas, which by the use of reflection, inquiry and critical thought, results in an evidence-based idea (as cited in King et al., 1998), it seems that reflective and critical thinking are incorporated in the thinking process. On top of that, King et al. (1998) stated that higher order thinking is represented by the following types of thinking: critical, logical, reflective, creative and metacognitive. Which suggests that higher order thinking is an encompassing term for the other types of thinking.

Therefore, this paper, will use the term higher order thinking to represent the high quality thinking and the quality of thinking will follow the definition of Lewis and Smith (1993). According to them, the quality of thinking should cover both earlier mentioned perspectives

(psychological and philosophical). Therefore, they chose the term higher order thinking to represent the quality of thinking and they defined it as follows: "higher order thinking occurs when a person takes new information and information stored in memory and interrelates and/or rearranges and extends this information to achieve a purpose or find possible answers in perplexing situations" (Lewis & Smith, 1993, p. 136).

Existing frameworks.

Different frameworks have been used in the past (as assessment tools) to analyze the quality of thinking. Some of these tools specifically state that they analyze higher order thinking, though others used different terms to refer to the quality of thinking. Some of the frameworks were designed for the analysis of thinking processes in online discussions, some were designed for different contexts of use. Also, differences were made in the way the frameworks present the quality of thinking. Some present it by making the distinction between high quality and low quality of thinking, some present different processes of thinking, or types of thinking. In order to identify higher order thinking in the online discussions of students in MOOCs, a closer look has to be taken at each of these frameworks to see how each of these frameworks propose to analyze the quality of thinking.

The first framework which is presented here is the revised taxonomy of Bloom (Krathwohl, 2002).This framework has been widely used to evaluate student's thinking, although the framework is not designed with the specific purpose of analyzing thinking, nor identifying higher order thinking in online discussion. Its original purpose was to serve teachers in designing educational objectives in the classroom (as cited in Meyer, 2004). This framework has six hierarchical cognitive levels based on complexity. In this case, the higher order thinking, which entails analyzing, evaluating, and creating, is separated in this framework from lower order thinking, which encompasses remembering, understanding, and applying (King et al., 1998; Krathwohl, 2002). Furthermore, each higher level of thinking builds on the antecedent levels. Therefore, higher order thinking is grounded in the lower order thinking applications (Bailin, Case, Coombs, & Daniels, 1999; King et al., 1998; Krathwohl, 2002). The specific details of the indicators of higher order thinking from this framework will be presented in Table 1.

The second framework presented here is the framework of Garrison, Anderson, and Archer (2001). As presented in the paper of Meyer (2004), this framework was designed for the analysis of cognitive processes present in online discussions. The quality of thinking here is represented

by critical thinking skills. These skills are part of a process composed of 4 phases named Triggering, Exploration, Integration and Solution. These phases can as well be used as indicators to determine the quality of thinking. The in-depth explanation of each phase can be found in Table 1.

Fourth, Marland, Patching, and Putt (1992) analyzed students' thought processes in the context of distance education which resulted in a classification tool composing of six categories of thinking processes in which students were found to engage in during their studies. The categories of thinking are: Evaluation, Linking, Strategy planning, Generating, Metacognition and Affective. The specifications of these six categories can be found in Table 1.

Last, Herrington and Oliver (1999) developed an instrument for the classification of student's talk based on Resnick' (1987) nine characteristics of higher order thinking (as cited in Herrington & Oliver, 1999). According to these characteristics, higher order thinking is non-algorithmic and complex, offers multiple solutions and applications of multiple criteria, involves judgement and interpretation, uncertainties, self-regulation and effort, and it gives meaning and structure (Resnick, 1987, as cited in Herrington & Oliver, 1999). The instrument differentiates between lower order thinking and higher order thinking through different indicators representing each category. The indicators of higher order thinking composing the instrument of Herrington and Oliver (1999) are: Uncertainty, Deciding on a path of action, Judgement and interpretation, Multiple perspectives, Imposing meaning, effortful thinking and multiple solutions, and Self-regulation of thinking. See Table 1 for further elaboration on each indicator.

Table 1
Frameworks of higher order thinking

Frameworks			
Bloom	Garrison	Marland	Herrington
Remember "recall information without engaging in a cognitive process of understanding"	**Triggering** "the correct identification of the problem that is discussed, students having a "sense of puzzlement" towards the subject"	**Evaluation** "a mental process that indicates judgement towards concepts."	**Uncertainty** "asking questions and clarifications"
Understand **"**constructing meaning of the new information"	**Exploration** "presenting multiple ideas (brainstorming) in one message"	**Linking** "synthesizing or connecting concepts, experiences and ideas"	**Deciding on a path of action** "planning what needs to be done"
Apply **"**making use of the information in a new situation**"**	**Integration** "connecting ideas and synthesizing information and constructing meaning from the ideas generated in the Exploration phase"	**Strategy planning** "planning study materials."	**Judgement and interpretation** "defending an issue or opinion, making connections and giving definitions."
Analyze "understanding the structure of something, making inferences, searching for evidence and explanations"	**Resolution** "defending the solutions found or giving argumentation and reasoning based on real world experiences"	**Generating** "mental processes like reasoning, making predictions, or elaborating"	**Multiple perspectives** "seeing both parts of an issue, challenging different ideas, and giving alternatives."
Evaluate "judging the new information by comparing it with information from past experiences"		**Metacognition** "being aware of thinking processes and self-directing thinking through reflections or evaluations"	**Imposing meaning, effortful thinking and multiple solutions** "synthesizing information, giving conclusions, presenting believes and alternative solutions."
Create: "combines ideas from prior knowledge to form new ideas or products into a new structure or product"		**Affective** "awareness towards own feelings and towards the learning process."	**Self-regulation of thinking** "awareness of their own thinking processes and understandings."

In conclusion, there is a big amount of research on the nature of thinking that produced frameworks to define and classify higher order thinking (Herrington & Oliver, 1999). However, not enough frameworks are designed specifically for the analysis of higher order thinking processes of students in online discussions (Meyer, 2004) nor for an automatized analysis of higher order thinking. Moreover, deciding on which framework fits best with this study could be challenging because as each approach has its own convincing rationale Newmann (1990) (as cited in Herrington & Oliver, 1999). Therefore, this paper aimed to explore how higher order thinking can be identified in a way that can be used to evaluate online discussions in a MOOC and the research question deriving from that, will be presented further**.**

Research Questions

The first research question that this study will try to answer is:

> **RQ1. How can higher order thinking processes be identified in the online discussions of a MOOC?**

The outcome of this research question can then be used to try to automatically identify the quality of student's thinking through their responses in online discussions in MOOCs. The following research question is therefore employed:

> **RQ2. To what extent can the identification of higher order thinking processes in online discussions of a MOOC be automatized?**

This research question will be answered through answering two sub questions. First, we are interested if a computer is able to make a distinction between Lower and Higher order thinking in online discussions within a MOOC. Next, we are interested if a computer can identify multiple levels of thinking. The resulting sub questions are:

> **SQ1: To what extent can a computer identify multiple levels of thinking from the higher order thinking process in online discussions within a MOOC?**
> **SQ 2: To what extent can a computer make a distinction between lower and higher order thinking in online discussions within a MOOC?**

Methodology

Research Design

As this research aimed to explore a relatively new field of investigation which has been very little explored until now, the research design used was an exploratory design (Greller & Drachsler, 2012). With this, it is strived to gain insight into a relatively new field. The approach used was a mixed method approach in which both qualitative analysis is used to identify higher order thinking processes in the online discussions between participants in a MOOC and quantitative analysis to see to what extent this analysis can be automatized.

