University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]
Evaluation and Automatic Analysis of MOOC Forum Comments

by

Tim O’Riordan

Thesis for the degree of Doctor of Philosophy

February 2018
EVALUATION AND AUTOMATIC ANALYSIS OF MOOC FORUM COMMENTS

by Tim O’Riordan

Moderators of Massive Open Online Courses (MOOCs) undertake a dual role. Their work entails not just facilitating an effective learning environment, but also identifying excelling and struggling learners, and providing pedagogical encouragement and direction. Supporting learners is a critical part of moderators’ work, and identifying learners’ level of critical thinking is an important part of this process. As many thousands of learners may communicate 24 hours a day, 7 days a week using MOOC comment forums, providing support in this environment is a significant challenge for the small numbers of moderators typically engaged in this work. In order to address this challenge, I adopt established coding schemes used for pedagogical content analysis of online discussions to classifying comments, and report on several studies I have undertaken which seek to ascertain the reliability of these approaches, establishing associations with these methods and linguistic and other indicators of critical thinking. I develop a simple algorithmic method of classification based on automatically sorting comments according to their linguistic composition, and evaluate an interview-based case study, where this algorithm is applied to an on-going MOOC. The algorithm method achieved good reliability when applied to a prepared test data set, and when applied to unlabelled comments in a live MOOC and evaluated by MOOC moderators, it was considered to have provided useful, actionable feedback. This thesis provides contributions that help to understand the usefulness of automatic analysis of levels of critical thinking in MOOC comment forums, and as such has implications for future learning analytics research, and e-learning policy-making.
Table of Contents

Table of Contents ............................................................................................................. i
List of Tables ................................................................................................................... v
List of Figures ................................................................................................................... vii
DECLARATION OF AUTHORSHIP .............................................................................. ix
Acknowledgements .......................................................................................................... xi
Abbreviations .................................................................................................................. xiii

Chapter 1: Introduction ................................................................................................. 1
  1.1 Making sense of the “Data Deluge” .................................................................... 2
  1.2 Making sense of online discussion forums ....................................................... 3
  1.3 Why writing and critical thinking? .................................................................... 3
  1.4 Content analysis ................................................................................................. 4
  1.5 Outline of this thesis ......................................................................................... 4
  1.6 Contribution ....................................................................................................... 6

Chapter 2: Background and Literature Review ........................................................... 9
  2.1 Measuring and analysing learning .................................................................. 9
    2.1.1 Automatic Essay Scoring ........................................................................... 9
    2.1.2 Learning Analytics .................................................................................. 11
    2.1.3 Privacy issues ......................................................................................... 12
  2.2 Massive Open Online Courses ...................................................................... 13
  2.3 Pedagogic content analysis ............................................................................ 16
    2.3.1 Community of Inquiry ............................................................................ 16
    2.3.2 Bloom’s Taxonomy of the Cognitive Domain ...................................... 18
    2.3.3 SOLO Taxonomy .................................................................................... 19
    2.3.4 Digital Artefacts for Learning Engagement ......................................... 20
    2.3.5 Discussion on pedagogic content analysis methods ......................... 21
  2.4 Language analysis ............................................................................................ 23
    2.4.1 Linguistic Inquiry and Word Count ....................................................... 24
    2.4.2 LIWC categories .................................................................................... 24
  2.5 Interaction analysis .......................................................................................... 28
  2.6 Machine Learning ............................................................................................ 29
2.6.1 Data processing – some ethical considerations ........................................... 30

2.7 Summary ......................................................................................................... 31

Chapter 3: Pilot stage – Content Analysis of Portus MOOC Comments ............................................ 33

3.1 Methods ......................................................................................................... 33

3.1.1 Data collection and analysis ................................................................. 33

3.1.2 Procedure and analysis ........................................................................... 35

3.1.2 Language analysis .................................................................................. 36

3.2 Findings ......................................................................................................... 36

3.3 Conclusions ................................................................................................... 43

3.4 Summary ....................................................................................................... 45

Chapter 4: Large-scale Study – Content Analysis of Multiple MOOCs ........................................ 46

4.1 Methods ......................................................................................................... 46

4.1.1 Research challenges ............................................................................... 47

4.2 Findings ......................................................................................................... 53

4.2.1 Sample analysis ......................................................................................... 53

4.2.2 Research questions .................................................................................. 56

4.3 Conclusions ................................................................................................... 60

4.4 Summary ....................................................................................................... 64

Chapter 5: Automated Evaluation of MOOC Comments ............................................. 67

5.1 Method ........................................................................................................... 68

5.1.1 Machine Learning ..................................................................................... 68

5.1.2 Random Forest classification ................................................................. 68

5.1.3 Data processing ......................................................................................... 69

5.1.4 Validation .................................................................................................. 71

5.2 Procedure and classification ......................................................................... 71

5.3 Findings ......................................................................................................... 72

5.3.1 SMOTE .................................................................................................... 72

5.3.2 Building the best classifier ...................................................................... 73

5.3.3 Validation .................................................................................................. 74

5.4 Conclusions ................................................................................................... 75
5.4.2 Further Machine Learning ................................................................. 76

5.5 Summary ................................................................................................. 78

Chapter 6: ‘Live’ MOOC Case Study ......................................................... 79

6.1 MOOC moderators’ reflections on their practice .................................. 79
6.2 Method ..................................................................................................... 81
6.3 Findings ................................................................................................... 84
  6.3.1 First interviews .................................................................................. 84
  6.3.2 Second interviews ............................................................................ 91
6.4 Discussion ............................................................................................... 95
  6.4.1 Using the comment forum ............................................................... 95
  6.4.2 Identifying critical thinking ............................................................. 96
  6.4.3 Accuracy of automatic rating ........................................................... 97
6.5 Limitations and potential issues ......................................................... 100
6.6 Summary ............................................................................................... 101

Chapter 7: Overall discussion and conclusions .................................... 103

7.1.1 Results Review .................................................................................. 103
  7.1.1 Pilot and large scale studies ............................................................ 104
  7.1.2 Building machine learning classifier ............................................. 104
  7.1.3 MOOC case study ............................................................................. 105
7.2 Research questions ................................................................................. 105
7.3 Discussion ............................................................................................... 107
7.4 Developing an automated analysis for online learning ....................... 107
  7.4.1 Validity of automatic methods ......................................................... 108
  7.4.2 Ethical considerations ..................................................................... 109
7.5 Limitations and future work ................................................................. 110
7.6 Final remarks and conclusions ............................................................. 111
7.7 Contribution ........................................................................................... 111

Appendix A .................................................................................................. 113
Appendix B .................................................................................................. 114
Appendix C .................................................................................................. 115
Appendix D .................................................................................................. 116
List of Tables

Table 1: Community of Inquiry: Cognitive Presence ................................................................. 18
Table 2: Bloom’s Taxonomy ........................................................................................................ 19
Table 3: SOLO Taxonomy ........................................................................................................... 20
Table 4: Adapted DiAL-e Framework ......................................................................................... 21
Table 5: Hyland’s rhetorical features mapped to LIWC2015 categories. .................................... 28
Table 6: Intra-rater reliability values ............................................................................................ 35
Table 7: Percentage occurrence of academic rhetorical features in Tse & Hyland 2009.......... 38
Table 8: Coding scheme correlations ............................................................................................. 39
Table 9: LIWC2007 analysis ......................................................................................................... 40
Table 10: Percentage occurrence of academic rhetorical features in MOOC2015 corpus........ 56
Table 11: Correlation and prediction results ................................................................................. 57
Table 12: Attributes used to build the classifier ........................................................................... 74
Table 13: Reliability results (results for test set are highlighted) .................................................. 75
Table 14: Confusion Matrix for the test set (n=300) ................................................................. 75
Table 15: Reliability results for n=1500/70:30 dataset .............................................................. 77
Table 16: Confusion Matrix n=1500/70:30 test set ................................................................. 78
Table 17: Identified themes in first interviews .............................................................................. 84
Table 18: Identified themes in second interviews ......................................................................... 91
List of Figures

Figure 1: Weekly distribution of sample comments.................................................................37
Figure 2: Total pedagogical rating for Portus MOOC............................................................37
Figure 3: Correlation between Bloom and SOLO.................................................................39
Figure 4: Correlation between Bloom and CoI .......................................................................39
Figure 5: Correlation between CoI and pronouns by average scores for 12 steps..............41
Figure 6: Correlation between CoI and present tense by average scores for 12 steps..........41
Figure 7: Correlation between CoI and pronouns by individual scores for >100 comments....41
Figure 8: Correlation between CoI and present tense by individual scores for >100 comments.41
Figure 9: Correlation between CoI and word count per comment.........................................42
Figure 10: Correlation between CoI and prepositions by average scores for 12 steps.........42
Figure 11: Correlation between CoI and positive words by average scores for 12 steps.......43
Figure 12: Correlation between CoI and Likes.......................................................................43
Figure 13: Sample selection process....................................................................................50
Figure 14: Comment rating process.......................................................................................52
Figure 15: Normal P-P plot of Bloom average.................................................................53
Figure 16: Normal P-P plot of CoI average.......................................................................53
Figure 17: Weekly distribution of sample comments...............................................................54
Figure 18: Total scores using Community of Inquiry: Cognitive Presence rubric...............54
Figure 19: Total scores using Bloom’s Taxonomy rubric....................................................55
Figure 20: Total scores for Social Science MOOC using CoI..............................................55
Figure 21: Total scores for Humanities MOOC using CoI...................................................55
Figure 22: Total scores for Natural Science MOOC using CoI............................................55
Figure 23: Total scores for Social Science MOOC using Bloom.......................................55
Figure 24: Total scores for Humanities MOOC using Bloom .................................................. 55
Figure 25: Total scores for Natural Science MOOC using Bloom ........................................ 55
Figure 26: Scatter plot showing correlation between word count and Bloom score .......... 56
Figure 27: Scatter plot showing correlation between first-person singular and Bloom score ........................................ 56
Figure 28: Scatter plot showing correlation between causal words and Bloom score .......... 56
Figure 29: Decision Tree showing possible categorisation judgements based on Word Count .. 69
Figure 30: Random Forest creates an ensemble of decision trees ........................................ 69
Figure 31: SMOTE pre-processing for class balancing in training set (n=300). ................. 73
Figure 32: SMOTE pre-processing for class balancing in testing set (n=300). ................. 73
Figure 33: Automatic ratings for Step 1, Week 1 ................................................................. 83
Figure 44: Overall research design ................................................................. 104
List of Accompanying Materials

1.1 Datasets:

Pedagogical rating and LIWC analysis of 1500 comments, and Pedagogical Rating and LIWC Analysis of 600 comments are published here:
DOI: 10.5258/SOTON/D0380
DECLARATION OF AUTHORSHIP

I, Timothy James O’Riordan declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Evaluation and Automatic Analysis of MOOC Forum Comments

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

Signed: …………………………………………………………………………………………………………………

Date: …………………………………………………………………………………………………………………
Acknowledgements

Heartfelt thanks to everyone who has helped and supported my work in the production of this thesis. I am extremely grateful to my wife and family, who have provided unfailing support, receptive ears, and critical feedback over the past four years as I worked through the many intellectual challenges provided by my exploration. A huge ‘thank you’ also to my supervisors, Dr David Millard and Dr John Schulz, for their critical, learned, dedicated, and practical help and support. I am very grateful for the advice and feedback they provided throughout my research, which has been invaluable in the production of this thesis. In addition, the lecturers, professors, and administrators within the Web and Internet Science (WAIS) research group, Electronics and Computer Science, the School of Education, and the University beyond, deserve many thanks for providing me with inspiration, encouragement and assistance in my research journey. Thanks also to fellow IPhD cohort members who have shared the journey since we started the Web Science MSc back in 2013.

I am also very grateful to the MOOC moderators for the time they gave to my user study, providing crucial insights into the usefulness of my research.

Final thanks go to the Web Science Centre for Doctoral Training (CDT) and WAIS, who provided the necessary space and support, and the EPSRC Digital Economy Programme, who gave me the wherewithal to undertake this research. The CDT especially has been a vital source of assistance, and I thank everyone who contribute to this important project.
Abbreviations

Automated Essay Scoring (AES)

Community of Inquiry: Cognitive Presence (CoI)

Digital Artefacts for Learning Engagement Framework (DiAL-e)

Computer Supported Collaborative Learning (CSCL)

General Data Protection Regulation (GDPR)

Intraclass Correlation Coefficient (ICC)

Intra-rater reliability (IRR)

Learning Analytics (LA)

Linguistic Inquiry and Word Count (LIWC)

Machine Learning (ML)

Massive Open Online Course (MOOC)

Project Essay Grade (PEG)

Structure of Observed Learning Outcomes (SOLO)

Synthetic Minority Over-sampling Technique (SMOTE)

Waikato Environment for Knowledge Analysis (WEKA)
Chapter 1: Introduction

The emergence of Massive Open Online Courses (MOOCs) have been proclaimed as a “game changer” for education (Boven, 2013, p. 1). Adopting the Web to disseminate high quality learning, and embracing the characteristically informal and social aspects of the ‘read/write’ Web, MOOCs provide opportunities for institutions and educators to reach a new, large, global audience of learners, but also create ‘Big Data’ challenges.

Commercial and institutional MOOCs are characteristically developed by subject experts with support from learning designers and technologists who provide registered learners with access to learning resources. Resources may be presented using a variety of media, and/or text formats, and may include peer review assignments, as well as formative and summative tests. Learners are encouraged to contribute to online forums, and participate in online conversations with fellow learners as well as the educators and mentors running the MOOC (Laurillard, 2002).

In face-to-face and small cohort online teaching and learning environments, pedagogy typically involves providing timely, relevant feedback to learners. However, within MOOC discussion forums, many thousands of learners may communicate 24 hours a day, 7 days a week. Attempting to provide support in this environment is currently haphazard, and ultimately unsatisfactory for those educators and mentors engaged in running MOOCs.

Although there is evidence that learners can successfully support each other through peer networking, forum facilitation and direction is viewed as fundamental in encouraging them, and orchestrating successful learning experiences (Garrison, Anderson and Archer, 1999). However, keeping up with the relentless flood of comments in MOOC forums is a significant challenge, with studies exploring MOOC moderation reporting a less than 0.7% reply rate, and difficulty in assessing which comments to respond to and which to ignore (León et al., 2015).