Respondents

The respondents of this research were teachers participating in the MOOC called "Growth mindsets" offered by a world-wide educational provider specialized in teachers' learning and development. The data was automatically gathered during the course by asking the participants questions related to the course content as part of the course' activities. The participants voluntarily posted comments in response to the questions from the course, therefore, this research used an existing dataset. There is no exact information about the respondents as the data about them was anonymous.

Method

In order to automatically identify higher order thinking levels in online discussion of MOOCs the choice was made to analyze the comments from these discussions through text mining. Text mining is an interdisciplinary field which uses techniques from different fields like machine learning, natural language processing, information retrieval and statistics with the purpose of automatically extracting information from unstructured text documents (Gupta & Lehal, 2009). A computer uses pattern recognition methods to find interesting information in large databases, which otherwise would require a lot of work and time spent by a human to manually process the text (Gupta & Lehal, 2009). The text mining methods are very promising in the context of online learning, as there is a big amount of data which could be explored and used for purposes like evaluation of learning effectiveness.

Text classification is an approach in the field of text mining that could be used in assigning textual objects from a total data set to two or more classes (He, 2013). According to He (2013), this approach is mainly divided into two categories: Supervised Text Classification (STC) and

Unsupervised Text Classification (UTC). UTC is used in the case when the data has not yet been classified and the labels are obtained only by finding patterns in the dataset (e.g. clustering) (He, 2013). Whereas STC is a method that can be very useful in the analysis of the online discussions, as data should first be labeled according to rules and then used as training data for the computer to learn how to classify (Awuor & Oboko, 2012). In the case of the present study, the rules would be made based on the findings of the literature review on higher order thinking identification.

STC involves feature (keywords) extraction (the terms keywords and feature will be interchanged as they refer to the same thing) and machine learning techniques to program computers in how to classify the text based on training data (Gupta & Lehal, 2009). Feature extraction is a procedure in which textual units are transformed into structured data features and labeled with one or more classes (He, 2013). Based on the labeled data, the most relevant and informative features are selected to contribute to the performance of the model by using statistical methods (e.g. chi-square selection algorithm) (He, 2013). Machine learning algorithms are then used to allocate the text documents to specific classes (Wiegersma, et all, 2017). Examples of such machine learning algorithms are: decision trees, naïve Bayes, support vector machines and the *K* nearest neighbor model (Wiegersma, et al, 2017) . These applications result in different models which are then evaluated for their performance and finally the best model is selected (He, 2013). The model can be further applied on unclassified data having the goal to gain new insight about the data or to make predictions (Awuor & Oboko, 2012). Therefore, the STC could be used to classify the data by first manually analyzing data and using the outcomes of the manual analysis to train a model.

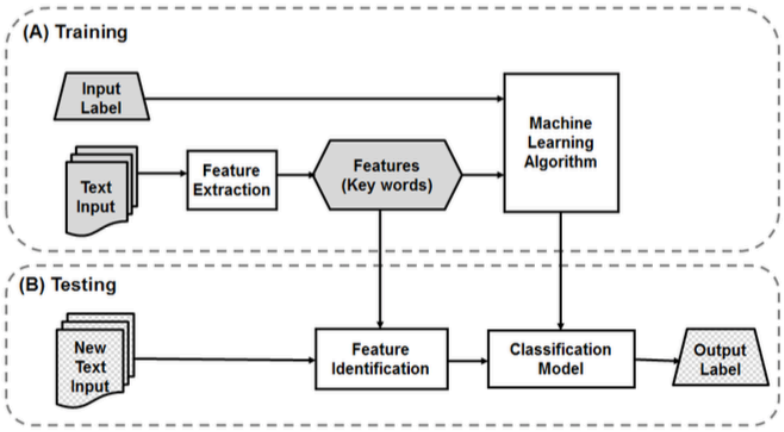

Figure 1. Framework of Supervised Text Classification as presented in He (2013)

Procedure

Development of a coding schema.

In order to analyze the comments through STC, a coding schema which presents the identification rules of higher order thinking was needed to manually classify the data. The development of this coding schema was then based on the theoretical framework. In order to create a coding schema, a literature review was conducted into different frameworks of higher order thinking, initially developed for varying contexts. This literature review had the objective to explore the existing theories on thinking and frameworks for analysis. However, choosing one of the frameworks presented in the literature review would not have been enough because they are not designed specifically for the analysis of higher order thinking processes of students in online discussions. Therefore, a compilation of the common elements from the existing higher order thinking frameworks was thought to be a better solution (Newmann, 1990, as cited in Herrington & Oliver, 1999). Therefore, in understanding how higher order thinking can be identified in the online discussions of students a decision was made to develop a coding schema and test it. As it was stated above, the development of the coding schema is grounded in the theories from the literature review. Therefore, its construction was based on the rationale that in MOOCs, students' learning processes are represented by their thought processes, which in turn, are expressed in the online discussions. Therefore, student's learning effectiveness could be evaluated by looking at the extent to which they engage in the learning processes, and by checking for the quality of their thinking processes during the course. It was desirable, therefore, to map the higher order thinking process (from the chosen definition) to the learning processes of the SOI model (as this model results in meaningful learning) and then add the indicators from the presented frameworks for analysis of higher order thinking. This mapping is explained in the Instrumentation under the coding schema development section.

Supervised text classification.

The description of the STC model development process is made according to Wiegersma, et al (2017). The model development process usually comprises of two main stages: model selection and model evaluation. First, a validation strategy is needed in order to split the data into separate samples to be used in each of the stages of the process, hence generalization problems are avoided. Validation strategies commonly used are holdout validation and K-fold cross validation.

Next, a text classification pipeline is used to select the model which has the best set of parameters. The main phases of the classification pipeline are training and testing (prediction). Both phases involve 3 steps: preprocessing, feature extraction and feature selection. The data needs again to be split in order to avoid generalization problems. Thus, different sets of data are used for training and testing, as the model needs to be tested on data which was not used in training (unseen data).

The process of training including the 3 steps is presented further. First, preprocessing strategies are commonly used to improve the efficiency of the training (or testing). This may involve tokenization – splitting documents of text in paragraphs, sentences and words, or normalization - removing punctuation marks and capitals or stemming words. Second, feature extraction is a procedure in which the most informative keywords for each class are extracted from the already labeled data (supervised). In this phase, different text representation procedures can be used depending on the context (e.g. language model representations such as N-(multi)grams for phrases, the bag-of-words model for single words and/or the linguistic variables like number of words, sentence length) (for more information see (Wiegersma, et al, 2017). Third, feature selection has the objective to find, by using statistical methods such as Chi-squared test, the optimal subset of keywords that contribute maximally to the efficiency of the model. Then the input for the machine learning algorithm is a set of pre-specified keywords and the accompanied classes. Thus, the computer learns how the keywords are related to the classes.

The process of testing starts first by preprocessing the data and it is followed by a scanning procedure in which the computer is searching for the keywords selected in the training phase in the new documents. The testing has the objective of checking how well the model works on a new dataset. Next, the systematically recognized keywords are selected and used as input for the trained model, which in turn predicts the most likely label for each unlabeled document.

Finally, the different models are compared for their performance and the one which has the best combination of parameters is selected as the final model. Next, for the model evaluation stage, the final model is once more trained on the full amount of data (model selection - data sample). Then the model is assessed for its performance on new, unclassified data (the evaluation - data sample).

Preprocessing the Data

Before any analysis, the text was processed using data cleaning methods. For example, data containing comments uploaded by the teacher of the course were deleted from the dataset, as it did not serve the purposes of the analysis. Also, comments posted by the developers of the course, which had the purpose to test if the platform works, were deleted from the dataset. Then, the text was normalized by removing punctuation and spaces, and the text was combined into paragraphs. Next, the format of the data was changed in order to fit with the requirements of the script for automatization. In this way, each comment became a separate document.

The Manual Analysis Procedure

Then, a sample data was analyzed with the developed coding schema (835 documents out of 19.633 total number of documents), by assigning manually the comments indicating the processes of different levels of thinking, to the representative category. A second coder was also instructed in how to use the coding schema and then used it to analyze a sample (100 documents) from the data. The results were compared, then a discussion took place and the final results had a very good interrater reliability score of 0.953, therefore the coding schema was validated by a second coder. The outcome of the manual analysis shows how higher order thinking processes can be identified in the online discussions and the classified data was further used for the development of the supervised classification model. After the final STC multiclass model was developed and used to classify the whole dataset, a manual final check on 100 comments was done.

The Automatized Analysis Procedure

The development data was split into different samples to serve the purposes of each stage of the model development process. The procedure of developing the model had two main stages: model selection and model evaluation, for which the first step was to choose a validation technique. Then the supervised classification pipeline follows, by training a range of different models and testing them for performance on a different subset. In training the model, the labeled documents were used as labeled features sets. The same feature sets were then used for prediction on the unlabeled data. The model selection compared the different models and selected the one with the best combination of parameters in a grid-search. The model was then trained again on the full development data set. This resulted in a final model which was analyzed for its performance on the rest of the data.

The performance of the model was calculated based on the difference between the true and the predicted class labels. Therefore, the outcomes show to what extent the analysis could be automatized using a STC tool. The whole procedure was done twice for the 3-class task and twice for the binary classification, first with a smaller dataset (835 documents), then more data was analyzed manually and added to the automatization process (879 added documents). Results were then analyzed and as the outcomes from the second round provided a new (better) 3-class classification model which then were taken as the final model for the 3-class classification and it was used to classify the whole dataset (19.633 total number of documents). Then, the final results and conclusions were then presented. The same procedure was done for the binary classification, only once with the full amount of development data (1714 documents).

Instrumentation

Manual Analysis Tool

Framework design.

Mapping the higher order thinking process and the indicators to the SOI model

As discussed in the beginning of the Theoretical Framework, three learning processes are presented in the SOI model. For each of these learning processes, indicators for the quality of thinking were found in the earlier presented frameworks (see Table 1) and three levels of thinking were found in the definition of the higher order thinking process. According to the definition, higher order thinking seems to occur as a process which contains three levels of thinking: (1) taking new information, (2) interrelating and rearranging this information, and (3) combining new information with existing information and extending it to find possible answers. These processes seemed to relate to the processes of the SOI model and therefore a mapping of higher order thinking to the learning processes has been made which consists of three levels. This mapping will be discussed in the following section.

Level 1.

Selecting, the first learning process, can be related to the first level of thinking which, according to the definition of higher order thinking, is "taking new information". Both of them suggest that the student engages in a thinking process that enable him/her to capture new information. Indicators for this thinking process were found also in the frameworks for analysis. For example, within Bloom's framework, the first level of thinking (Remembering) is defined as recalling relevant information without engaging in a cognitive process of understanding, this also fits with indicators of the framework from Garrison. For detailed information see Table 2.

Level 2.

The second learning process (Organizing) can be related to a second level of thinking as the definition of the higher order thinking process suggests that students would "interrelate and/or rearrange the new information". Which is in accordance with the idea of connecting information into a coherent structure from the process of learning. In order for people to be able to interrelate and rearrange information, they first need to be able to take information. So, level 1 thinking is needed for level 2 thinking, which is in accordance with Bloom's taxonomy and other literature on thinking processes (Bailin et al., 1999; King et al., 1998; Krathwohl, 2002). Indicators from different frameworks were also found to fit this description. Garrison's indicator of higher order thinking (Integration) refers to a meaning making process in which information is connected and synthesized. And Bloom's indicator, Analyze, suggests that students would enter

the process of higher order thinking by understanding the information, inferencing and explaining different concepts. In addition, it seems that the second level of thinking is represented by a critical discourse. Lai (2011) stated that critical thinking includes skills like analyzing arguments, inferring using inductive and deductive reasoning, judging and evaluating. These definitions can be linked to the indicators from the frameworks of Bloom, Garrison, Marland and Herrington. See Table 2.

Level 3.

The SOI model suggests that by going through all three learning processes, learning will become meaningful, which in turn means that students will be able to use the learned information within their own contexts (Mayer, 1996). In order to use the learned information within their own contexts, students should be able to integrate (the third learning process) their personal experiences into their learning process. Therefore, the student would engage in meaningful learning by relating new knowledge to the existing information gathered through past experiences. The third learning process (Integrating) is reflected in the definition of the higher order thinking process by the idea that the "information stored in the memory" is connected to "new information" and extended with the purpose to find answers or create solutions. Combining these two results in the following definition for level 3 thinking: "Extend the use of the new information to existing knowledge or past experiences to achieve a purpose or find possible answers".

Level 3 of thinking can be related through similar indicators on the quality of thinking from the frameworks of Bloom, Garrison, Marland and Herrington. For example, Bloom's indicators, Evaluate and Create imply that one is integrating information from past experiences with new information, to support new ideas with evidence or create products. However, as this study wants to look into student's online discussions, the products would take the form of ideas which are presented in the answers to the online discussions' questions. See Table 2 for the full mapping of all frameworks to this level.

As with level 2 thinking, level 3 thinking requires the previous levels as well. In order to be able to combine new with old information, one first needs to take the new information, understand the information and critically discuss it. This idea is well captured in the following statement: information and memory work as "a refrigerator in which to store a stock of meanings for future use," while judgment "selects and adopts the one to be used in an emergency. . ." (Dewey, 1933, p. 125 as cited in King et al., 1998). Through levels 1 and 2, people end up with a

cognitive model of new information that can be stored and then used, in level 3, in combination with existing knowledge or past experiences.

Coding schema design.

Adding Keywords.

Additionally, as the ultimate goal in the current study was to automatize the identification of the higher order thinking processes in online discussions and this was made by using a Supervised Text Classification, different dispositions of the words humans use was needed to be taken into account in the development of the coding schema. For choosing keywords for each category of thinking, the theory of Tausczik and Pennebaker (2010) was used. They argued that the way people think, how they process information and interpret it in order to make sense of their experiences is reflected by the words they use to connect thoughts. Therefore, cognitive complexity is reflected in people's reasoning. Their reasoning is composed of two processes represented by exclusive words (but, without, exclude, etc.) and conjunctions (and, also, although, etc.). Prepositions and cognitive mechanisms (cause, know, ought, etc.) indicate language complexity. Whereas the use of causal words (because, effect, hence, etc.) represent cognitive mechanisms used for giving explanation. Moreover, Davis and Brock (1975) state that a person who is self-aware is inclined to use more first person pronouns. For more information about how words usage is connected to people's thinking, see the paper of Tausczik and Pennebaker (2010). More specifically, no keywords were assigned to the first category, first-person pronouns were only dedicated to the third level of thinking, whereas the second and the third level share the keywords for language complexity. In addition, a rule referring to the length of the comments was added, expecting the comments to show an ascending length based on the quality of thinking.