Supporting learners is also a critical part of MOOC moderators’ work, and identifying learners level of critical engagement is an important part of this process. Critical thinking is considered by many educational researchers to be fundamental to successful learning (Ennis, 1993; Silva, 2009; Lai, 2011), and in this thesis I argue that an important deciding factor for educators and mentors when evaluating learners’ written contributions is the level of critical thinking identifiable within comments. Computational solutions to measuring critical thinking in learners’ comments have been proposed, but research has predominately been focused on developing interventions for formal learning environments (Wise and Paulus, 2016). In addition, there is a strong indication that complex algorithm-based solutions are a barrier to implementation (Kovanovic et al., 2017), and new regulations requiring ‘black box’ algorithmic decision making to be more accountable and transparent than at present have significant implications for the design of automatic data-driven solutions (European Commission, 2017). This thesis argues that in order to be successfully applied,
these solutions need to be readily understandable and explainable to data subjects, as well as the educators, administrators, and other users of resulting analyses.

In order to address these issues, I adopt a pedagogically sound approach to classifying comments, and report on several studies I have undertaken which seek to ascertain the reliability of these established content analysis approaches, and establish associations with these methods and linguistic and other indicators of critical thinking. I discuss the development of a simple algorithmic method of classification based on automatically sorting comments according to their linguistic composition, and then evaluate an interview-based case study, where this algorithm is applied to an on-going MOOC. Finally, I explore critical debates on these issues, and the implications of automated rating of MOOC learners’ comments for public policy, educational practitioners, and the research community are discussed.

1.1 Making sense of the ‘Data Deluge’

Learners and educators have actively engaged with formal educational content on the Web, as well as informal content with educational value, for more than 25 years. However, the success of the Web as a learning environment has led to the proliferation of online learning objects, fuelled in recent years by the development of MOOCs, which is causing problems for its effective use. Principally, the large amount of content and the difficulties in evaluating their efficacy provides challenges for people seeking resources that are relevant to their personal learning needs (Dichev et al., 2011), provide challenges for learners and educators in interpreting what is taking place (Greller and Drachsler, 2012), and is a significant hurdle for learners to overcome (Walraven, Brand-Gruwel and Boshuizen, 2009). This ‘data deluge’ (Anderson, 2008), which affects all aspects of developed society, is driving the wide adoption of automated ‘data-driven decision making’ in education (Picciano, 2012), and provides complex challenges regarding how data derived from learning environments is measured, analysed and disseminated (Koene, Webb and Patel, 2017).

While some suggest that almost any feedback provided by technology helps students learn better (Belardi, 2015), others argue for a more nuanced development of methods that combines data analysis and feedback with deep knowledge of the teaching and learning process (Ferguson, 2014), and for an evidence-based approach to setting evaluation standards (Scheffel, 2015). In addition, concerns regarding the potential of automated data analysis methods to ‘hard-wire’ unintentional bias into decision-making are informing the creation of robust ethical standards to control algorithm development (Koene, 2017), as well as regulation aimed at removing obscurity from algorithmic evaluations (Goodman and Flaxman, 2016).
1.2 Making sense of online discussion forums

Online discussions within Computer-Supported Collaborative Learning (CSCL) environments are a useful source of data for those involved in analysing educational interventions, as they reflect engagement with learning and offer a “gold mine of information concerning…the acquisition of knowledge and skills” (Henri, 1992, p.118). Dialogue helps learners build personal social capital and gain exposure to new ideas (Kovanovic et al., 2014), and language used in these environments has been shown to indicate depth of critical thinking (Allen, Snow and McNamara, 2015), a key component of collaborative learning. Discussion forums have been identified as rich seams of instructor and learner interaction data that can be mined to monitor levels of participation, but they can also reveal significant aspects regarding the quality of interaction through the adoption of content analysis techniques (Chan et al., 2002; Meyer, 2004; Joksimovic et al., 2014). However, educational research has predominately focused on discussion within formal learning environments, rather than the informal setting provided by MOOCs (Wise and Paulus, 2016), and has been dominated by assessments of the quantity rather than the quality of interaction.

Weber defines content analysis techniques as research methods that build on “procedures to make valid inference[s] from text” (Weber, 1990, p.1), and in this thesis, I employ human raters to interpret and categorise critical thinking within MOOC comment forums, and evaluate an algorithm derived from these categorisations in a ‘live’ setting.

1.3 Why writing and critical thinking?

Investigating user-generated text in MOOC discussion forums provides insights into learners’ engagement with discussion topics. While not taking place in a formal educational environment, writing in this context can act as an indicator of depth of critical thinking, an essential component of study (Silva, 2009). The act of writing works as “vehicles for thought” (Menary, 2007, p. 622) that allow writers to sort, edit, and clarify ideas into a communicable form (Tyler, 1949). Writing enables the construction of external representations more concrete than internal, abstract thought. The words we use, their order and the ideas they contain, contribute to an indication of the depth of our cognitive engagement, as well as the level of importance we give to any associated topic (Oatley and Djikic, 2008). In this way, writing, word volume and language use become proxies for our level of cognition and understanding.

Critical thinking is conceptualised in different ways by different scholastic traditions. Lai (2011) identifies three similar but divergent viewpoints, where the Philosophical tradition views critical thinking as perfecting thought to develop critical “traits of mind” (e.g. clarity, precision, accuracy, etc.) (Paul, 1992, p. 4), Cognitive Psychology defines critical thinking by how critical thinkers behave (e.g. adopting problem solving strategies), and Education applies frameworks developed from classroom observation to identify critical thinking (e.g. the three highest levels of Bloom’s
taxonomy) and it is recognised as a key objective when encouraging learners to adopt in-depth, rather than surface, learning approaches (Newman, Webb and Cochrane, 1995). While there is some dispute regarding the transferability of critical thinking skills across disciplines (e.g. domain-specific Mathematics skills may not be directly applicable to English), there is general agreement that critical thinking is typically revealed through asking questions for clarification, defining terms, identifying assumptions, analysing arguments, making reasoned inferences, evaluating, making decisions or solving problems, interpreting and explaining, predicting, and seeing both sides of an issue (Lai, 2011).

My thesis is in agreement with Lipman, who argues that critical thinking is best acquired within the social context of a community of inquiry (Lipman, 2003), as well as Biggs association of deep learning with ‘affective involvement’ through interaction (Biggs, 1987). Dialogue between learners stimulates cognitive conflict which encourage reflection, assimilation of new knowledge, and continued interaction. In this context ‘critical thinking’ is perhaps best defined as, “reasonable and reflective thinking that is focused upon deciding what to do or believe” (Norris and Ennis, 1989, p. 1). Measuring levels of critical thinking may provide an indication of the extent to which learners are building on their knowledge and can enable MOOC moderators to sort learners’ comments, identify their progress, respond appropriately, and make inferences about learning design success.

1.4 Content analysis

While the high volume of MOOC data provides an unprecedented opportunity for insight into how CSCL is used in practice, the reliability of coding schemes used for pedagogical content analysis used to explore this data is questioned. Specifically, some argue that methods lack coherence and validity (De Wever et al., 2006), and others identify a research-inhibiting lack of consistency in their application (Weltzer-Ward, 2011). High volume of data combined with uncertainty regarding analysis methods emphasises the importance of constructing theoretically sound methods that can reliably, and automatically, analyse this data. To address these issues, my research employs appropriate pedagogical content analysis methods, using instruments that have previously been adopted in studies exploring depth of critical thinking evidenced in CSCL. By this, I mean the understanding that observations submitted by learners to MOOC discussion forums act as an indicator of the level of their critical thinking, from which alignment to coding schemes used for pedagogical content analysis can be inferred.

1.5 Outline of this thesis

This introduction has provided the context from which my research originated. Specifically, my thesis focuses on an evaluation of the methods used in developing an algorithm that can be used to automatically rate levels of critical thinking in CSCL, and its usefulness to educators running a
MOOC. In order to undertake an exploration of the issues outlined above, I adopt a mixed methods approach to investigate pedagogical and linguistic content analysis methods, validate methods that indicate levels of critical thinking in MOOC comments, develop an algorithmic approach to automatically rate these comments, and explore the viability of algorithmic decision making in support of informal learning. This is an area of vital interest to administrators, moderators, learners, and learning analytics researchers that requires further understanding in order to support the implementation of automated rating methods in informal online learning environments.

The primary motivation for this thesis is to test hypotheses regarding the adoption of comments in MOOC discussion forums as pedagogically meaningful annotations. The aim being to develop a means of automatically categorising comments in terms of their inferred depth of critical thinking. The value of this is to develop meaningful feedback methods for high volume discussion forums that will improve moderators’ decision making when engaging with learners. To achieve this outcome, this thesis reports on the outputs of a four-stage project:

1. a pilot phase which builds on earlier work that compared coding schemes used for pedagogical content analysis with common learning analytics and language analysis methods;

2. a large-scale study which develops the pilot and employs multiple raters to evaluate a large number comments from discipline-diverse discussion forums;

3. a machine learning experiment using data from stage 2 to automate comment classification; and

4. a case study to elicit reflections from MOOC practitioners on outputs from the automatic comment classification method developed in stage 3.

The overall study sets out to answer four key research questions:

RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forum comments? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?

RQ2: Are linguistic content analysis measures significant indicators of levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?

RQ3: To what extent do typical measures of attention to learning (such as social interactions) indicate levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?

RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments?
The first two stages seek to answer the first three questions. They investigate the reliability of coding schemes used for pedagogical content analysis, and in particular seek to explore the potential of established content analysis methods for identifying levels of critical thinking in MOOC discussion forums. They lay the foundation for the final stage of the thesis, which seeks to answer the fourth question and reports on the development of an algorithmic approach to automatically analyse learner comments. The views of MOOC educators and mentors on the usefulness of this approach will then be considered.

The remainder of this document sets out the work undertaken and is structured as follows. Chapter 2 gives an overview on the background literature on how discussion forums in CSCL are analysed, the study of MOOCs, and the application of algorithmic decision-making in this area. The thesis then reports on the main outputs of this research. Chapter 3 presents a pilot study that explores the use of content analysis methods to rate comments in a MOOC. Chapter 4 evaluates a large-scale content analysis study employing a number of human raters to categorise a number of discussion forum comments in a variety of MOOCs. The purpose of this study is to establish key attributes in comments that indicate levels of critical thinking for use in the development of a rating algorithm. Chapter 5 explores the construction of this algorithm, using a machine learning approach. Then Chapter 6 reports on the results of a small-scale case study carried out to obtain the views and motivations of practitioners running a MOOC, regarding the outputs of the comment rating algorithm scores which were provided to them, and their attitudes to managing learner comments.

Results from the studies are critically assessed in Chapter 7, including a discussion on their meaning and impact, before limitations of this work are explored and final conclusions and recommendations for further study are made.

### 1.6 Contribution

This thesis provides insights into a number of areas that affect the development and use of automated analysis methods in CSCL, and contributes to current research, practitioner, and policy debates. The examination of content analysis methods I have undertaken, the new datasets created, and my novel approach to comment analysis, contribute to current deliberations in learning analytics and CSCL research. In my user study, practitioners reflect on their needs when running MOOCs, and my pedagogical theory-based analysis contributes to on going debates on the application of automated feedback in learning. Through the development and application of an unambiguous analytical tool, my thesis provides policy makers with a coherent method to evaluate learner engagement, as well as a learned reflection on the value of such methods.

In particular, it presents three studies that assess methods of evaluating critical thinking in MOOC forums, using human raters in the first two instances. An automated comment rating method is
derived from these studies, applied to a ‘live’ MOOC, and the views of practitioners supporting the MOOC are analysed and evaluated.

As a result of this work, I have had one paper published in an online journal, and two published in conference proceedings:


The next chapter provides information sourced from background literature on how discussion forums in CSCL are analysed, the study of MOOCs, and a reflection on the application of algorithmic decision-making in this area.
Chapter 2: Background and Literature Review

This review of relevant literature will examine the main issues surrounding the evaluation of learner interaction in CSCL settings, pedagogical issues in the context of CSCL in general, and MOOCs in particular, and manual and computational approaches to measuring and evaluating activity in CSCL learning environments. The motivation for collecting and analysing learner activity data will be examined, as will pedagogical classification and frameworks relevant to understanding how learners share their learning in comment forums. Ethical considerations regarding the collection and processing of personal data will also be considered.

Because establishing relevant learning analytic methods that can reliably measure cognitive activity is central to this thesis, a sensible starting point is to explore studies that have been undertaken in this area.

2.1 Measuring and analysing learning

Grading, ranking, classifying, and recording student activity are fundamental activities in formal education wherever institutions or teachers need to know how students are developing, and where students require feedback, either formative or summative, on their progress. Computational approaches to measuring and analysing this activity holds the promise of relieving human effort and dealing with large amounts of data at speed, but is a controversial topic that demands a multidisciplinary perspective encompassing “not only psychometrics and statistics, but also linguistics, English composition, computer science, educational psychology, natural-language analysis, curriculum, and more” (Page, 1966, p. 88).

2.1.1 Automatic Essay Scoring

In the mid-1960s, as part of Project Essay Grade (PEG), a number of experiments to assess the reliability of machine-based essay grading were undertaken, adopting a word and punctuation count method of “actuarial optimization” to “simulate the behaviour of qualified judges” (Page, 1966, p. 90). Using 30 features that approximated to values previously identified by human experts, PEG explored essays written by high school students, and found statistically significant associations with human criteria. The highest associations were average word length ($r = 0.51$), use of words commonly found in literature ($r = -0.48$), word count ($r = 0.32$), and prepositions ($r = 0.25$) (p. 93). While costly in terms of the time taken to input, these initial experiments were highly successful, showing strong multiple correlations between computational evaluations and human judges ($r = 0.71$). In the face of hostility and suspicion from progressive as well as established interests, and hampered by the rudimentary computing facilities available at the time, further development of the project waned (Wresch, 1993).
As computers became ubiquitous and software improved in the decades that followed, PEG was revived and applied to large-scale datasets. These experiments resulted in algorithms that were shown to surpass the reliability of human expert rating (Page, 1994). In recent years, the focus of developing automated essay scoring (AES) algorithms has shifted from academic faculties to the research and development departments of corporations. AES has been successfully marketed, and different systems are currently used to assess students’ writing in professional training, formal education, and MOOCs, primarily in the United States (Williamson, 2003; Anson et al., 2013; Whithaus, 2015; Balfour, 2013).

While the details of proprietary AES algorithm design is a matter of commercial confidentiality, systems continue to be based on word and punctuation counts and word lists, with the addition of Natural Language Processing techniques (Burstein et al., 1998), Latent Sentiment Analysis (Landauer, Foltz and Laham, 1998), and Machine Learning methods (Turnitin, LLC, 2015; McCann Associates, 2016).

Controversy and criticism of AES has focused on the inability of machines to recognise or judge the variety of complex elements associated with good writing, the training of humans to mimic computer scoring, over-emphasis on word count and flamboyant language, and the ease with which students can be coached to ‘game the system’ (Haswell and Wilson, 2013; Anson et al., 2013; Perelman, 2014).