In conclusion, based on the new framework, the coding schema is comprising of the learning process (SOI model), the levels of thinking, the indicators from the other frameworks and keywords (see Table 2).

Table 2
Coding Schema

Mayer's SOI model	***Levels of the higher order thinking process***	*Bloom*	*Garrison*	*Marland*	*Herrington*	***Keywords and Other rules***
Selecting "focusing attention on relevant pieces of information"	Level 1 "Taking new information"	**Remember** "recall relevant information without engaging in a cognitive process of understanding"	**Triggering** "the correct identification of the problem that is discussed, students having a "sense of puzzlement" towards the subject"	n.a.	n.a.	- Short length of the comment - No KW from L2 and L3
Organizing "forming a coherent structure from the construction of internal connections between the selected information"	Level 2 "interrelate and/or rearrange the new information"	**Understand** **"**constructing meaning of the new information" **Analyze** "understanding the structure of something, making inferences, searching for evidence and explanations"	I**ntegration** "connecting ideas and synthesizing information and constructing meaning	**Evaluation** "judgement towards concepts" **Linking** "synthesizing or connecting concepts, experiences and ideas" **Generating** "reasoning, making predictions, or elaborating"	**Judgement & interpretation** "defending an issue or opinion, making connections and giving definitions" **Multiple perspectives** "seeing both parts of an issue, challenging different ideas, and giving alternatives" **Imposing meaning** "synthesizing information, giving conclusions, presenting believes and alternative solutions"	**-** Medium – long length of the comment - because - however - if – then - so - hence - as - though - whereas - on one hand – on the other hand - whereby - as long as - unless - effect - cause - know - ought - in order to - rather than
Integrating "relating the new knowledge to the existing information"	Level 3 "Extend the use of the new information to existing knowledge or past experiences to achieve a purpose or find possible answers"	**Evaluate** "judging the new information by comparing it with information from past experiences" **Create** "combines ideas from prior knowledge to form new ideas or products into a new structure or product"	**Resolution** "defending the solutions found or giving argumentation and reasoning based on real world experiences"	**Metacognition** "aware of their thinking processes and self-directing their thinking through reflections or evaluations"	**Self-regulation of thinking** "awareness of their own thinking processes and understandings"	- Long length of the comment - past tense - KW from L2 + - I - My - Experience

The Supervised Text Classification Tool

The script for supervised text classification (Wiegersma, et al, 2017) is written in Python and designed for its use in the classification of text data from different psychological contexts. Therefore, it was used in this study in order to investigate to what extent an automatization of the text analysis is possible.

The validation strategy used in model selection was nested 5-fold cross validation, using a 5-fold cross validated grid-search for model selection (inner loop) and 5-fold cross validation for model evaluation (outer loop).

The classification pipeline uses the following text processing elements and a machine learning algorithm. Preprocessing the data was done by using tokenization and normalization strategies. The feature extraction steps were: removing "stop words" (e.g. I, to), document representation through N-(multi)grams and the bag-of-words model. Additionally, the terms were weighted using the vectorization strategies term frequency (tfi_j) and term frequency-inverse document frequency ($tfidf_{ij}$) (for more information see Wiegersma, Van Noije, Sools, & Veldkamp, 2017).

For feature selection, the filter method was used to score each feature independently and then rank the most informative features. Pearson's chi-squared statistical test was used (X^2) as a metric for ranking the features and compare the difference between the observed and expected occurrences of the features in the three classes. Next, for the machine learning, the Support Vector Machine algorithm was used to predict the class labels.

Results

Results of the First Research Question

The first research question that this study aimed to investigate was: How can higher order thinking processes be identified in the online discussions of a MOOC? In order to respond to this question, we first wanted to identify and classify higher order thinking.

Therefore, the first step was to conduct a literature review on the theories of learning, thinking, and existing frameworks to assess the quality of thinking. The results from this literature review were combined into a framework on higher order thinking which is comprised of the process of learning (SOI), levels of thinking with their explanations (according to the definition of higher order thinking), and indicators of higher order thinking from other frameworks. The next step was to explore how the resulting three levels of higher order thinking within this framework could be identified within online discussions. Therefore, keywords related to language usage theories were added to the different levels within the constructed framework resulting in a coding schema that can be used for analysis.

By using this coding schema, the responses of students to the online discussions' questions can be evaluated from multiple perspectives (learning process, thinking process, levels of higher order thinking, etc). In order to identify the higher order thinking process, a researcher can manually classify student's responses into level one, two or three of thinking by looking for patterns in the data according to the rules of the coding schema. For example, when a student, at a specific moment during the course, does not use any keywords which are representative for causation, the thinking at that moment can be classified as the first level of thinking. Lack of causation, namely, implies that the student is not giving an explanation/argumentation which is needed for both level 2 and level 3 thinking. After this, the categorization can be checked by evaluating whether the comment is in accordance with the definition of the first level of thinking. In the case that the coder is not sure about the category in which the comment should be classified, (s)he can compare the comment with the indicators form the other frameworks and see if a match can be found, which can clarify uncertainty based on more in-depth explanations from multiple sources.

Results of the Second Research Question

Next, in order to find an answer to the second research question: To what extent can the identification of higher order thinking processes in online discussions of a MOOC be automatized, the following sub-questions needed to be answered: SQ 1) To what extent can a computer identify multiple levels of thinking from the Higher order thinking process in online discussions within a MOOC? and SQ2) To what extent can a computer make a distinction between Lower and Higher order thinking in online discussions within a MOOC?

The first sub-question.

First multiclass classification.

The total amount of development data was composed of 835 text documents (out of 19.633 total number of documents). The model selection for the 3-class is done based on a 5-fold cross-validated grid-search, alternately using four folds for training and one-fold for validation. The grid search is guided by the weighted F_1 metric. The parameter combination that generated the highest cross-validated weighted F_1 score is shown in the Table 7. The best results on the validation set were generated by the Linear Support Vector Classifier with a weighted F_1= 0.584 and the penalty factor C= 1. Removing stop words did not result in a higher mean cross-validated weighted F_1 score. Additionally, the grid search showed that documents could be best represented by N-multigrams ranging from one to three words, using term frequency to weight the terms.

Table 7

Best parameter Values for the Final Multiclass Classification Model (first and second round)

	Best parameter value	
Parameter	First multiclass model	Second multiclass model
Remove stop words	No	No
Minimal x documents	3	3
Representation schemes	N-multigrams range (1, 3)	N-multigrams range (1, 3)
Term weights	tf_{ij}	$tfidf_{ij}$
Select k best features	390	210
Classifier	SVC	SVC
Regularization parameter C	1	1
Class weights	balanced	balanced

The model comprises of only 390 best features, the 50 most informative features for each class have high x^2 values, with significant differences in occurrences between the 3 classes. The selected model was then evaluated in the outer 5-fold cross-validation loop, alternately using four folds as development set and the one remaining fold as test set. The Confusion Matrix (Table 8) shows the results per class. The model predicted well 20 comments out of 44 for the first category, 23 out of 40 for the second category and 41 out of 49 for the third category.

Table 8
Confusion Matrix Final Multiclass Classification Model (first round)

	Predicted thinking level		
True thinking level	First Level of thinking	Second Level of thinking	Third level of thinking
First level of thinking	**20**	6	18
Second level of thinking	8	**23**	9
Third level of thinking	4	4	**41**

The best obtained model classifies the data moderately well having a weighted F_1 score of 0.622 with an accuracy of 0.632, a weighted recall of 0.632 and weighted precision of 0.639 (Table 9).

Table 9
Performance Scores Final Multiclass Classification Model (first round)

	Precision	Recall	F_1	Accuracy	N documents in test set
First level of thinking	0.62	0.45	0.53		44
Second level of thinking	0.70	0.57	0.63		40
Third level of thinking	0.60	0.84	0.70		49
Weighted average total	0.64	0.63	0.62	0.63	133

After these results, it was decided to manually classify more data and increase the input data for the development of the model.

Second multiclass classification (after adding more data).

The total amount of development data in this round is composed of 1714 documents (out of 19.633 total number of documents). The model selection for the 3-class is done based on a 5-fold cross-validated grid-search, alternately using four folds for training and one-fold for validation. The grid search is guided by the weighted F_1 metric. The parameter combination that generated the highest cross-validated weighted F_1 score is shown in the Table 7. The best results on the validation set were generated by the Linear Support Vector Classifier with a weighted F_1= 0.643 and the penalty factor C= 1. Removing stop words did not result in a higher mean cross-validated weighted F_1 score. Additionally, the grid search showed that documents could be best represented by N-multigrams ranging from one to three words, using (tfidf) term frequency and inverse document frequency to weight the terms.

The model comprises of only 210 best features, the 50 most informative features for each class are shown in the Table 10. The x^2 values are high, starting from 6.4391 with significant differences in occurrences between the 3 classes and in accordance with the coding schema from the manual analysis.

Table 10

The Most Informative Features per Class for the Final Multiclass Model (second round)

Feature	X^2 value	Feature counts		
		First level of thinking	Second level of thinking	Third level of thinking
Resource_name	6.4391	21	5	2
I	5.9094	241	232	715
Resourc	5.8947	25	5	15
I have	5.5448	11	9	78
Toolkit	5.4321	16	1	1
Because	5.1159	5	66	132
Challeng toolkit	4.9471	15	1	1
My	4.6188	59	32	167
The challeng toolkit	4.4493	12	0	1
Blog	4.3560	12	1	0
As	4.3423	42	167	256
Success learner	4.0271	34	16	1
we	3.9475	46	43	175
Resource_name speech	3.9198	7	0	0
me	3.7224	27	22	104
you	3.7064	22	98	62
it	3.3832	79	218	313
was	3.3510	15	7	70
thank	3.3016	11	1	7
Resource_name	3.2932	13	5	2
Resource_name	3.2932	13	5	2
To solv	3.2801	0	11	0
They	3.2588	139	349	411
Student	3.1768	51	130	241

Like the	3.1310	32	14	12
Wrong	3.1083	4	14	52
Them	3.0984	61	71	210
So	3.0583	14	34	95
Speech	3.0464	8	1	1
If	2.9673	13	77	94
Essay	2.9069	0	0	13
Mindset	2.8346	9	36	88
Help me	2.6954	11	4	37
Way to help	2.6875	5	0	1
At	2.6833	19	48	119
She	2.6808	0	2	22
That	2.6589	104	198	374
Solv	2.6471	4	25	7
The blog	2.6080	5	0	0
Use resourc	2.5743	3	0	0
Question and	2.5584	9	0	1
Howev	2.5567	1	34	36
As I	2.5563	0	4	19
Success learner are	2.5490	15	4	0
Amaz	2.5451	3	0	0
Lo	2.5089	22	40	13
The articl	2.5041	6	0	0
Some student	2.4880	0	24	19
Time	2.4852	16	49	99
When I	2.4774	3	1	20

Table 11 shows the Confusion Matrix for the final model based on 249 documents (20%) from 1714 documents (the total amount of training data). The diagonal cells show that the model predicted the correct class labels for 168 documents, leading to an accuracy of 0.68. More specific, the multiclass classifier predicted correctly 61 out of 91 documents for the first level of thinking, 50 out of 85 for the second level of thinking and 57 out of 73 for the third level of thinking.

Table 11
Confusion Matrix Final Multiclass Classification Model (second round)

	Predicted thinking level		
True thinking level	First Level of thinking	Second Level of thinking	Third level of thinking
First level of thinking	**61**	12	18
Second level of thinking	23	**50**	12
Third level of thinking	7	9	**57**

The best model classifies the data well having a weighted F_1 score of 0.673 with a weighted recall of 0.67 and weighted precision of 0.68 (Table 12).

Table 12
Performance Scores Final Multiclass Classification Model (second round)

	Precision	Recall	F_1	Accuracy	N documents in test set
First level of thinking	0.67	0.67	0.67		91
Second level of thinking	0.70	0.59	0.64		85
Third level of thinking	0.66	0.78	0.71		73
Weighted average total	0.68	0.67	0.67	0.68	249

Comparing with the previous dataset analysis it seems that adding more data improved the model (first round 835 total documents in the training dataset vs. 1714 total documents in the training dataset second round). As the second model results are better, it is taken as the final model for the 3-class classification. The model is then used for classifying the whole data set (19.633 documents) and manually evaluated for performance on a sample of 113 documents. The results show that the model predicted well the thinking levels with 29.2% wrongly classified data. In conclusion, the tool generates very stable results.

The second sub-question.

Binary classification.

The total amount of development data was composed of 1714 text documents (out of 19.633 total number of documents). The model selection for the binary is done based on a 5-fold cross-validated grid-search, alternately using four folds for training and one-fold for validation. The grid search is guided by the F_1 metric. The parameter combination that generated the highest cross-validated F_1 score is shown in the Table 3. The best results on the validation set were generated by the Linear Support Vector Classifier with a F_1= 0.842 and the penalty factor C= 10. Removing stop words did not result in a higher mean cross-validated F_1 score. Additionally, the grid search showed that documents could be best represented by N-multigrams ranging from one to three words, using $tfidf_{ij}$ to weight the terms.

Table 3

Best parameter Values for the Final Binary Classification Model

Parameter	Best parameter value
Remove stop words	No
Minimal x documents	1
Representation schemes	N-multigrams range (1, 3)
Term weights	$tfidf_{ij}$
Select k best features	'all'
Classifier	SVC
Regularization parameter C	10
Class weights	balanced

The model finds 'all' features as informative features, the 50 most informative features for each class are shown in Table 4. These features have high x^2 values, with significant differences in occurrences between the 2 classes.