However, many of these criticisms are levelled at the wide-spread application of computational methods to replace human rating, criticisms which were clearly addressed early in the development of AES. Page argues that computational approaches are based on established experimental methods that privilege, “data concerning behaviour, rather than internal states, and the insistence upon operational definitions, rather than idealistic definitions” (Page, 1969, p. 3), and that machine grading simply replicated the behaviour of human experts. In response to arguments that machines where not capable of judging creativity, Wresch cites Slotnick’s support for the use of AES to indicate deviations from norms and highlight unusual writing, which could then be referred for further human assessment (Wresch, 1993). In recent work exploring the use of automated assessment in MOOCs, while recognising the limitations of AES in assessing unique writing (e.g. individually selected topics, poetry, original research), Balfour suggests the use of computational methods to correct mechanical writing problems, combined with a final, human, peer review (Balfour, 2013).

As well as grading purposeful learner activity, learners’ Web interactions, such as within Learning Management Systems, CSCL, and MOOCs, is viewed as a source of potentially valuable data from which learners’ progress can be observed, measured, and predicted (Siemens and Long, 2011). Within the emerging field of Learning Analytics (LA), the analysis of learner interactions within networks has become widely recognised as a valuable tool for providing feedback on learner
progress (Klisener and Fortenbacher, 2015), predicting learner behaviour (Beck and Woolf, 2000), developing collective intelligence (De Liddo, Sándor and Buckingham Shum, 2012) and automating metadata annotation (Downes, 2004).

2.1.2 Learning Analytics

Through the aggregation of online activity associated with learning resources, LA techniques provide additional rich and effective means of evaluation, and offer opportunities for enhanced discoverability and personalisation (Siemens, 2012a). Learning analytics is a relatively new area of research that is comparable with other fields, such as Big Data (Cox and Ellsworth, 1997), e-science (Roure, Jennings and Shadbolt, 2001), Web analytics (Pitkow, 1997), linguistic analysis (Pennebaker, 1993), and Educational Data Mining (Baker, Corbett and Koedinger, 2002). All of these fields use large collections of in-depth data to identify patterns. In LA, data derived from learners online interactions are analysed to build prediction models with the aim of understanding and enhancing learning experiences (Kizilcec, Piech and Schneider, 2013; Wen, Yang and Rosé, 2014). As with other big data fields the application of LA outputs have critical implications for future teaching and learning practice, and have the potential to disrupt practice.

The underlying assumptions of LA are based on the belief that Web-based proxies for behaviour can be used as evidence of knowledge, competence and learning. Through the collection and analysis of ‘trace data’ (e.g. learners’ search profiles, their website selections, and how they construct, use and move information on the Web) learning analysts explore “how students interact with information, make sense of it in their context and co-construct meaning in shared contexts” (Knight, Buckingham Shum and Littleton, 2014, p. 10). LA methods that focus on discussion forums include processes that identify learners’ attention, sentiment analysis (agreement or disagreement), learner activity, and relationships between learners within forums (De Liddo et al., 2011).

As with AES, design of LA methods is not neutral, but inevitably reflects the ideology, epistemology and pedagogical assumptions of the designers. Data are not value free; they require interpretation and are subject to ‘interpretative flexibility’ as much as any other technological development (Collins, 1983; Hamilton and Feenberg, 2005).

Historically, information and communication technology interventions in education have been based on objectivist assumptions that learners’ ability to represent or mirror reality are key to judging evidence of knowing and learning. While still maintaining a strong position in summative assessment, over the past 30 years the assumptions underlying objectivism have been challenged by a growing body of constructivist thought which holds that the key to understanding how knowledge is built is through examining the interpretive process of learning (Jonassen, 1991). The practice of LA broadly adheres to either an objectivist perspective, which prioritises the use of trace data to
make evaluations of knowledge acquisition (assessment of learning), or a constructivist position, which values the provision of feedback to facilitate improved learner self-awareness (assessment for learning).

Critics have focused on a number of problems with the outcomes of analysing learning. The reliable validation of human and automatic annotation is problematic and unresolved (Rourke et al., 2001; De Wever et al., 2006); crude feedback mechanisms can lead to efforts to ‘game the system’, so that educators design learning objects to elicit positive responses regardless of the overall benefit to learners; and analytics can lead to learner-dependence on feedback rather than their own understanding (Buckingham Shum and Ferguson, 2012). Furthermore, the ethical implications of processing personal data are not fully understood (Ferguson and Clow, 2017).

2.1.3 Privacy issues

Hildebrandt argues for what US legal scholars call ‘technological due process’ in the handling of learners’ data, and recommends that pseudonymisation, encryption, non-discriminatory data mining, and profile transparency, are essential for ethical and legal implementation of learning analytics (Hildebrandt, 2017). She suggests that the processing of LA operates at two levels which have implications for how learners’ data are handled.

First order LA involves acting on inferences which affect identifiable learners. Privacy may be an issue whenever data is processed (i.e. shared) in a way that infringes on the legitimate expectations of learners, e.g., if data are shared out of context, enabling non-critical staff, funding agencies, or law enforcement agencies access to personal, identifiable data.

Second level LA explores patterns in pseudonymised data (e.g. where personal names are replaced with arbitrary identifiers) and may associate specific features, behaviours and contexts to specific performance metrics. These may be used to review outcomes from past events, predict future performance, and provide general insight into what specific interventions and circumstances may improve future learner performance.

In my thesis, I use pseudonymised data to identify patterns in levels of critical thinking in order to develop a method which would ultimately enable MOOC moderators to intervene and respond more effectively to comments made by individual learners. Therefore, while not intentionally identifying individual students, future operationalisation of methods developed in this study would require strict safeguards to be implemented. With the introduction of the European Unions’ General Data Protection Regulation (GDPR) in 2018 (see section 2.7.1 below), this has become an issue of increased urgency for those involved in implementing learning analytics interventions.
## 2.2 Massive Open Online Courses

A key area of learning analytics research focus, MOOCs have been heralded by commentators as a ‘revolution’ and ‘game changer’ for education (Pappano, 2012; Boven, 2013) which has initiated considerable debate regarding the delivery and reception of online learning. In July 2017, *MOOC List* (an online MOOC inventory) cited more than 90 providers based in over 50 countries running in excess of 5,900 courses, in more than 30 languages (CoToNet Information Technologies Ltd, 2016). While assertions have been made of enrolments of up to 230,000 participants (Jordan, 2015), completion rates are typically about 6.5% of initial subscribers (Jordan, 2014).

MOOCs are usually described in terms of their approaches to pedagogy. One of the first MOOCs, *Connectivism and Connective Knowledge Online Course* (Downes and Siemens, 2008), developed from debates on the nature of knowledge and the dissemination of Open Educational Resources (OER) in Web environments. Downes (2005) asserts that a new category of knowledge has evolved from the ‘emergence’ of knowledge resulting from interactions between networked educators and learners – a process resulting in “distributed knowledge” (Downes, 2013, p. 84).

MOOCs typically provide opportunities for learners to join online courses without paying a registration fee. Once registered, learners are given access to learning resources provided in a number of stages (or ‘steps’) distributed over several weeks. Learning resources are presented in video, audio, and/or text formats, and may include quizzes, as well as peer review exercises. Learners are encouraged to post comments within discussion forums, and to take part in online debates with fellow learners, as well as the educators and mentors running the MOOC.

Some claim that MOOCs are not a new phenomenon, as openly available, online tutorials, resources and discussion forums existed on the pre-Web Internet (Masson, 2012). However, the apparent ‘massive’ uptake of MOOCs and the potential not only to disseminate knowledge to a huge audience, but also to facilitate analysis of learning on a large scale, has generated substantial interest from platform developers and educational policymakers.

The launch of for-profit platforms Udacity and Coursera in 2012 was followed by a wider uptake of the format by universities and corporate institutions, and a shift towards traditional, objectivist teaching methods, including video lectures, quizzes and testing – a format that Siemens (2012b) has labelled ‘xMOOC’ (as opposed to ‘cMOOC’ which apply connectivist approaches to learning). While there are differences in pedagogical methods, MOOCs follow similar patterns which are potentially disruptive to established approaches to university teaching: “MOOCs allow free and easy registration, do not require formal withdrawals, and include a large number of students who may not have any interest in completing assignments and assessments.” (Breslow et al., 2013, p. 14).

While research indicates the vast majority of MOOC learners do not take part in discussion forums (Breslow et al., 2013; Gillani et al., 2014), evidence suggests that there are significant levels of
interaction between those learners who activity take part (De Waard et al., 2012; Gillani and Eynon, 2014). Participation in online discussions can support higher learning (Hannafin, Land and Oliver, 1999), Schwartz (1999) argues that the production of knowledge that happens when people communicate is fundamental to the appropriation of knowledge, and it is asserted that online discussion enhances performance in problem solving tasks (Oh and Jonassen, 2007). In addition, research into language use in MOOC discussion forums indicates that learners who use sophisticated linguistic skills are more likely to complete a MOOC (Crossley et al., 2015).

Although a small proportion of learners take part in discussion forums, because of the often large number of registrations involved, the quantity of social learners (i.e. learners who post one or more comments [Adams, 2017]) is often very high, leading to unique challenges for those running MOOCs. In Beetham and Sharp’s review of the pedagogic requirements for online learning, the authors observe that contemporary learners are no longer viewed as, “passive recipients of knowledge and skills but as active participants in the learning process” (2007, p. 2), and that “even the most self-directed of adult learners can benefit from the support of others” (2007, p. 1). In a MOOC environment, this support is occasionally provided by other learners, but most MOOCs deliver significant support through moderation by MOOC educators and mentors.

MOOC moderation is essential in supporting the learning communities that develop as a MOOC progresses, and primarily involves reading and responding to learners’ comments. Those carrying out this role are referred to by different titles in the literature. Some refer to them as facilitators (Ross et al., 2014), and others call them teaching assistants (Hazlett, 2013; Mueller, 2014). The FutureLearn platform use the terms educator and mentor to describe this role (FutureLearn, 2017c). Educators are subject matter experts, typically experienced teachers who are closely involved in structuring and selecting course content (White and White, 2016). Mentors are often postgraduate students with some teaching experience, but not closely involved in the learning design aspect of a MOOC (León et al., 2015). As my research focuses on FutureLearn MOOCs, I have adopted this platforms’ terminology when referring to these distinct roles, and use ‘moderator’ to discuss the combined effect of the roles they undertake.

Throughout the duration of time-limited MOOCs (which normally run for no longer than 6 weeks) moderators have a fundamental role to play in reading and responding to learners’ comments and questions. They provide much needed encouragement for learners who may feel disconnected to the course content and isolated from other learners. They typically attempt to encourage learners to engage in conversation, help develop understanding of new concepts, and seek out learners’ comments that express important ideas. Following the constructivist perspective, rather than simply convey information, moderators aim to encourage learners to be reflective and inquiring, to construct their own path to deeper understanding (Crow, 2012).
Moderation is therefore an important part of what the Community of Inquiry model refers to as teaching presence: the “binding element” in CSCL that coordinates the learning experience, and frames the social and cognitive presence aspects of a course (Garrison, Anderson and Archer, 1999, p. 96). Teaching presence includes the expectation that moderators guide discussion “toward higher levels of learning through reflective participation as well as by challenging assumptions and diagnosing misconception” (Anderson et al., 2001, p. 3). In addition to designing and creating the learning environment, this is achieved through the two distinct activities of, ‘facilitating discourse’ and ‘directing instruction’, normally via comments within online forums (Anderson et al., 2001).

‘Facilitation’ principally involves supporting and guiding learners’ discussion, and is comprised of comments that prompt dialogue, acknowledge learner comments, seek consensus, and evaluate discussion. On the other hand, ‘direction’ engages educators’ and mentors’ subject expertise and leadership in the delivery of course content. These comments normally include interjections that present relevant content for discussion, provide explanatory feedback, correct misinterpretations, suggest further reading, summarise debate, and answer technical queries (Garrison, 2007).

The significant role of educators and mentors in moderating informal MOOC environments is also recognised by Cormier and Siemens, who assert that: “the true benefit of the academy is the interaction, the access to the debate, to the negotiation of knowledge—not to the stale cataloging of content” (2010, p. 32).

However, keeping up with the flow of comments is a major challenge for MOOC moderators. León et al. report that mentors of a highly-subscribed MOOC had to make rapid judgements on which comments to respond to, were only able to reply to less than 0.7% of all comments, and that, “it was difficult to determine the most interesting, engaging or current issues for learners at this scale” (León et al., 2015, p. 16).

León et al. suggest automated analysis as a solution to managing the torrent of comments. However, providing meaningful feedback to course mentors raises further issues.

The enormous amount of digital content emerging from MOOCs creates large datasets concerning learner logins, interaction, test scores, and other examples of online behaviour, which provide opportunities for analysis of learning activity. However, carrying out effective analyses of MOOC data requires significant time investment and advanced data science abilities, which are often out of the reach of the educators, mentors, and learning designers who create and run MOOCs. As Kovanovic et al. discovered when deploying their MOOCito tool, developing analytic instruments which produce actionable outputs that support the improvement of learning, require solutions that are pedagogically relevant, and are readily interpretable by users who are inexperienced in data science (Kovanovic et al., 2017).
Because the ultimate aim of this thesis is to develop automated methods of assessing comments that can be readily comprehended by educators, mentors, and others involved in MOOC development, it is important to ground this method in established pedagogic theory.

2.3 Pedagogic content analysis

Computer-supported collaborative learning (CSCL) research has witnessed a change in learning design focus in recent years, from instructor-led teaching and learning practices to greater prominence of learner-centred approaches. Concurrent with this there have been significant developments in analysing and evaluating collaborative argumentation (Chinn and Clark, 2013), including social (Asterhan and Babichenko, 2014) and cognitive (Chinn et al., 2006) explanations of the benefits of online discussion, transactivity (Ai et al., 2010), critical thinking skills (Garrison, Anderson and Archer, 2001), arguing to learn (Andriessen, 2013), assessing creative collaboration (Dawson, Tan and McWilliam, 2011), and online community development (Haythornthwaite et al., 2000). In order to develop a comment evaluation method of relevance to educators and others involved in MOOC moderation, this thesis explores a number of appropriate pedagogical content analysis methods (i.e. methods that facilitate the evaluation of learners’ critical thinking when discussing course topics).

In Weltzer-Ward's (2011) analysis of 56 content analysis coding schemes used between 2002 and 2010, Community of Inquiry (Garrison, Anderson and Archer, 2010), and analyses adopting Bloom’s taxonomy (Bloom et al., 1956) or the Structure of Observed Learning Outcomes (SOLO — Biggs and Collis, 1982) were identified as widely used methods with high citation counts, accounting for close to 65% of the papers reviewed. They are therefore considered good choices for content analysis in this thesis, and inform the rubrics I developed, following in the tradition of similar work that also adopt these methods.

In addition to these three instruments, a novel content analysis method developed from the Digital Artefacts for Learning Engagement Framework (DiAL-e — Atkinson, 2009) is employed in the pilot phase of this thesis. DiAL-e is a “pedagogically sound” framework (Burden and Atkinson, 2008, p. 1) based on explicitly constructivist and situated learning principles, and developed to support the use of digital video in higher education. It is suitable for adaption and is transferable to other digital learning environments (Burden and Atkinson, 2007) and is thus appropriate for use in this thesis.

2.3.1 Community of Inquiry

Community of Inquiry (CoI) is based on the interaction of forms of engagement or ‘presence’ within Web-based learning communities: cognitive presence, social presence, and teaching presence (Garrison, Anderson and Archer, 2001). As this study looks for evidence of critical
thinking in MOOC forums, the focus is on the cognitive presence dimension, which Garrison et al. define as “critical, practical inquiry” as evidenced within four types of text-based dialogue: triggering, exploration, integration, and resolution (2001, p. 14). These categories refer to stages of dialogue – starting with an initiating ‘triggering’ comment and ending with assertions that resolve the discussion (see Table 1 below).

As an established pedagogic content analysis method, CoI has been applied to many studies. In her paper exploring the application of learning analytics, Dringus asserts that CoI provides “an array of meaningful and measurable qualities of productive learning and communication in online learning environments”, and suggests converting CoI dimensions into datatypes that can be mined to “draw out coherent patterns” (2012, p. 96) in online courses.

Tirado, Hernando and Aguaded (2012) apply CoI in their study on the quality of knowledge construction in social environments, and call for the strong validation of content analysis methods that evaluate the processes of the construction of knowledge in this setting. Shea et al. (2013) adapt the approach to measure students’ successful learning strategies, and compare their results with social network analysis methods. They recognise the importance of further research into the relationship between cognitive presence and interaction, and suggest that its detection contributes to understanding of students’ networking behaviours. Joksimovic et al. (2014) associates linguistic proxies for learning with CoI stages in discussion forums within small scale online courses. Their findings indicate the usefulness of further research that explore the effects of different levels of cognitive presence on learners with different levels of prior-knowledge.

Finally, Waters et al. (2015) implement a machine learning approach to predict students’ critical thinking levels in formal online discussions according to CoI. In their study, they adopt word count, post similarity, chronological order and other features to build a model that achieves a moderate level of accuracy. My thesis adopts Waters et al.’s suggestions for future work that include the use of Linguistic Inquiry and Word Count (LIWC) analysis to identify phases of critical thinking.

The research questions identified by these studies consider the utility of data derived from pedagogical content analysis and its potential in measuring performance in online courses. These are relevant to my thesis, which seeks to contribute to the development of robust and effective ways of understanding large-scale comment data that are based on established theory.

Table 1 presents the categories used by human raters in the pilot and large-scale research projects described in this thesis, and includes descriptions and examples of verbs associated with each category.
### CoI score | CoI descriptor
--- | ---
0 - Off-topic | There is written content, but it is not relevant to the subject under discussion.

1 – Trigger | A contribution that exhibits a sense of puzzlement deriving from an issue, dilemma or problem. Includes contributions that present background information, ask questions, or move the discussion in a new direction. Verbs: evoke, induce, contradict

2 – Exploration | A comment that is seeking a fuller explanation of relevant information. This can include brainstorming, questioning and exchanging information. Contributions are unstructured and may include: unsubstantiated contradictions of previous contributions, different unsupported ideas or themes, personal stories, and descriptions or facts that are not used as evidence. Verbs: inquire, diverge, search

3 – Integration | Previously developed ideas are connected. Contributions include: references to previous messages followed by substantiated agreements or disagreements; developing and justifying established themes; cautious hypotheses; combining different sources; providing a tentative solution to an issue. Verbs: test, conjecture, check

4 – Resolution | New ideas are applied, tested and defended with real world examples. This involves methodically testing hypotheses, critiquing content in a systematic manner, and expressing supported intuition and insight. Verbs: commit, settle, confirm

Table 1: Community of Inquiry: Cognitive Presence
(Garrison, Anderson and Archer, 2001; Park, 2009).

#### 2.3.2 Bloom’s Taxonomy of the Cognitive Domain

The *Taxonomy of Educational Objectives: Handbook 1: Cognitive Domain*, commonly referred to as ‘Bloom’s Taxonomy’ was developed to improve the: “exchange of ideas and materials among test workers, as well as other persons concerned with educational research” (Bloom et al., 1956, p.1), and promote the use of teaching methods that encourage higher order learning. Although directed by a small committee, the Taxonomy resulted from a collaborative effort that gathered feedback from a wide range of educators, educational psychologists, administrators and researchers. Further Taxonomies classifying the Affective and Psycho-Motor domains were published, but did not reach the high level of recognition and use achieved by Handbook 1.

As revised by Krathwohl (2002), Bloom’s Taxonomy consists of a hierarchy of categories of educational goals or outcomes, starting from the lower-order learning goals of ‘remember’ and ‘understand’, to the mid-level uses of knowledge as evidenced in ‘apply’ and ‘analyse’, with ‘evaluate’ and ‘create’ indicating the achievement of deeper understanding. Researchers have found these categories useful as a framework for analysing learning processes. Bloom himself used them to evaluate types of learning that take place in classroom discussions and compared them with lectures (Bloom, 1953). His key finding that learners spend more time engaged in higher order thinking in class discussion than in lectures led him to suggest that increasing opportunities for learner interaction would lead to improved development of problem-solving skills.

Kember’s (1999) association of Bloom’s dimensions with Mezirow's (1991) ‘thoughtful action’ category (e.g. writing); Gibson, Kitto, and Willis (2014) use of Bloom to map word types to levels of cognition, and Karaksha et al.’s (2014) use of Bloom to evaluate the impact of e-learning tools in a...
higher education setting support the use of the Taxonomy in this thesis. Further, in Chan et al.’s (2002) study two raters were employed to analyse essay papers and classroom discussions, applying Bloom, Structure of the Observed Learning Outcomes (SOLO) and the reflective thinking measurement model. Finding strong correlations between the models in long essays, but not in short discussions, they proposed further research using more than two raters to improve agreement and using the revised version of Bloom to improve the accuracy of assessing cognitive learning outcomes. By engaging a team of seven raters in my large-scale research project (see Chapter 4) as well as Krathwohl’s revised version of Bloom (Krathwohl, 2002), I aim to advance understanding of these research issues.

Table 2 presents the categories used by human raters in the pilot and large-scale research projects described in this thesis, and includes descriptions and verb examples associated with each category.

<table>
<thead>
<tr>
<th>Bloom score</th>
<th>Bloom descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – Off-topic</td>
<td>There is written content, but it is not relevant to the subject under discussion.</td>
</tr>
<tr>
<td>1 – Remember</td>
<td>Recall of specific learned content, including facts, methods, and theories. Verbs: name, describe, relate, find, list, write, tell.</td>
</tr>
<tr>
<td>2 – Understand</td>
<td>Perception of meaning and being able to make use of knowledge, without understanding full implications. Verbs: explain, compare, discuss, restate, predict, translate, outline.</td>
</tr>
<tr>
<td>3 – Apply</td>
<td>Tangible application of learned material in new settings. Verbs: show, complete, use, classify, examine, illustrate, implement, solve.</td>
</tr>
<tr>
<td>4 – Analyse</td>
<td>Deconstruct learned content into its constituent elements in order to clarify concepts and relationships between ideas. Verbs: explain, compare, contrast, examine, identify, investigate, categorise, differentiate, organise.</td>
</tr>
<tr>
<td>5 – Evaluate</td>
<td>Assess the significance of material and value in specific settings. Verbs: check, decide, rate, choose, recommend, justify, assess, prioritise, critique.</td>
</tr>
<tr>
<td>6 – Create</td>
<td>Judge the usefulness of different parts of content, and producing a new arrangement. Verbs: synthesise, invent, plan, compose, construct, design, imagine, generate.</td>
</tr>
</tbody>
</table>

Table 2: Bloom’s Taxonomy (Bloom et al., 1956; Chan et al., 2002; Krathwohl, 2002).

### 2.3.3 SOLO Taxonomy

Similar to Bloom, the Structure of Observed Learning Outcome (SOLO) taxonomy (Biggs and Collis, 1982) is an hierarchical classification framework that describes levels of complexity in a learners’ knowledge acquisition as evidenced in their writing. SOLO adopts five categories to distinguish levels of comprehension (Table 3). The Taxonomy has been adopted as a coding scheme in a number of studies. Slack et al.’s study of comments made on a health-related online course noted that shorter comments were difficult to evaluate, but that use of the taxonomy facilitated the identification of in-depth discussion which enabled them to detect weaker and stronger elements of the course (2003). Holmes identified 50% of comments in an online postgraduate course as “high level task oriented” according to SOLO evaluation, the effectiveness of mentors in generating learner interest, and the importance of timely mentor intervention (2005,
Shea et al. (2011) adopt CoI and SOLO to evaluate discussion within online courses, concluding that the frameworks may detect complimentary attributes in student comments.

In addition to facilitating the comparison of SOLO with other frameworks and testing for complimentary attributes, adopting this framework in the pilot stage of my research provides an opportunity to validate the other frameworks’ capacities to distinguish levels of critical thinking.

Table 3 presents the categories used in the pilot research project described in this thesis, and includes descriptions of comment types associated with each category.

<table>
<thead>
<tr>
<th>SOLO score</th>
<th>SOLO descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – Off-topic</td>
<td>There is written content, but it is not relevant to the subject under discussion.</td>
</tr>
<tr>
<td>1 – Prestructural</td>
<td>No evidence of any kind of understanding. Irrelevant information is used, the topic is misunderstood, or arguments are unorganised.</td>
</tr>
<tr>
<td>2 – Unistructural</td>
<td>A single aspect is explored and obvious inferences drawn. Evidence of recall of terms, methods, and names.</td>
</tr>
<tr>
<td>3 – Multistructural</td>
<td>Several facets are explored, but are not connected. Evidence of descriptions, classifications, use of methods and structured arguments.</td>
</tr>
<tr>
<td>4 – Relational</td>
<td>Evidence of understanding of relationships between several aspects and how they may combine to create a fuller understanding. Evidence of comparisons, analysis, explanations of cause and effect, evaluations, and theoretical considerations.</td>
</tr>
<tr>
<td>5 – Extended abstract</td>
<td>Arguments are structured from different standpoints and ideas transferred in novel ways. Evidence of generalisation, hypothesis formation, theorising, and critiquing.</td>
</tr>
</tbody>
</table>

Table 3: SOLO Taxonomy (Karaksha et al., 2014).

2.3.4 Digital Artefacts for Learning Engagement

The least well-established method of analysis in this study, the DiAL-e framework (Atkinson, 2009), was devised to support the creation of pedagogically effective learning interventions using Web-based digital content. It adopts 10 overlapping learning design categories to describe engagement with pedagogic activities, and is pragmatically grounded “in terms of what the learner does, actively, cognitively, with a digital artefact” (Atkinson, 2009). Case study research indicates that practitioners gain value from using the framework (Burden and Atkinson, 2008), and DiAL-e has been adopted as a pedagogical model in studies that evaluate Web-based learning environments (Kobayashi, 2013; O’Riordan, Millard, and Schulz, 2015). In this thesis, I adapt the framework and use it as a comment classification method, whereby comments are rated according to inferred engagement with learning in terms of the DiAL-e categories. Table 4 presents the categories used in the pilot research project described in this thesis, and includes descriptions and illustrative comments associated with each category.
Table 4: Adapted DiAL-e Framework
(Atkinson, 2009).

2.3.5 Discussion on coding schemes used for pedagogic content analysis of online discussions

Weltzer-Ward (2011) argues that understanding of online environments may be improved through the use of pedagogical content analysis methods, and calls for research on their application outside of formal learning settings, as well as the exploration of opportunities for synthesis. My thesis addresses these research areas by using different methods that adopt complementary approaches to analyse discussion in large scale, informal settings. While Garrison, Anderson and Archer (1999) acknowledge the consistency of their CoI framework with socio-constructivist learning theory emerging from Dewey’s (1897) ideas on the importance of sociological as well as psychological aspects of learning, and Biggs and Collis recognise the influence of Piagetian stages of development on their levels of learning quality (Biggs and Collis, 1982), Bloom et al., do not explicitly recognise a primary theoretical basis. However, the authors of these three frameworks adopt hierarchical approaches identifying changes in learners’ behaviour that have much in common with Piaget’s theory of staged development (1959), DiAL-e’s focus on evaluating pedagogical affordances of
emerging technologies draws from Piaget’s ideas on adaption (Burden and Atkinson, 2010), and all four frameworks implicitly recognise the value of social learning explored by Vygotsky (1978).

While there are similarities, there are also distinct differences in their focus as well as approaches to the evaluation of critical thinking: Bloom facilitating generalisable assessments of educational outcomes that can be applied to evaluating learners in any number of settings, CoI focusing on appraisal of participation in CSCL environments, SOLO evaluating the quality of learners’ responses, and DiAL-e’s assessing learning design categories inferred from comments. In addition, some educational psychologists argue that individual and distributed cognition are two distinct, interrelated processes (Moore and Rocklin, 1998), and the methods adopted in this study emphasise these different aspects: Bloom and SOLO – individual, and CoI and DiAL-e – distributed.

As I use these methods to identify levels of critical thinking in comments, it is valid to assess the value of these methods used individually, and reflect on possibilities for combining methods, as well as to consider the practical use of the outcomes of my research. This raises a potential conflict when balancing a methods-centred approach to my research, with a problem-centred consideration of the practical issues involved. As Maslow observes, methods-based research can provide more elegant results, but may restrict interpretations, and divert attention from resolving real problems (Maslow, 1966). Further, Mayer cautions that use of particular analysis methods tends to focus on the frameworks’ perspective to the impairment of other viewpoints. In addition, there is a possibility that evaluations may be unhelpfully interpreted according to a framework’s perspective simply to find and justify meanings associated with the framework, with the outcome that analyses of comments may relate more to the framework than to the actual meaning of the comment (Meyer, 2004). In practical terms, it is important to consider if a MOOC moderator managing high volumes of comments finds classifying a ‘Resolve’ statement (according to CoI) more useful than labelling a ‘Create’ comment (according to Bloom), or if a general indication of critical engagement with a topic is of greater use in facilitating actionable feedback.

By comparing distinct approaches to measuring critical thinking I seek to establish if different methods yield significantly different results, and identify opportunities for synthesis. This leads to my first research question (RQ1): are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forum comments? In particular, can different people consistently apply them, and do different frameworks identify the same types of learning activity?