Table 4

The Most Informative Features per Class for the Final Binary Model

		Feature counts	
Feature	X^2 value	Lower order thinking	Higher order thinking
Resourc	6.8704	27	13
It	5.0980	78	537
Because	4.7184	5	176
As	4.7116	48	382
They	3.9374	162	740
If	3.5157	14	172
Resource_name	3.4992	18	8
To	3.4750	474	1777
Success learner	3.2488	38	16
Thank	3.2108	14	9
That	3.1824	113	575
Good resourc	2.9227	6	0
blog	2.8785	15	2
Think	2.8283	38	270
Toolkit	2.7193	14	4
more	2.6577	31	237
this	2.5430	60	361
I think	2.3994	13	139
mindset	2.3852	9	138
be	2.2822	63	328
have	2.2181	68	346
Resource_name	2.1321	12	7
Resource_name	2.1321	12	7
Challeng toolkit	2.0766	12	4
I	2.0593	223	905
would	2.0001	15	135
howev	1.8833	2	65
yoga	1.8621	1	0
The challeng tookit	1.8466	10	2
student	1.8230	58	322
a	1.8158	230	905
Success learner are	1.8092	17	3
do	1.7932	31	211
inspir	1.7882	13	10
so	1.7692	15	122
Resource_name speech	1.7463	5	0
question	1.7449	47	62
make	1.7423	38	221
It is	1.7363	10	112
Can	1.7132	60	295
Like the	1.6944	25	24
interest	1.6708	18	18
are	1.6424	106	445
video	1.6216	11	8
If they	1.5770	2	54
Thank you	1.5461	5	3
will	1.5144	61	282
feel	1.5011	5	71
at	1.4954	21	152
I like the	1.4808	15	12

The selected model was then evaluated in the outer 5-fold cross-validation loop, alternately using four folds as development set and the one remaining fold as test set. The Confusion Matrix (Table 5) shows the results per class. The model predicted well 35 comments out of 81 for the first category, and 158 out of 168 for the second category.

Table 5

Confusion Matrix Final Binary Classification Model

	Predicted thinking level	
True thinking level	First Level of thinking	Second Level of thinking
First level of thinking	**35**	46
Second level of thinking	10	**158**

The best obtained model classifies the data very well, having a F_1 score of 0.849 with an accuracy of 0.775, a recall of 0.78 and precision of 0.78 (Table 6).

Table 6

Performance Scores Final Binary Classification Model

	Precision	Recall	F_1	Accuracy	N documents in test set
First level of thinking	0.78	0.43	0.56		81
Second level of thinking	0.77	0.94	0.85		168
average total	0.78	0.78	0.75	0.78	249

Discussion

In the following section the results based on the research questions will be discussed, conclusions will be made and the implications of the results will be presented.

The First Research Question

The answer to the first question in embodied in a coding schema which was used to manually classify the comments of students participating in the online discussions of a MOOC. In this way, higher order thinking processes were identified under the second category and the third category from the coding schema.

The outcome of this research question was then used to automatically identify the quality of student's thinking through their responses in online discussions in MOOCs.

The Second Research Question

The second research question was answered by answering its sub-questions. This will be discussed in detail in the following section.

The first sub-question.

The first sub-question was: To what extend can a computer identify multiple levels of thinking from the Higher order thinking process in online discussions within a MOOC?

The classification in general.

First, the extent to which the STC model classified first level of thinking comments as first category will be discussed. These comments lack keywords related to the second and third levels of thinking. The results showed that indeed the final model successfully classified 60 comments out of 91 (from the test set), which did not have the keywords assigned for the second category and the third, as first category. Though, as demonstrated in Table 11, the model also mistakenly classified 12 comments part of the second category and 18 as part of the third category instead of the first category. For the wrongly classified level 1 comments as level 2, an explanation can be that in the first level of thinking students also can give arguments based on the resources from the course. These arguments, however, don't have to represent level 2 thinking. For example, a sentence containing an argument but belonging to the first category could be: "As the teacher stated, a growth mindset is needed in schools.". In this case, the student

only recalls information from the course without forming a judgement or giving an explanation but still uses a keyword representative for the second category.

More interesting, however, are the wrongly classified comments that should have been classified as level one but were classified as level 3 instead of 2 since level 1 is closer to level 2 then level 3. Looking at more specific results (Table 10), a pattern can be identified in the communalities between informative keywords for level 1 and level 3. It seems that students are inclined to use first person pronouns in both cases, which explains why the model mistakenly classified more comments that should have been classified as level 1, as level 3 instead of level 2. An explanation for the use of first pronouns in level 1 could be that students expressing their opinions without judging the information or giving evidence from personal experience. Therefore, they could use combinations of words like: “I think that”, “I like that”, “I would do”. For level 3, on the other hand, students could use combinations of words like: “I have done”, “I used”, “I did not”, to explain something from personal experience. In both cases they use words like “I”, though in the case of third level thinking the comments involve reflection on past experience.

In order to classify comments as first category, the coding schema did not use any keywords, but rather the absence of keywords representing categories two and three. However, during the learning process for the automatization, the computer did find keywords that match with the first category and used this to classify comments in the later stages. As can be seen in Table 10, these keywords mostly represent the idea of students recalling information as many of these keywords represent the course content. More specifically, the strongest keyword for the first category is the name of a learning resource from the course. This identification of keywords is in accordance with the coding schema, since level 1 thinking represents recalling information of the course. This indicates that that level one thinking is actually represented by keywords and that the absence of keywords in the coding schema is therefore not correct. However, where some new keywords, that were found by the Classification, can be used to extend the existing keywords set (representing the different levels in the coding schema), adding keywords related to the course content can result in generalizability issues due to the specific context of the evaluated course. Therefore, it still seems to be correct to not identify any keywords for the first level of thinking in the coding schema, as this coding schema is intended to be used for a variety of MOOCs.

Second, the extent to which the STC model classified the comments which include the keywords related to the second level of thinking as the second category will be discussed. According to the results presented in the Table 11, the model managed to classify 50 out of 85 comments (from the test set) correctly into the second category. This means that 35 comments were mistakenly classified as the first category (23 comments) and the third (12 comments). For wrongly classifying comments representing level 2 thinking as the first category, an explanation could be that when students in an argumentation refer to resources from the course, the computer classified these comments based on the level 1 keywords as some of these keywords, based on the model, only represent level 1. Therefore, comments containing these keywords cannot be classified as another category based on the best model. A solution could be to choose an accepTable 3inimum number of keywords (e.g. 2) in order to classify the comment as the second level of thinking. However, in this case, one must be aware to not miss comments that actually belong to the second category due to not containing enough casual keywords.

Third, the extent to which the STC model classified the comments which include the keywords related to the third level of thinking as the third category will be discussed. The results illustrate (Table 11) that the model identified 57 out of 73 comments (from the test set) correctly. This implies that the model classified wrongly classified 16 comments belonging to the third category as the first category (7 comments) and the second (9 comments). It seems that the third category is best classified and therefore it could be that the complexity (e.g. causal words plus first-person pronouns and past tense) of this part of the coding schema helps in making the category more distinctive. This complexity can, however, at the same time also serve as an explanation for the wrongly classified comments. It could be that the combinations of keywords required for level 3 from the coding schema were sometimes too complex.

Altogether, it seems that the computer was able to sufficiently identify the three levels of thinking in the comments from the online discussions. Even though some comments were wrongly classified, it is not expected that the amount of wrongly classified comments significantly influences the overall distribution of thinking levels.

Recognition of keywords.

Next, we will discuss to what extent the STC model recognized the keywords from the coding schema as being informative features. Regarding this, it was expected that the supervised

text classification tool will find the same keywords from the coding schema as being the most informative features.

For the first level of thinking, Table 10 shows that the most informative keywords for lower order thinking are words representing resources from the course such as "Resource_name", "toolkit" "resources", "the blog", "the article", "speech". This, in fact, is in accordance with the definition and indicators from the coding schema, as the first level of thinking is about recalling information. Even if the keywords "resources", "the blog" and "the article" do not have such high X^2 values, they are very strong indicators for the first category one, as they did not occur at all in the other two categories. At the same time, "Resource _name" occurs 5 times for the second category and two times for the third category in comparison with 21 times for the first category. Which is interesting because it sustains the idea that the learners who engage in higher order thinking processes, become critical and use personal experience to argument their ideas rather than recalling information from the course. Another feature for the first category is "like the" which also fits to the characteristics from the coding schema, as students might give examples from the course when they express their opinions.

For the second level of thinking, Table 10 shows that "to solve" is an informative keyword as it occurs 11 times for the second category and none for the first or third category, which makes sense, as according to the literature review, critical thinkers are focusing on problem solving strategies. Furthermore, as expected, "however", "if", "so", "as", and "because" are informative keywords for the second level of thinking (and the third level of thinking). This is explained in the coding schema through the fact that students use in a critical discourse causal words to explain, argument or reason their choices.

For the third level of thinking, Table 10 illustrates that "as I" is a strong indicator for the third category. Which confirms the idea from the coding schema that higher order thinking is composed of 2 levels of thinking, by the fact that first, students engage in a critical discourse and then they are integrating past experiences in their making meaning process. Additionally, in accordance with the coding schema the keywords "I", "My", "I have", "we", "me", "when I", "help me" are identified, which confirms the fact that in the third level of thinking students reflect on their past experiences as these are personal. Moreover, "was" is an informative keyword, which reflects the use of past tense. This keyword occurred 70 times in the third category, 15 times in the first and 7 times in the second. Lastly, "wrong" was found to be an informative keyword, which could be explained by the idea that students use their past

knowledge to decide on what could be good or wrong. Also, it confirms the idea that in level 3, students are critical and use their past knowledge to infer. For both second and third level of thinking, informative keywords are the second person pronouns, “it” and “student”. However, they are stronger indicators for the third category, which could be explained by the idea that, through higher order thinking, the participants are thinking of their own context while learning. More specifically, as the participants in this course were teachers learning about growth mindsets in the classroom, using these keywords could mean that they were thinking of how implementing different things (learned in the course) in the classroom is affecting the users (their students).

Indicators that were both found for the second and third category were “so”, “it”, and “student”. Even if “so” is an indicator for both the second category and the third category, it is a stronger indicator for the third category. This could be explained by the idea that when you have experience you are more inclined to infer, which can be done by using the word “so”. This, however, does not mean that inferring does not happen in level 2 thinking, but it happens less often.

The second sub-question.

The second sub-question asked to which extent a computer can make a distinction between lower and higher order thinking in an online discussion within a MOOC. The results showed that by using the coding schema constructed in this study and the supervised text classification tool, a computer is able to distinguish very well between higher and lower order thinking. The results for classifying only two levels of thinking were better than the results for classifying three levels of thinking. This is understandable based on the idea that the chances for mistakes the computer can make decreases. For example, mistakes made by misplacing category 2 comments in category 3 and category 3 comments in category 2 are not observed anymore in the binary classification. Since the binary classification, compared to the multiclass classification, resulted in better results, a balance must be found between quality (can the computer identify the process of higher order thinking by distinguishing the three levels as represented in the coding schema) and quantity (how many correct classifications).

When it comes to the keywords that were identified in the binary classification task, it seems that the same patterns were found as in the multiclass classification task. More specifically, the model again found that words representing resources and recalling the information correspond to the first level of thinking. Additionally, for the higher levels of

thinking, the computer again found keywords representing argumentation and causation (e.g. because, as, if) and the keywords representing past experiences (e.g. “I”) as being informative. However, by not distinguishing anymore between the second and the third level of thinking, the amount of keywords representative for level 3 (compared to the multiclass classification) decreased. More specifically, these seem to be the keywords that represent past experiences (e.g. “was” and “when I”). This seems to support the proposition that level 3 thinking represents an extension of level 2 thinking where it combines judgements etc. with personal experiences.

Whereas some personal keywords representing past experience were missing in the binary classification, a new personal keyword (“will”) was found to be informative. This keyword was not part of the coding schema, but it is connected to the idea of achieving meaningful learning. According to this idea, meaningful learning occurs when students know how to use learned information in their own contexts. This seems to support the idea that the third level of thinking is not only represented by combining past experiences and knowledge, but might also include being able to predicting or respond to situations. Therefore, the coding schema could be improved by expanding the third level of thinking with other personal keywords representing the future tense, but also keywords referring to prior knowledge.

In conclusion, it seems that with both the Multiclass Classifier and the Binary Classifier, the computer was able to identify similar keywords to the ones from the coding schema. This is a positive indication for the extent to which the classification of online comments can be successful and recognition of higher order thinking processes can be automatized.

Limitations

In spite of trying to overcome the challenges that occurred in this research, the current study ended up having the following limitations: 1) Not covering all dimensions of thinking in the coding schema, and 2) a limited amount of data used as input for the automatization process.

A dimension that influences the quality of students' thinking that could not be fully analyzed within this study, is the domain specific knowledge of the participants. The reason why this dimension was not taken into account is the fact that the data used in this study did not contain personal information in order to protect the anonymity of the respondents. The importance of domain specific knowledge is well funded in the idea that one can form reasonable judgements about something based on his/her knowledge about the specific subject (Bailin et al., 1999; Facione, 1990a, 1990b; King et al., 1998; Willingham, 2007). However, an indicator for higher order thinking which was added to the coding schema is reflection on past experiences and integrating background knowledge in their reasoning. By adding this indicator, the problem was partly addressed, as the use of domain specific knowledge in their argumentative discourses was identified even without looking at the specific content of students past knowledge.

Adding more data after the first multiclass task seemed to improve the model. Therefore, it is clear that having a big amount of data to use as input for the model development can increase the usefulness of the model. Hence, a limitation in this study was the amount of data that could be used to develop the model. However, the idea of automatizing the process of analysis was founded in the fact that a manual analysis it is time consuming and requires a lot of effort in the first place. Still the results of the multiclass classification are promising, as they are indicating that automatization based on less data is still possible to some extent.

As for the binary classifier, the development dataset was composed of more data for the Higher order thinking and less for the Lower order thinking because two levels of thinking were combined into one. Hence, the binary model is better at recognizing the higher order thinking then it is at recognizing lower order thinking. However, as the objective of this study was to automatically recognize higher order thinking processes, the results are still very good as the present study provides a tool that can be used already in real world practices.

Practical Recommendation and Future Research

As stated in the previous section of the paper, the present study delivers a tool that can be used already in current practices. However, it could also be used as a basis for future research and be improved by using more data in the development of the model or changing other elements in the automatization process.

Practical Recommendations

First, in order to analyze the higher order thinking process in online learning discussions, the current study meant to combine existing theories and practices with exploration of new ideas. However, the theoretical background of the relationship between language usage and higher order thinking processes is not vast. Therefore, extensive research on how the two domains are connected could help to improve the tool. For example, research could provide more keywords representative for the categories of the coding schema which in turn could make the manual analysis more time efficient. As a consequence, if the manual analysis is more efficient, it may be easier to manually code a bigger amount of data to be used as input for the automatization.