In this thesis, MOOC forum comments are evaluated in terms of the extent to which they provide evidence of deep learning approaches that reflect critical thinking through the lens of each method and attempt to identify significant differences in outcome from their use. Specifically, I seek to establish if these coding schemes used for pedagogical content analysis identify the same types of learning activity, and in the large-scale study, if CoI and Bloom can be applied consistently by
different people, with the aim to establish a robust means of measuring levels of critical thinking in informal comments.

In addition to comparing and critically evaluating different coding schemes used for pedagogical content analysis, I compare the outcomes of this analysis with established proxies for learning in the form of linguistic analysis and typical interaction analysis. These are explored in the next two sections.

### 2.4 Language Analysis

Language analysis is a branch of discursive psychology, which typically involves the study of a collection of text data (referred to as a corpus). Language is viewed as “a medium of action” (Potter and te Molder, 2005, p. 3) and recurring themes which are seen to be characteristic of particular forms of communication are explored (Stubbs, 2001). In terms of this thesis and methods for analysing cognitive activity, texts are not considered externalisations of inner states (which cannot be reliably inferred), but are regarded as subjective interpretations of meaning in context (which may be validated) (Sanders, 2005).

The content and style of language used in everyday communications can provide important indicators of psychological and social meaning which may be measured using quantitative methods, including thematic analysis and word pattern analysis (Pennebaker, Mehl and Niederhoffer, 2003). Characteristic approaches to quantitative language analysis involve the identification and coding of similar patterns and the interpretation of content supported by statistical tests of significance (Khawaja, Chen and Marcus, 2010). Grammatical features, lexical categories, punctuation and word counts are seen as important proxies for emotional and cognitive processes (Pennebaker and Francis, 1996).

A recent, increased emphasis of content analysis studies exploring language use and discussion in CSCL has been identified (Weltzer-Ward, 2011). For example, Delfino and Manca (2007) discuss the use of ‘figurative’ language in online social contexts, Miller (2004) explores gender-related language patterns in CSCL, Uzuner (2007) identifies educationally valuable talk in CSCL, Tausczik and Pennebaker (2013) adopt a real-time language feedback system to improve learner collaboration, Robinson, Navea and Ickes (2013) correlate student language use with educational attainment, and Allen, Snow and Menamara (2015) employ linguistic indicators to predict learners’ reading comprehension abilities. Evidence that pedagogically meaningful dialogue in Web-based environments can be automatically identified using learning analytic techniques has importance for my thesis (De Liddo et al., 2011), as does the use of mixed linguistic and interactional data to identify potentially ‘at risk’ learners (Wen, Yang and Rosé, 2014b).
2.4.1 Linguistic Inquiry and Word Count

While content analysis has historically been undertaken by human annotators using widely available research tools (e.g. QSR Internation Pty Ltd, 2017), in recent years, substantial developments in automated language analysis has been made which appear to demonstrate significant correlations between human raters and automatic tools. Among these computational approaches to language analysis, Linguistic Inquiry and Word Count (LIWC) (Francis and Pennebaker, 1993; Pennebaker Conglomerates Inc., 2015) has emerged as a significant tool for analysis of online discussion, and evaluation of cognitive processes. LIWC2015 software was developed from studies into the therapeutic effects of writing about psychological traumas (Pennebaker, Colder and Sharp, 1990; Spera, Buhrfeind and Pennebaker, 1994). The application adopts a quantitative, word count approach (as distinct from qualitative thematic, or word pattern analysis) which aims to reveal the psychological meaning of words taken out of context from their original settings (Pennebaker, Mehl and Niederhoffer, 2003). The application searches within text files for over 2,300 words or word stems, tracking stylistic aspects of language use classified into 82 dimensions (e.g. articles, prepositions, pronouns), psychological processes (e.g. positive and negative emotion categories, cognitive processes), and other categories, including punctuation.

LIWC2015 processes text by counting words and calculating the percentage of each of its categories, so that for example, an individual text file may be said to contain 2.34% pronouns, 3.33% verbs, and so on. The word categories are derived from the programs’ internal dictionaries that contain collections of words that have been assigned to the categories by expert evaluators (Tausczik and Pennebaker, 2010). Once a text has been processed, the application produces a list of categories and percentages as a tab-delimited text file that can be directly read in application programs, such as IBM SPSS Statistics (SPSS) or Microsoft Excel.

In addition to the developers’ experiments aimed at validating the program and demonstrating its potential for discovering, for example, attentional focus, emotional status and thinking styles in word use (Tausczik and Pennebaker, 2010), there are a number of studies which suggest the usefulness of LIWC in detecting the meaning of words. While not seen as a replacement for qualitative analysis, Carroll (2007) found an earlier version of the program, LIWC2007, “easy to use” (p. 226) in his analysis of essays written for a critical thinking course. He found learners’ final essays demonstrated less use of pronouns and words related to insight (think, know, consider), discrepancy (should, would, could), and tentativeness (maybe, perhaps, and guess), and were more likely to express causal thinking (because, effect, hence), compared with their earlier writing.

2.4.2 LIWC Categories

In this thesis, all correlations with LIWC categories were explored, however research indicates that specific categories are more closely associated with critical thinking than others. Therefore, as I set
out to answer the question on whether the linguistic characteristics of comments are reliable proxies for depth of learning (RQ2), I expected to find significant associations between the results of our pedagogical analysis and a number of LIWC characteristics.

### 2.4.2.1 Word Count

The number of words used in comments is often understood as a rough guide to levels of participation (Asterhan and Babichenko, 2014) and is commonly associated with intensity of engagement (Sexton and Helmreich, 2000; Darabi et al., 2011). Ferguson and Buckingham Shum's (2011) research into synchronous text chat, and Joksimovic et al.'s (2014) linguistic analysis of online discussions similarly suggest close associations between high word counts and thoughtful, ‘exploratory’ exchanges.

### 2.4.2.2 Pronouns

In their comparison of self-assessment with traditional (non-reflective) assignments, Peden and Carroll (2008) found that learners writing self-assessment essays included more pronouns, insight and emotion words and used simpler language than expressed in traditional academic assignments. Kacewicz et al. (2014) suggest that high status contributors use fewer first-person singular pronouns, and Vosecky, Leung and Ng’s (2012) research into tweet quality suggests that ‘I-talk’ signifies ‘low quality’, non-factual communication. Robinson, Navea and Ickes (2013) discovered that they could predict learners’ course performance on their use of “word simplicity, first-person singular pronouns, present tense, details concerning home and social life, and words pertaining to eating, drinking, and sex” (p. 469), concluding that low-performing learners tended to exhibit egocentricity in their writing.

### 2.4.2.3 Causal Words

Within LIWC dictionaries, causal words are categorised as a sub-group of cognitive process words, which suggest engagement with active reappraisal, or processing of information (Pennebaker, Mayne and Francis, 1997). While Joksimovic et al. (2014) found counts of causal words were not significant between higher phases of CoI, several studies (e.g. Pennebaker, Mayne and Francis, 1997; Creswell et al., 2007; Ferguson and Buckingham Shum, 2011) have found that causal words are related to the level of cognition. Linguistic analysis of journals and essays indicates that causal words are more evident in precise and concise descriptions, and indicate progress in the level of cognition and understanding (Lengelle et al., 2013; Spera, Buhreinf and Pennebaker, 1994). In addition, increased levels of differentiating between competing ideas has been linked to higher levels of cognition (Tausczik and Pennebaker, 2013).
2.4.2.4 Power and Affiliation

The LIWC categories of power and affiliation, are developed from thematic apperception test (TAT) research and relate to assessments of an individuals’ unconscious drives and social motives, where the affiliation motive is related to friendliness and establishing rapport, and power is associated with making an impact and exerting control (Winter et al., 1998). In LIWC, higher incidence of words associated with power suggests the writers’ perception of themselves as having high status or expertise.

2.4.2.5 Emotion Words

Using sentiment analysis to measure relationships between mood and different variables – from consumer confidence to managing disaster relief – is commonplace wherever people’s behaviour is under scrutiny. Research suggests that while positive language can imply focus on group cohesion, which may encourage individuals to work harder (Winter et al., 1998; Lebie, Rhoades and McGrath, 1995), it has been noted that correlation with positive sentiment can suggest disconnection, and that high levels of empathetic discussion may distract learners from key tasks (Leshed et al., 2007). Conversely, the expression of negative sentiment has been associated with ‘cognitive disequilibrium’ and higher levels of critical thinking (Leshed et al., 2007; Schwarz, 2000).

2.4.2.6 Word length

Whereas complex cognitive processes and critical thinking are often associated with using long words (Spera, Buhrfeind and Pennebaker, 1994; D’Mello and Graesser, 2012), some researchers have found that counts of long words are not significant indicators of cognitive load, but have use in supporting analysis that include other significant features (Khawaja et al., 2009). In addition, long sentences do not necessarily suggest increased cognitive attention. In their research into predictors of students’ reading comprehension Allen et al. assert that shorter sentences can suggest more sophisticated writing strategies (2015).

2.4.2.7 Other Word Types

Researchers have found negation, auxiliary verbs, and conjunctions to be significant indicators of cognitive load (Khawaja, Chen and Marcus, 2010), and in analysis of undergraduate writing these categories have shown significant differences between triggering and other phases of CoI (Joksimovic et al., 2014). Research also indicates a high incidence of prepositions associated with attention to reflective behaviour. High use of prepositions are identified as significant indicators of increased cognitive load (Khawaja et al., 2009), and their prevalence in the discussion sections of journal articles which are “often the most complex part of an article” (Allen, Snow and McNamara, 2015, p. 35).
In addition, Joksimovic et al.’s (2014) study found distinct use of dictionary, functional, inhibition, inclusive and cognitive words, as well as articles, prepositions, and conjunctions in the triggering phase of CoI, but found no significant difference in the use of pronouns or insight words throughout the four phases.

### 2.4.2.8 Limitations

Although supported by numerous research outputs, linguistic analysis is limited in its reliability (Tausczik and Pennebaker, 2010). The percentages of word characteristics per comment calculated by LIWC2015 are not useful in brief messages containing fewer than 50 words (Pennebaker Conglomerates Inc., 2017), and as well as uncertainty over the meaning of higher numbers of words per sentence counts, as referred to above, analysis of word categories may also be compromised. Contributors to discussion forums often use symbolic, oblique and indirect ways of communicating meaning which may lead to classification errors (Pennebaker and Francis, 1996). Multiple meanings of words, complicated sentence formation, and unclear use of pronouns, may obscure meaning and require more complex methods to resolve uncertainty than are available in the software used in this study (Indurkhya and Damerau, 2010). However, notwithstanding the potential for error, I am in agreement with Pennebaker and Francis’ claim that LIWC analysis is “as valid as a judge-based system that requires multiple judges who, themselves, are prone to error” (Pennebaker and Francis, 1996, p. 622).

### 2.5.2.9 Other linguistic analysis methods

Linguistic analysis is used in many other fields in addition to identifying pedagogically meaningful dialogue in online environments. Disciplinary studies research suggests that different disciplines adopt different rhetorical styles in academic writing, and have identified variations in the use of linguistic categories that are similar to LIWC categories, but which are characteristic of different disciplines rather than cognition (Bazerman, 2002; Fløttum, Dahl and Kinn, 2006; Hyland, 2007).

In his study of leading journals from eight disciplines (four humanities and social sciences, and four natural sciences journals), Hyland (2007) classifies different behaviours which writers from different disciplines adopt when representing themselves, their work, and how they address their readers. He argues that rhetorical differences can be observed in two key linguistic features: authorial ‘stance’, and ‘engagement’. Stance comprises of ‘hedges’ (linguistic features expressing uncertainty), ‘boosters’ (linguistic features expressing certainty), attitude markers (opinions), and self-mention (personal pronouns). Engagement includes reader pronouns (e.g. *we, you*), directives, personal asides, appeals to shared knowledge, and questions. Hyland asserts that there are significant differences between what he refers to as the ‘soft’ and ‘hard’ sciences, with hedges, boosters, self-mentions, reader pronouns and questions tending to be more common in ‘soft’ humanities (e.g. economics, history, education) and social science papers than in ‘hard’ natural sciences (e.g. physics, engineering, life sciences) (Becher and Trowler, 2001; Hyland and Diani, 2009).
As this discipline-specific tendency adopts linguistic features that are also associated with levels of critical thinking (e.g. pronoun use), and to avoid potential bias, secondary analysis is carried out in order to identify diverse discipline-specific language characteristics. While it is beyond the scope of this thesis to undertake a full evaluation following Hyland and Diani (2009), five of Hyland’s rhetorical features can be closely mapped to LIWC categories (Table 5).

<table>
<thead>
<tr>
<th>Hyland’s rhetorical features</th>
<th>LIWC categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedges</td>
<td>Tentative + Discrepancy</td>
</tr>
<tr>
<td>Boosters</td>
<td>Certainty</td>
</tr>
<tr>
<td>Self-mention</td>
<td>Personal pronoun (ppron)</td>
</tr>
<tr>
<td>Engagement markers</td>
<td>You + We + Qmark</td>
</tr>
</tbody>
</table>

Table 5: Hyland’s rhetorical features mapped to LIWC2015 categories.

2.5 Interaction analysis

Where use of language acts as an unconscious indicator of mood, interaction analysis explores more direct actions. For example, ‘likes’ are a common intentional rating mechanism often adopted by social media platforms to enable users to rapidly signify personal feelings (Indurkhya and Damerau, 2010; Qi et al., 2012). Some research suggests that this metric is ambiguous (Leibowitz, 2013) and unreliable (Matthews et al., 2015), and my early research produced insignificant results for this metric (O’Riordan, Millard and Schulz, 2015). However, this indicator as well as sentiment analysis are widely used in learning analytics, for example to identify learner attrition (Wen, Yang and Rosé, 2014), self-confidence (Siemens, 2012a), and learners’ opinions of courseware (Song, Lin and Yang, 2007). In addition, there is some evidence that this cumulative rating system may provide learners with timely prompts that can lead to higher levels of learning (Darabi et al., 2011), and FutureLearn, the platform hosting the MOOCs explored in the large-scale research project reported in this thesis, provides a ‘like’ button feature associated with all individual comments which allows learners to provide immediate, simple feedback.

Further, by placing discussion forums within the context of each activity and providing mentor support (León et al., 2015), the platform encourages sharing and situated debate (Ferguson and Sharples, 2014), with the explicit intent of building communities of inquiry and inspiring higher-level learning. In this context, I set out to discover if learners’ use of the ‘like’ feature was significantly associated with coding schemes used for pedagogical content analysis. Specifically, I aim to answer the question of whether the number of likes awarded to comments or the sentiment of posts are a reliable indicator of critical thinking (RQ3).
2.6 Machine Learning

A key approach to automating analysis in Automatic Essay Scoring, Pedagogic Content Analysis and Linguistic Analysis is Machine Learning (ML). The field has emerged from a multidisciplinary combination of philosophy, biology, statistics, cognitive science, information theory, and artificial intelligence. ML techniques approach analysis as a search problem that applies relevant domain-specific search strategies with the aim of creating computation methods that automatically improve with experience (Mitchell, 1997). The essential aim of supervised ML is to automatically categorise large amounts of unlabelled data (i.e. to generalise) using a small number of human labelled data (the ‘training set’) as a template, essentially acting as a “knowledge lever [turning] a small amount of input knowledge into a large amount of output knowledge” (Domingos, 2012, p. 81).

Applications of ML include, predicting consumer behaviour (Yaeli et al., 2014), automatic fraud detection (Akoglu and Faloutsos, 2013), product recommendation (Li and Chen, 2013), driverless cars (Stavens and Thrun, 2006) and personalised learning (Smith, Davies and Evans, 2013).

Research in Automatic Essay Scoring and Learning Analytics suggests that Machine Learning techniques are effective in automatically identifying patterns of pedagogical activity from texts (Forbes-Riley and Litman, 2009; Hassan and Mihalcea, 2011; Kovanovic et al., 2016; Turnitin, LLC, 2015).

The central computation core of any ML process is the algorithm it uses, and various types have been designed to manage different learning tasks, including: Naïve Bayes, a classification system based on statistical analysis, used by Hassan and Mihalcea (2011) to establish effective search methods for finding Web-based educational resources; the K-Means Cluster Analysis method, adopted in Hogo’s (2010) study evaluating activity in e-learning settings; and Support Vector Machines (SVM), which achieved highly accurate results in Mayfield and Rosé's (2011) research into identifying authoritative dialogue in CSCL. However, among the large number of available classifiers, Random Forest appears to offer opportunities for high accuracy and reliability (Caruana and Niculescu-Mizil, 2006; Fernández-Delgado et al., 2014).

Random Forest classifications are recognised as accurate and relatively straightforward to apply (Walker, 2013), they are among the most widely adopted classifiers in Learning Analytics research (Siemens and Baker, 2012), and a Random Forest approach is adopted by Kovanovic et al. (2016) to automatically classify phases of cognitive presence in distance learning discussion forums, achieving a high classification accuracy of 70.3% and Cohen’s kappa of 0.78. In addition, this method is supported in the open source WEKA (Waikato Environment for Knowledge Analysis) machine learning ‘workbench’ application (Witten, Frank and Hall, 2011).

The four key elements to designing a ML program are: the training set (e.g. a collection of comments), the target function (e.g. correctly rated comments), the learned function (e.g. pedagogical and language content analysis methods) and the learning algorithm (e.g. Naïve Bayes,
k-means, SVM, Random Forest) (Mitchell, 1997). These steps require the development of well-formed hypotheses aimed at answering clearly defined questions, and the creation of human rated datasets that have assigned specific scores to specific states (e.g. values for CoI, Bloom, and linguistic features of text), in order to ‘train’ the learning algorithms. Algorithms are then applied to the datasets, and outputs assessed for validity by human judges.

In 2013, the limitations of ML methods were highlighted by the well-publicised failure of Google Flu Trends. This machine approach, designed to predicting outbreaks of influenza using outputs from Google’s search algorithms, failed to identify the peak of the 2013 ‘flu season by 140 percent. Lazer et al. (2014) argue that this was caused by Google’s algorithm vulnerability to ‘overfitting’ to unrelated terms, and not recognising changes in human behaviour (brought about by alterations to Google search design). As well as avoiding the pitfalls of overfitting (where the statistical model includes too many irrelevant correlations), they suggest that designers of ML approaches should use a variety of appropriate data sources, not just readily available large datasets.

2.6.1 Data processing – some ethical considerations

While my research has been carried out on anonymised data, where comment authors are not identified, and where learners have given prior consent for their data to be used for academic research (FutureLearn, 2017a), my ultimate aim is to develop a means of automatically processing personal data (i.e. learners’ comments) in order to enable MOOC mentors to intervene at an individual learner level. This has implications for future development of my work.

Public concern regarding the unethical use of personal data was highlighted in 2014 by Facebook’s ‘contagion study’ (Kramer, Guillory and Hancock, 2014), where the newsfeeds of more than 600,000 Facebook users were deliberately altered, and their emotional responses measured. Kramer, Guillory and Hancock claimed that informed consent had been obtained from all participants, as they had agreed to Facebook’s Data Use Policy prior to joining the network. However, as well as the negative reaction from Facebook users to their uninformed participation (Griffiths et al., 2016), bioethics scholars assert that this research activity was conducted unethically, as specific, informed consent had not been obtained, and that independent ethics oversight was not in place (Kleinsman and Buckley, 2015).

The Learning Analytics Community Exchange’s (LACE) work on the ethics of data processing in support of learning stresses the complex range of issues involved in protecting personal rights in this environment. Among the many ambiguities thrown up by the application of learning analytics to teaching practice, LACE researchers assert that the opportunistic nature of the field, which combines data sources in unintended ways, makes the collection of informed consent problematic, raising concerns about the potential for unintended bias, and unfair profiling of individual learners (Griffiths et al., 2016). In terms of my work with comment data, although prior consent for use of
personal data for academic research has been obtained, and this work has been carried out with oversight from a University Ethics Committee, application of its outputs in a real learning environment requires further attention to data protection concerns.

Many of these concerns are addressed in the European Union’s General Data Protection Regulation (GDPR), which is due to come into force in 2018. The new regulations place specific restrictions on automated decision-making, and gives greater emphasis to individual rights, including rules governing the giving and withdrawal of informed consent, as well as rights to access to personal data and information on decision-making processes. While the GDPR has wide implications for data controllers and processors, it has specific consequences for MOOC providers who wish to use personal data to improve the learner experience, as well as data scientists engaged in designing algorithms for use in this setting.

In particular, the GDPR requires that individuals should not have decisions made about them that produce a legal, or similar, effect, and which are based on exclusively automated decisions that have not been subject to human mediation. People will gain the right to human mediation in automated decision-making systems, to give their point of view, and obtain an explanation of the decision, and challenge it. In addition, appropriate mathematical and statistical procedures should be used to minimise errors, and prevent discriminatory effects (Council of the European Union, 2016).

Although the ‘legal or similar’ effects of the automated decisions resulting from my research are uncertain, my belief is that the outcomes of the computational method reported in this thesis fall short of this test. However, on completion of a Privacy Impact Assessment (Morlière, 2015), MOOC providers would be able to put in place measures that provide human intervention to the automated-evaluations provided by the computational methods described in this thesis, and facilitate arbitration processes to manage individual concerns and challenges.

Of greater significance to my study is the ‘right to an explanation’, and the implication that automatic decision-making should be transparent and readily understood (Goodman and Flaxman, 2016). By adopting well-known pedagogical analysis methods and using robust approaches to identify correlations between these analyses and the linguistic features of comments, I have taken substantial steps to ensure that the automated decisions described in this thesis are intelligible to potential users.

2.7 Summary

The study of relevant literature reveals that the analysis of cognitive activity in Web-based environments is a complex and evolving area which is driven by desires to enhance monitoring as well as learner engagement. The key to understanding how people learn is through examining the
process of learning, and those engaged in Learning Analytics use Web-based proxies as evidence of knowledge, competence and learning.

Analysing learner interaction and providing feedback mechanisms have the potential to disrupt learning on the Web, with far reaching, but little understood outcomes. In addition, what metrics to use, how to automate analysis, and how to provide outputs in order to bring about the improvements indicated by current research is uncertain. However, while no approach can accurately show what is going on in a person’s brain while they are learning, the traces people leave in comment data may indicate levels of critical thinking. Therefore, as well as counting mouse clicks and other proxies for attention, Learning Analytics is also interested in identifying cognitive activity within online discussion forums. Research indicates that automatic decision-making is a promising field of research that adds value to human efforts. An important focus of this thesis is to report on research studies that extract meaningful measurements suggestive of critical thinking from comment data, which may be used to provide actionable feedback within MOOCs.

The next chapter describes the research strategy, data collection techniques and conclusions on the pilot stage of a research project to identify attributes which may be used to design an algorithmic approach to evaluating levels of learners’ critical thinking in discussion forum comments.
Chapter 3: Pilot stage – Content Analysis of Portus MOOC Comments

This initial pilot stage of my thesis investigates the reliability and viability of four coding schemes used for pedagogical content analysis in identifying levels of critical thinking in a MOOC discussion forum, and explores correlations between these measures and linguistic and interaction characteristics. Following earlier work suggesting the usefulness of further research into rating based on pedagogical frameworks (O’Riordan, Millard and Schulz, 2015), the purpose of this stage of the overall study is to make an initial small-scale test of my assumptions and scope the potential for answering the study’s overall research questions in advance of undertaking a wide-ranging investigation.

3.1 Methods

In order to explore possible correlations between inferred levels of critical thinking (from scores derived from pedagogical content analysis) with typical measures of engagement, the level of intentional user-feedback (‘likes’ per comment), and linguistic features (e.g. word count, pronoun usage, sentiment) were also evaluated.

3.1.1 Data collection and analysis

An anonymised dataset derived from comment fields associated with ‘The Archaeology of Portus’ MOOC offered on the FutureLearn platform in June 2014 was used in this pilot study. More than 20,000 asynchronous comments, generated by nearly 1,850 contributors (both learners and educators) occurred within the comment fields of each of the 110 learning ‘steps’ offered during the six-week duration of the course.

Qualitative and quantitative pedagogical content analyses were undertaken manually on a purposive sample of 600 comments (the MOOC2014 corpus), using four different methods. Qualitative analysis comprised of assessing and rating the MOOC2014 corpus using a content analysis scheme developed from the Community of Inquiry (CoI) Cognitive Presence dimension (Table 1), as well as thematic analysis schemes developed from Bloom’s Taxonomy (Table 2), SOLO Taxonomy (Table 3) and the DiAL-e Framework (Table 4). In total, each entire comment was rated four times using all methods, with a seven-day interval between assessments using each method.
De Wever et al. (2006) argue that reliability is the primary test of objectivity in content studies, where establishing high replicability is important. Intra-rater reliability (IRR) tests were undertaken on 120 comments selected from the MOOC2014 corpus using a random number generator (Haahr and Haahr, 2015). This sample represents 20% of the total and follows typical IRR sample selection practice (De Laat and Lally, 2004). Similar to inter-rater reliability, which measures the degree of agreement between two or more raters, intra-rater reliability quantifies the level of agreement achieved with one rater assessing a sample more than once, after a period of time has elapsed. While not as robust as methods employing multiple raters, testing for intra-rater reliability is viewed as an early stage in establishing replicability, provides an indication of rater stability (Rourke et al., 2001), and is an appropriate measure for this small-scale, pilot study.

Many different indicators are used to report on IRR, including percent agreement, Cohen’s kappa, Krippendorff’s alpha, and Intraclass Correlation Coefficient (ICC). Percent agreement is a simple and popular test that is derived from the result of the proportion of ratings that are agreed upon and the total number (the sum of agreed and disagreed) of ratings. Its main drawbacks are that it does not take into account agreement by chance or rating that is close to agreement (Cohen, 1960; Krippendorff, 2004a). Unlike Cohen’s kappa, which is more appropriate for use with nominal data, the latter two methods may be applied to ordinal variables (Krippendorff, 2004a; Hallgren, 2012). Because the content analysis dimensions used in this study may be considered as a system of ranking comments in order of status, they can be classed as ordinal, rather than nominal variables (which treat data as discrete unrelated values). However, while this distinction is recognised by researchers in the learning analytics field, values for Cohen’s kappa are typically reported in learning analytics literature when reporting outputs from machine learning experiments (Dr V Kovanović, 2017, pers. comm., 14 July). Therefore, where appropriate, and for completeness, different reliability measures are reported in this thesis. ICC is reported for outputs from the comment rating research projects, and both ICC and Cohen’s kappa are reported in the ML section of this study.

While acknowledging the essentially arbitrary nature of benchmarking, Landis and Koch have suggested the following standards for interpreting kappa: 0.41 – 0.60 suggest ‘moderate’ agreement, 0.61 – 0.80, ‘substantial’ and 0.81 – 1.00, ‘almost perfect’ agreement (1977, p. 165). On the other hand, Krippendorff (2004b) suggests that values below 0.67 should indicate non-agreement and be discounted, while results above 0.8 may form the basis for definite conclusions. In the case of ICC used to evaluate agreement in psychological assessments, Cicchetti (1994) argues that, when the reliability coefficient is below 0.40, the level of clinical significance is poor, between 0.41 and 0.59 it is fair; between 0.60 and 0.74, good; and above 0.75, excellent.

To provide a comprehensive account of reliability, in this pilot study, all four measures of reliability commonly used in content analysis research (De Wever et al., 2006) were calculated using IBM SPSS Statistics predictive analysis software (SPSS) (Table 6): percent agreement, Cohen’s kappa,
Krippendorff’s alpha, and ICC. As expected, treating content analysis dimensions as nominal variables produce moderate or unacceptable levels of agreement using percent agreement and Cohen’s kappa approaches, but treating the dimensions as ordinal produces more convincing results (e.g. results for Bloom: $\alpha = 0.897$, ICC = 0.951).

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Percent agreement</th>
<th>Cohen’s k (nominal)</th>
<th>Krippendorff’s $\alpha$ (ordinal)</th>
<th>Interclass Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community of Inquiry</td>
<td>65.0%</td>
<td>0.538</td>
<td>0.831</td>
<td>0.900</td>
</tr>
<tr>
<td>SOLO</td>
<td>69.2%</td>
<td>0.585</td>
<td>0.867</td>
<td>0.928</td>
</tr>
<tr>
<td>Bloom</td>
<td>58.3%</td>
<td>0.492</td>
<td>0.897</td>
<td>0.951</td>
</tr>
<tr>
<td>DiAL-e</td>
<td>59.2%</td>
<td>0.497</td>
<td>0.859</td>
<td>0.918</td>
</tr>
</tbody>
</table>

Table 6: Intra-rater reliability values.

### 3.1.2 Procedure and analysis

Quantitative analysis comprised a number of rating activities leading to statistical analysis. Comment data was manually rated and appropriate software was used to automate those parts of the procedure that required consistent and repeatable approaches to data search and numerical calculation.

#### 3.1.1.1 Data consolidation

In an effort to locate typical comment threads, a purposive sampling approach was chosen (Silverman, 2013). Six steps were selected based on their closeness to average word and comment count, and where less than 5% of comments were made by the most frequently posting contributor. A further six steps were selected: three with the highest number of comments and highest word count and three with the lowest number of comments and lowest word count.