Second, the instrument used in this study used specific parameters in the grid search which can influence the results according to the context. Adding or changing some of these parameters might improve results of the automatization. Introducing specific keywords into the system instead of letting the computer identify all the keywords itself, for example, can improve the automatized classification process. When introducing content specific keywords, however, it decreases the generalizability. Therefore, not introducing these types of keywords, keeps the tool generalizable so it can be used in more contexts. Creating a new category for the comments that do not fit in any category of thinking could improve the automatization. In addition, in order to make the analysis more reliable, a length parameter can be added. It was part of the coding schema a rule to classify the shortest comments to the first category, medium length comments to the second category and the longest comments to the third category. However, the length of comments was not part of the text processing elements of the pipeline, but it could be added in future use if needed. In order to improve the coding schema, it could be attempted to make the criteria for categories 1 and 2 more complex and, therefore, more distinctive. As with the third category, this may result in a more accurate classification of the comments in the three categories. A challenge in doing this, might be to connect research in linguistics to thinking

theories and at the same time maintain the idea of presenting student's thinking as a process in which the lower levels are subsumed by the higher levels.

In conclusion, depending on practitioners' needs, this tool can be either improved by adding context specific features, making it only usable in a specific context, or by improving general features, which ensures a more general use.

Currently assessments in MOOCs are content based, evaluating students through peer assessment and/or rubrics. The current study proposed a different approach, to look at their thinking processes during the course since research shows that if students engage in higher levels of thinking, they acquire knowledge and produce new ideas. However, the approach discussed in this paper does not take into account what specific information students need to have at the end of the course. Therefore, combining the two approaches might also be a good idea, giving a more in depth understanding of their learning process. In this way, by combining the two evaluation systems, their learning process will be evaluated from both thinking process and knowledge acquisition (Chapman et al., 2016). Therefore, a correlation between the two can be investigated, which in the end will lead to more accurate and personalized solutions.

Future Research

Current evaluation systems lack in monitoring and providing accurate information about students' learning effectiveness. By formatively assessing student's individual learning progress through an analysis of their thinking processes (demonstrated in the online discussions), a deeper understanding of the quality of their learning could be achieved. This can eventually help to give students more accurate feedback and improve the courses according to their needs.

Future researches should therefore focus on the correlation between design features of the course and student's higher order thinking. For example, by evaluating which types of instructions (video/text) were used at the moment when students engaged in higher order thinking during a course, the course design can be improved accordingly. According to Lai (2011), a few empirical studies have shown that time could have an independent effect on the development of thinking skills. Therefore, the evolution of the thinking process over time has also an important role in the understanding of the higher order thinking skills. Hence, future research could investigate if students' thinking changes during a course and what influences those changes.

As it was previously stated that domain specific knowledge plays an important role in the application of higher order thinking skills, it may be interesting for future research to investigate its influence. More specifically, to what extent the domain specific knowledge of a person can influence his/her engagement in higher order thinking during a course.

By contributing to the research in this field, teachers can be aided in providing the learners with real-time feedback. Therefore, future research is needed to investigate which types of feedback strategies are favorable for enhancing higher order thinking, what type of infrastructure should be used to provide real-time feedback, and how this should be designed to achieve easy access and use. In conclusion, all these future contributions will help to develop learning technologies for personalized learning. Ultimately, this paper provides a first step in that direction, helping students to become better higher order thinkers and course designers to improve their courses.

References

Admiraal, W., Huisman, B., & Pilli, O. (2015). Assessment in Massive Open Online Courses. Electronic Journal of E-learning, 13(4), 207-216.

Awuor, Y., & Oboko, R. (2012). Automatic assessment of online discussions using text mining. International Journal of Machine Learning and Applications, 1(1), 7 pages.

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. Journal of curriculum studies, 31(3), 285-302.

Beigi, M., Wang, J., & Shirmohammadi, M. (2015). AHRD! Take the opportunity and pioneer vocational MOOCs. Human Resource Development International, 18(2), 203-212.

Capuano, N., & Caballé, S. (2015). Towards Adaptive Peer Assessment for MOOCs. In P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2015 10th International Conference on (pp.64-69). IEEE.

Carol, R. (2002). Defining reflection: Another look at John Dewey and reflective thinking. Teachers college record, 104(4), 842-866.

Chapman, S., Goodman, S., Jawitz, J., & Deacon, A. (2016). A strategy for monitoring and evaluating massive open online courses. Evaluation and program planning, 57, 55-63.

Davis, D., & Brock, T. C. (1975). Use of first person pronouns as a function of increased objective self-awareness and performance feedback. Journal of Experimental Social Psychology, 11(4), 381-388.

Dodson, M. N., Kitburi, K., & Berge, Z. L. (2015). Possibilities for MOOCs in corporate training and development. Performance Improvement, 54(10), 14-21.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. Educational leadership, 43(2), 44-48.

Facione, P. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report).

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. American Journal of distance education, 15(1), 7-23.

Girard, J., & Girard, J. (2015). Defining knowledge management: Toward an applied compendium. Online Journal of Applied Knowledge Management, 3(1), 1-20.

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. Journal of Educational Technology & Society, 15(3), 42.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), 60-76.

He, Q. (2013). Text mining and IRT for psychiatric and psychological assessment (Vol. 15). University of Twente [Host].

Herrington, J., & Oliver, R. (1999). Using situated learning and multimedia to investigate higher-order thinking. Journal of Interactive Learning Research, 10(1), 3.

Khoshneshin, Z. (2011). Collaborative critical thinking in online environment. Procedia-Social and Behavioral Sciences, 30, 1881-1887.

King, F., Goodson, L., & Rohani, F. (1998). Higher order thinking skills: Definition, teaching strategies, assessment. Publication of the Educational Services Program, now known as the Center for Advancement of Learning and Assessment. Obtido de: www. cala. fsu. edu.

Koller, V., Harvey, S., & Magnotta, M. (2006). Technology-based learning strategies. Social Policy Research Associates Inc. http://www/. doleta. gov/reports/papers/TBL_Paper_FINAL. pdf.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory into practice, 41(4), 212-218.

Lai, E. R. (2011). Critical thinking: A literature review. Pearson's Research Reports, 6, 40-41.

Lewis, A., & Smith, D. (1993). Defining higher order thinking. Theory into practice, 32(3), 131-137.

Marland, P., Patching, W., & Putt, I. (1992). Thinking while studying: A process tracing study of distance learners. Distance education, 13(2), 193-217.

Mayer, R. E. (1996). Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. Educational psychology review, 8(4), 357-371.

Meyer, K. A. (2004). Evaluating online discussions: Four different frames of analysis. Journal of Asynchronous Learning Networks, 8(2), 101-114.

Moseley, D., Elliott, J., Gregson, M., & Higgins, S. (2005). Thinking skills frameworks for use in education and training. British educational research journal, 31(3), 367-390.

Pursel, B. K., Zhang, L., Jablokow, K. W., Choi, G., & Velegol, D. (2016). Understanding MOOC students: motivations and behaviours indicative of MOOC completion. Journal of Computer Assisted Learning, 32(3), 202-217.

Sutton, L. (2013). A MOOC of Our Own. Library Journal, 138(20), 41-42.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

Wiegersma, S., Van Noije, A.J., Sools, A.M., & Veldkamp, B.P. (2017). Supervised Text Classification: A Tool and Tutorial for Model Selection and Evaluation. Manuscript in preparation

Willingham, D. T. (2007). Critical thinking. American Educator, 31(3), 8-19.