#### 3.1.1.2 Count and categorise pedagogic activity

The first 50 comments from each of these 12 steps were then rated, amounting to 600 out of the total 20,253 comments – a 3% sample. Rating for Bloom, SOLO and CoI was based on the application of values to whole comments, whereas the adapted DiAL-e model used aggregated scores derived from the number of DiAL-e category examples observed in each comment. For example, a comment rated using DiAL-e may contain a question, an indication of research activity (e.g. a hyper-link to a relevant resource), and a statement supporting a previous comment. This would result in an aggregated Pedagogic Score (PS) of three – one for inquiry, one for research and one for collaboration.
3.1.1.3  Collect and analyse typical learning measures

The number of ‘likes’ per comment were counted, and sentiment data for each comment (calculated using LIWC2007 software) were aligned with each comment and analysed using SPSS.

3.1.1.4  Language analysis

LIWC2007 analysis of all 70 word count and linguistic categories was carried out on aggregated rated comments for each step, and on individual comments containing 100 words or more.

3.1.1.5  Correlate Pedagogic Scores for each method

The Pedagogical Scores for each content analysis method were analysed using IBM SPSS software.

3.1.2  Language analysis

Language analysis typically involves selecting a ‘corpus’ containing an arbitrary number of words from a collection. Yates (1996) asserts that in Computer Mediated Communication research an arbitrary cut-off of between 2,000 to 5,000 words is in keeping with established methods, however Banko and Brill (2001) suggest that larger samples are more reliable. Because this study explores comments, in context and in detail, the MOOC2014 corpus comprises a relatively large sample – some 33,648 words.

The linguistic qualities of the MOOC2014 corpus were calculated using LIWC2007 software. Due to the explorative nature of this study, all 82 categories in LIWC’s internal dictionary 2007 were selected, with no text segmentation (Pennebaker Conglomerates Inc., 2015). When this study was carried out, the advice from the application’s developers was that LIWC was designed to analyse texts containing 100 words or more, however this was revised in later advice to 50 words or more (Pennebaker Conglomerates Inc., 2017). Because many of the comments contain short interjections with low word counts, it was decided to analyse the corpus in two ways. Firstly, text files containing aggregated rated comments from each of the steps included in this study were created and analysed. Then, individual files for each comment containing 100 or more words (93 comments in total) were created and analysed separately from aggregated comments.

Ethical clearance for this study was granted by the University of Southampton’s Faculty of Physical Science and Engineering Ethics Committee (ERGO/FPSE/10898).

3.2  Findings

3.3  A total of 600 comments were rated according to four different coding schemes, resulting in 2,400 ratings being carried out. The purposive sampling technique used in this study resulted in an unequal distribution of comments across the 6 weeks of the course. This
meant that no comments from week three were rated, and the same number of comments in the final week of the MOOC were rated as in the first two weeks (Figure 1).

![Weekly distribution of sample comments](image1.png)

Figure 1: Weekly distribution of sample comments.

Comment rating resulted in an uneven distribution of scores, with a large number of 0 rated (‘off-topic’) comments being recorded in all methods, with the exception of those rated according to DiAL-e (Figure 2).

![Total pedagogical rating for Portus MOOC as per schemas (Tables 1-4)](image2.png)

Figure 2: Total pedagogical rating for Portus MOOC as per schemas (Tables 1-4).

<table>
<thead>
<tr>
<th>Rating/schema</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Col</strong></td>
<td>Off-topic</td>
<td>Triggering</td>
<td>Exploration</td>
<td>Integration</td>
<td>Resolution</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SOLO</strong></td>
<td>Off-topic</td>
<td>Prestructural</td>
<td>Unstructural</td>
<td>Multistructural</td>
<td>Relational</td>
<td>Extended Abstract</td>
<td></td>
</tr>
<tr>
<td><strong>Bloom</strong></td>
<td>Off-topic</td>
<td>Remember</td>
<td>Understand</td>
<td>Apply</td>
<td>Analyse</td>
<td>Evaluate</td>
<td>Create</td>
</tr>
</tbody>
</table>

DiAL-e ratings are aggregates of scores derived from DiAL-e as shown in Table 4.

While the definition of off-topic is unsettled, the occurrence of this form of discourse in this study (e.g. 35% of all comments rated according to SOLO) has similarities to findings in other MOOC studies (e.g. Wang et al., 2015; Huang et al., 2014). In addition, the idiosyncrasies of the small purposive sample collected for this exploratory study resulted in a high proportion of comments...
from the start of the course being rated (25%). In common with other MOOCs, at this stage of the Portus MOOC, learners were encouraged to introduce themselves and discuss their personal interests and background – subjects which may not focus on the main topic of the course.

In an attempt to account for potential influence of discipline-specific language on language analysis aimed at identifying levels of critical thinking, an abbreviated evaluation was undertaken by mapping five of Hyland's (2007) rhetorical features to the LIWC categories listed in Table 5. Hyland (2007) claims that there are significant differences between ‘soft’ and ‘hard’ sciences, with hedges, boosters, and engagement markers (e.g. reader pronouns and questions) tending to be more frequent in ‘soft’ humanities discourse than in ‘hard’ natural sciences, as demonstrated in Tse and Hyland's (2009) study of disciplinary variation in reviews of academic books (Table 7).

<table>
<thead>
<tr>
<th></th>
<th>Philosophy (soft)</th>
<th>Biology (hard)</th>
<th>MOOC2014 corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedges (tentative + discrepancy)</td>
<td>2.24%</td>
<td>1.67%</td>
<td>5.47%</td>
</tr>
<tr>
<td>Boosters (certainty)</td>
<td>1.69%</td>
<td>1.35%</td>
<td>1.25%</td>
</tr>
<tr>
<td>Engagement markers (you + we + qmark)</td>
<td>3.19%</td>
<td>1.94%</td>
<td>2.28%</td>
</tr>
</tbody>
</table>

Table 7: Percentage occurrence of academic rhetorical features in Tse and Hyland 2009, mapped to LIWC categories and compared with MOOC2014 corpus.

Further, Nesi and Gardner's (2006) research into variations in disciplinary cultures indicates that academic assessment in archaeology includes many features associated with soft disciplines (e.g. assessment by essay) as well as natural sciences (e.g. laboratory reports). By revealing a low occurrence of boosters (1.25%), and higher incidence of hedges (5.47%) and engagement markers (2.28%), linguistic analysis of the MOOC2014 corpus appears at least partially to substantiate this (Table 7). While not conclusive, this analysis indicates that the sample captured a mixture of distinctive hard and soft rhetorical features typical of academic discourse in archaeology suggesting that learners in this MOOC were adopting discipline-specific language. However, none of the LIWC characteristics associated with this analysis produced statistically significant results, and the influence of the idiosyncrasies of discipline-specific language on measuring levels of critical thinking in comments can be discounted.

In order to ascertain correlations between variables, statistical analysis software was used to generate scatter plots with lines of best fit, which identified the existence and intensity of simple linear regression.
3.3.1 RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forum comments? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?

Correlation coefficient (r) evaluations are based on Evans (1996), who classifies r values of less than 0.20 as very weak, 0.20 to 0.39 as weak, 0.40 to 0.59 as moderate, 0.60 to 0.79 as strong and 0.80 or greater as very strong correlations. Positive linear associations were made between all content analysis methods and each other (Table 8), all of which were very strongly correlated. The variables with the strongest statistically significant correlation across all three dimensions were the methods based on Bloom and SOLO taxonomies ($r = 0.868, p = 0.001$ - Figure 3), but correlation with scores based on the CoI model was also very strong ($r = 0.83, p = 0.001$ - Figure 4). This indicates that while the content analysis methods describe pedagogical activity in different ways, they appear somewhat consistent in identifying its presence and assessing its strength.

<table>
<thead>
<tr>
<th>Correlations</th>
<th>CoI</th>
<th>SOLO</th>
<th>Bloom</th>
<th>DiAL-e</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoI</td>
<td></td>
<td>$r = 0.811^{***}$</td>
<td>$r = 0.83^{***}$</td>
<td>$r = 0.673^{***}$</td>
</tr>
<tr>
<td>SOLO</td>
<td>$r = 0.811^{***}$</td>
<td></td>
<td>$r = 0.868^{***}$</td>
<td>$r = 0.693^{***}$</td>
</tr>
<tr>
<td>Bloom</td>
<td>$r = 0.830^{***}$</td>
<td>$r = 0.868^{***}$</td>
<td></td>
<td>$r = 0.711^{***}$</td>
</tr>
<tr>
<td>DiAL-e</td>
<td>$r = 0.673^{***}$</td>
<td>$r = 0.693^{***}$</td>
<td>$r = 0.711^{***}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Content analysis methods correlations. (** = p<0.01, *** = p<0.001)

![Figure 3: Correlation between Bloom and SOLO (ratings as per schemas – Tables 2 and 3)](image)

![Figure 4: Correlation between Bloom and CoI (ratings as per schemas – Tables 1 and 2)](image)
3.2.2 RQ 2: are linguistic content analysis measures (such as LIWC) significant indicators of levels of critical thinking as identified through pedagogical content analysis?

In this pilot study, the ability of LIWC to reliably analyse texts was inferred from the literature to be limited to those messages containing over 100 words (to conform with new information, this was revised down to 50 in the later study reported in Chapter 4). As many of the comments contain short interjections with low word counts, and to facilitate exploration within different contexts (first order [individual] and second level analytics [aggregated]) it was decided to investigate the corpus in three ways. First, all comments were explored, then text files containing aggregated rated comments from each of the steps included in this study were created and analysed. Finally, individual files for each comment containing 100 or more words (93 comments in total) were analysed separately from aggregated comments.

<table>
<thead>
<tr>
<th>All comments</th>
<th>Comments aggregated by step</th>
<th>≥100 word comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>(dof=599)</td>
<td>(dof=11)</td>
<td>(dof=92)</td>
</tr>
<tr>
<td>Likes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bloom</td>
<td>r=0.28</td>
<td></td>
</tr>
<tr>
<td>CoI</td>
<td>r=0.00</td>
<td></td>
</tr>
<tr>
<td>SOLO</td>
<td>r=0.02</td>
<td></td>
</tr>
<tr>
<td>DiAl-e</td>
<td>r=0.00</td>
<td></td>
</tr>
<tr>
<td>Word count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r=0.57</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.53</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.50</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.54</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Prepos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r=0.23</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.25</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.24</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.17</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Sxlr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r=0.15</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.17</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.18</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.18</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Pronoun</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r=0.32</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>r=0.33</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r=0.32</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>1st per.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r=0.33</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Posmo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r=0.29</td>
<td>***</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: LIWC2007 analysis. 
(ns: not significant, *: p<0.05, **: p<0.01, ***: p<0.001).
Correlation coefficient (r) evaluations: very strong – strong – moderate

This study sought to establish which word count indicators had significant correlations with levels of critical thinking using LIWC2007 software. Just seven of the 70 dimensions produced statistically significant results that had moderate to very strong correlations when compared with the pedagogic content analysis instruments: pronouns, present tense, word count, prepositions, first person singular, positive emotion words (posmo), and words containing more than six letters (sxlr) (Table 9).
Two moderate to very strong indicators were evident across aggregated and longer individual comments: pronouns (Figs. 5 and 7, highest correlation: \( r = 0.84, p<0.001 \)) and present tense (Figs. 6 and 8, \( r = -0.83, p<0.001 \)). Strong indicators in aggregated comments were observed, including: word count (Fig. 9, \( r = 0.63, p<0.001 \)), and prepositions (Fig. 10, e.g. to, with, above. \( r = 0.77, p<0.001 \)). First person singular pronouns were strongly or moderately associated with all analysis methods in the sample containing longer comments (\( r = 0.62, p<0.001 \)), word count had moderate associations with all analysis methods at an individual comment level (\( r = 0.59, p<0.001 \)), and the sixltr category achieved moderate correlations with the longer comment sample (\( r = 0.43, p<0.001 \)).

![Figure 5: Correlation between CoI and pronouns by average scores for 12 steps.](image1)

![Figure 6: Correlation between CoI and present tense by average scores for 12 steps.](image2)

![Figure 7: Correlation between CoI and pronouns by individual scores for >100 comments.](image3)

![Figure 8: Correlation between CoI and present tense by individual scores for >100 comments.](image4)
3.2.3 **RQ 3: to what extent do typical measures of attention to learning (such as social interactions) indicate levels of critical thinking as identified through pedagogical content analysis?**

All comparisons produced graphs that indicated approximate linear relationship between the four PS dependent variables and two explanatory variables (Figs. 11 and 12). Comment sentiment was analysed using LIWC2015 positive and negative emotion measures (posemo and negemo). There were variable negative correlations between all content analysis methods and positive sentiment (e.g. love, nice, sweet), with strong associations in aggregated comments ($r = -0.69$, $p<0.001$), and no statistically significant relationship between the methods and ‘likes’ (Table 9).

Analysis shows that pedagogical analysis ratings are not related to the number of ‘likes’ awarded to comments by users, which may indicate that within this learning environment, issues other than those strictly related to critical thinking received positive feedback. However, there were statistically significant negative associations between positive sentiment and higher PS, indicating that learners may employ a more formal approach to writing comments that indicate critical thinking, than when writing at a more surface level.
Figure 11: Correlation between CoI and positive words by average scores for 12 steps.

Figure 12: Correlation between CoI and Likes.

3.4 Conclusions

In preparation for the large-scale study reported in Chapter 4, undertaking this pilot study enabled me to identify levels of critical thinking within a MOOC comment forum. Simple measurements of this attribute were made (Pedagogical Scores) using four different pedagogic content analysis methods, which are distinct from measuring intentional user-feedback (‘likes’) and automated analysis of language complexity. In answer to RQ1 (are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forums? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?), because I was the sole comment rater in this initial study, it was not possible to explore inter-rater reliability. However, results suggest that while the content analysis instruments are designed to evaluate different aspects of online discussions (DiAL-e’s engagement with pedagogic activities, Bloom’s and SOLO’s increased complexity of understanding, and CoI’s development of reflective discourse), they are closely aligned with each other in terms evidencing very similar behaviours related to the levels of critical thinking. While these strong associations do not support assertions that the coding schemes are measuring the same things, the strong correlations between them, and between these schemes and learners’ use of language, suggests potential for developing real-time, automated feedback systems.

The practice of rating comments revealed styles of writing that are typical of this environment but which are not accounted for in all four pedagogical content analysis instruments. Because of the succinct nature of many comments in the sample, these are particularly evident at the lower end of the CoI, SOLO and Bloom scales. While the ‘Triggering’ and ‘Exploration’ dimensions of CoI explicitly facilitate rating for questions and some of the social dynamics characteristic of CSCL (aspects which are also classifiable using the ‘Inquiry’ and ‘Collaborate’ dimensions of DiAL-e), neither SOLO nor Bloom explicitly account for these features. In addition, these differences in focus inevitably lead to some instruments identifying some activities better than others, for example,
in the MOOC2014 corpus, examples of all dimensions in all instruments were identified with the exception of ‘Figurative’ in DiAL-e and ‘Extended Abstract’ in SOLO.

The close correlation of the Community of Inquiry instrument with linguistic proxies for critical thinking aligns with Joksimovic et al. (2014) and suggests that further research into this model is worthwhile. In addition, combining content analysis instruments may improve accuracy. Shea et al. (2011) recommend combining CoI and SOLO, although in my pilot study the two content analysis methods with the strongest correlations to the identified linguistic proxies are CoI and Bloom. As this study also found that SOLO was not appropriate for categorising behaviour in the informal settings typical of online discussion forums, this suggests that further analysis using CoI and Bloom are likely to produce more reliable results, and these two methods were adopted in the large-scale experiment described in Chapter 4.

Results derived from testing RQ2 (are linguistic content analysis measures significant indicators of levels of critical thinking in MOOC discussion forum comments as identified through pedagogical content analysis?) suggest that while the content analysis instruments are designed to evaluate different aspects of online discussions, they are closely aligned with each other in terms of evidencing very similar behaviours related to the depth and intensity of cognition, and with distinct word type patterns. Strong correlations with LIWC2007-based proxies for pedagogical activity and all content analysis methods have been identified. For example, Tausczik and Pennebaker’s (2010) research into the linguistic dynamics of a large, diverse dataset, indicates that pronoun use is a key indicator of emotional, cognitive and social awareness and, as in the current study, pronoun use is positively correlated with emotion words and negatively correlated with prepositions. Carroll’s (2007) study of linguistic patterns, found that students tend to use more complex words and fewer pronouns in their academic writing as they progressed through their studies, and Hartley, Pennebaker and Fox (2003) note that pronoun use in their study of journal articles accounted for 2% of the total number of words. This aligns with the current study where pronoun use is lower in comments rated for in-depth writing.

Finally, with regard to answering RQ3 (to what extent do typical measures of attention to learning (such as social interactions) indicate levels of critical thinking in MOOC discussion forum comments as identified through pedagogical content analysis?), in this study ‘typical interaction measures’ refer to intentional rating systems (‘like’ buttons) (Ferguson and Sharples, 2014) and ‘opinion mining’ techniques (sentiment analysis) (Ramesh et al., 2013). These measures are widely used in evaluations of learning as well as in other fields. ‘Likes’ are a commonly used rating mechanism that are adopted to measure personal attitudes (Kosinski, Stillwell and Graepel, 2013), and sentiment analysis has been used in social media research to explore people’s mood and attitudes towards politics, business and a number of different variables, including evaluating satisfaction with online courses (Wen, Yang and Rosé, 2014a). Findings from this study shows that PS is not related to the number of ‘likes’ awarded to comments by users. This accords with Kelly (2012) who argues that these measures suggest a variety of ambiguous meanings, and
Ringelhan, Wollersheim and Welpe (2015) who suggest that Facebook ‘likes’ are not reliable predictors of traditional academic impact measures. The most relevant outcome from correlation analysis of comparisons of outputs from LIWC2007 and the content analysis instruments is the clear, statistically significant, negative correlation between positive sentiment and PS. This negative association suggests that learners may employ a more formal approach to writing comments that indicate critical thinking, than when writing at a more surface level.

### 3.5 Summary

The pilot study contributed to understanding of the limits of the use of ‘likes’ as indicators of on-topic engagement, and establishes links between learners’ language use and their depth of learning. It strongly indicates the value of further content analysis using different datasets from a variety of courses, covering diverse subjects, with contributions from different participants, and analysed by multiple raters, in order to establish widely applicable techniques. This analysis was undertaken in the follow-up study, which is discussed in the next chapter.
Chapter 4: Large-scale Study – Content Analysis of Multiple MOOCs

This chapter reports on the second stage of a three-stage study, the objective of which was to expand on the work undertaken in the pilot phase, this time with a larger number of raters, evaluate two content analysis methods based on pedagogical activity, and compare them with linguistic and interaction analysis methods. The first two stages of this three-stage study set out to answer three questions (and set the stage for answering my fourth research question in the following chapters):

RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forum comments? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?

RQ2: Are linguistic content analysis measures significant indicators of levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?

RQ3: To what extent do typical measures of attention to learning (such as social interactions) indicate levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?

To achieve this, 1500 comments posted within three different MOOCs were manually scored using coding schemes based on Bloom’s Taxonomy, and Community of Inquiry: Cognitive Presence frameworks by several raters. The results of this coding were assessed for inter-rater reliability. These scores were then correlated against established linguistic analysis methods and measures of social interaction.

By comparing the rating of the two frameworks we can begin to address whether they are measuring the same sorts of activity (RQ1), and the inter-rater reliability scores will also show whether they can be applied consistently by different people (RQ1). Identifying linguistic, as well as interactional proxies for critical thinking (RQs 2 and 3), facilitates the development of an automatic approach to categorising comments and leads to tackling my fourth research question – RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in CSCL discussion forum comments?

4.1 Methods

Anonymised comment data from three MOOCs offered on the FutureLearn platform in 2014–15 were analysed. These MOOCs were chosen to enable the analysis of writing produced by learners
studying diverse disciplines: social science, humanities, and natural science. More than 41,500 registered learners engaged with the courses, with nearly 15,000 contributors posting over 174,500 comments containing more than 8.5 million words. Each MOOC was delivered via an average of 20 steps per week throughout each of the three to six-week courses, and each step provided the facility for instructors and registered learners to contribute to discussions within the steps’ comment field.

4.1.1 Research challenges

Undertaking qualitative research at scale provides several research challenges. Selecting samples that are large enough to reflect the volume of comments, and number of contributors, but small enough to allow the rating task to be completed within the limited time and budget of a PhD research project is a difficult balance to strike. In addition, choosing an appropriate diversity of academic disciplines in order to capture a variety of discipline-specific discourse is also problematic. Within budget constraints, a sufficiently large sample of random comments from MOOCs based in diverse disciplines was selected for rating (Figure 13).

4.1.1.1 Sample size

The choice of sample size was constrained by two key factors: the amount of time each rater could be expected to spend on-task in order to provide an authentic rating, the cost of employing raters to undertake the task, and the financial resources available for the project. In order to obtain accurate results, I anticipated that raters had to be motivated to undertake the tasks in an expert manner. Many studies support the proposition that among the intrinsic and extrinsic factors that motivate focus on work, the ability to earn money from undertaking tasks is the strongest. In his survey of more than 600 members of an online photography crowdsourcing community (iStockPhoto) Brabham (2009) describes 10 factors that motivate engagement with the platform, including: opportunities to learn, have fun, and improve reputation. However, the statement that most of the study’s participants agreed with was that the platform provided “the opportunity to make money”. Further, Kaufmann, Schulze and Veit (2011) found in their study on motivation among Mechanical Turk workers, that participants were motivated more by extrinsic than intrinsic factors. The study found that workers placed the promise of immediate payment for work above all other considerations when choosing which work to undertake, including having a choice regarding which tasks they undertook, and the opportunity to use a variety of skills.

Because the work I expected raters to undertake was relatively mundane, and provided little opportunity for choice, or using different skills, I decided that the best approach was to pay raters the normal rate used to pay university demonstrators. My assumption was that by paying raters at the normal rate for skilled teaching, they would consider the work as a professional activity, and give the attention required to provide accurate ratings in a reliable manner.
Having previously undertaken similar work myself (see Chapter 3), I estimated that, in order to provide a reliable rating, each rater would expect to spend an average of 30 seconds evaluating each comment. Using this estimate as a benchmark, and working within the constraints of my £700 research budget, as well as rating each comment twice (once for each analysis method), I estimated that raters could be reasonably expected to evaluate no more than 1500 comments in total (the MOOC2015 corpus).

Although amounting to less than 1% of the total number of comments submitted on the three chosen MOOCs, the MOOC2015 corpus is considerably larger than 20 – 140 sample sizes required to train ‘good classifiers’ suggested by Beleites et al.’s algorithm design research (2013). It was therefore considered large enough for the purposes of training the machine learning algorithm described in Chapter 5.

4.1.1.2 Data selection

Having established an achievable sample size, the next step was to select a MOOC, or MOOCs, from which to select the sample.

In Chapter 2 (Background and Literature Review) I explain that disciplinary studies research suggests that different disciplines adopt different rhetorical styles in academic writing, and have identified variations in the use of linguistic categories that are similar to LIWC categories, but which are characteristic of different disciplines rather than cognition. As analysis of discipline-specific language in the Archaeology of Portus MOOC pilot study (Chapter 3) indicated the presence of rhetorical discourse associated with its subject area, it is possible that, by selecting one MOOC from one discipline, rhetorical variations in linguistic practices may have adversely affect the outcome of this study.

Hyland (2007) asserts that there are significant differences between what he refers to as the ‘soft’ and ‘hard’ sciences, with hedges, boosters, self-mentions, reader pronouns and questions tending to be more common in ‘soft’ humanities and social science papers than in ‘hard’ natural sciences. These rhetorical features can be closely mapped to LIWC categories, which facilitates partial analysis in this study.

Because this discipline-specific tendency affects the use of linguistic features that are also associated with levels of critical thinking, and to avoid potential bias, I chose to evaluate comments from MOOCs based in different disciplines – one social science (Contract Management: Building Relationships in Business), one humanities (Understanding Language: Learning and Teaching), and one natural science (Exploring our Oceans). Rating comments from a diverse range of disciplines, and setting out to identify diverse discipline-specific language characteristics, facilitates effective evaluation.
4.1.1.3 Sampling method

Comment data from the three chosen MOOCs was provided by FutureLearn in three separate spreadsheet files. Unlike the pilot study which used a purposive sampling method in order to discover typical comments, this follow-up study adopted a random method in an attempt to achieve representative samples and remove the possibility of researcher bias from the selection process (Gravetter and Forzano, 2011).

Selecting 1,500 from the total of 174,500 comments involved adopting a variation of simple random sampling (Rowntree, 1981). While labelling each comment with a unique number and choosing 500 comments from each MOOC using a random number generator would have provided a satisfactorily random sample, during the training process raters expressed concern that comments selected using this technique would be presented out of context, and have a prejudicial effect on accurate rating. The expectation was that free-standing comments taken out of the context of on-going discussions could be misinterpreted by raters. In order to minimise this possibility, in addition to labelling each comment, they were also organised into numbered batches of 20 consecutive comments (the minimum number of comments considered to be large enough to facilitate context-based judgements). Eight batches from each MOOC were then selected for rating using a random number generator (Haahr and Haahr, 2015). Three randomly selected batches of 20 consecutive comments from each MOOC were also selected to facilitate test rating, prior to undertaking analysis of the rest of the sample (see Figure 13).
Figure 13: Sample selection process.

4.1.1.4 Comment rating process

Qualitative analysis was undertaken by a team of seven raters recruited from postgraduate students registered at the University of Southampton, using content analysis methods based on Community of Inquiry: Cognitive Presence (Table 1), and Bloom’s Taxonomy (Table 2), to rate whole comments (this process is shown in Figure 14). Two of the seven had backgrounds in education, two in anthropology, and one each from physics, psychology and languages. Five had previous experience of assessing written work.
The raters were provided with a one-hour face-to-face instruction workshop, where they rated a number of example comments following the coding scheme indicators shown in Tables 1 and 2. The raters carried out this process using an audience response system so they could observe how others scored the same comments and then discuss the reasons for each rating choice. Where comments exhibited traits associated with more than one category (for example, a comment may include a question (‘1 - Triggering’ in CoI) but also an unsupported contradiction of a previous statement (‘2 - Exploration’)), raters were instructed to rate for the characteristic with the highest value in each comment. Finally, raters were instructed to work alone when undertaking the rating task and not compare results with, or request advice from, anyone else.

To identify outlying scores and possible misunderstandings among raters, an initial test selection of 60 comments, comprising 20 randomly selected consecutive comments from each MOOC, were scored by the rating team. Intraclass correlation coefficients were calculated using a two-way mixed consistency, average measures definition and provided inter-rater reliability scores of 0.928 for Bloom and 0.880 for CoI suggesting “excellent” agreement (Cicchetti, 1994, p. 286), and that the pedagogical analysis methods were interpreted and applied similarly across raters.

Rating of a larger sample of 1,440 comments then went ahead. Comments from all three MOOCs were sorted into batches of 60 comments (3 x 20 consecutive comments) from which eight batches from each MOOC were randomly chosen and distributed among the raters (24 sets in total). Three raters rated 6 batches, two rated 7 batches, and two rated 8 batches. Raters rated each comment twice (once by Bloom and once by CoI), with raters scoring different batches for each coding scheme, and each batch scored by two raters working independently. To avoid confusion between analysis methods, scoring sheets were distributed with a 10-day time lag between each method, and only after the first pass had been completed by all raters.
Comments from the test and full sample were combined (n=1500 – the MOOC2015 corpus) and Pedagogical Scores (PS) for each comment were generated based on the mean scores of the raters who had evaluated each comment. In order to ascertain correlations between variables, statistical analysis software was used to generate probability (P–P) plots, and scatter plots with fitted lines to identify the existence and intensity of simple linear regression. P–P plots indicate the degree to which the distribution of ratings are skewed from normal distribution. If skewness is low, this allows some confidence in the precision of correlation and predictive values. As both coding schemes generate P–P plots with close to normal distribution (Figures 15 and 16), correlation and prediction values between variables (e.g. ratings, words per comment, likes and LIWC2015 categories) can be confidently made (Rowntree, 1981).

Since undertaking the pilot study, where LIWC analysis had been inferred to be unreliable for texts containing less than 100 words, advice provided with a newer version of the application, LIWC2015, suggests that texts containing less than 50 words “should be looked on with a certain degree of scepticism” (Pennebaker Conglomerates Inc., 2015). This allowed the analysis of comments containing 50 or more words – a larger proportion than had been undertaken in the pilot study. Because just 40% of comments contained less than 50 words, results were explored on
three levels: all 1500 comments (where analysis methods, likes, and word count were compared with individual comment scores), 150 aggregated batches of 10 contiguous comments (LIWC2015 compared with average scores for each batch), and 607 individual comments containing 50 or more words (LIWC2015 compared with individual comment scores). Adopting diverse viewpoints on the data facilitates exploration within different contexts. Exploration at an individual level enables the analysis of specific contributions (first order analytics), while investigating aggregated data offers indication of general course development (second order analytics).

![Figure 15: Normal P–P plot of Bloom average](image1)

![Figure 16: Normal P–P plot of CoI average](image2)

In addition, the pedagogical scores generated by the two different frameworks (CoI and Bloom) were correlated in order to explore whether the analysis methods were measuring similar levels of critical thinking.

Ethical clearance for this study was granted by the University of Southampton’s Faculty of Physical Science and Engineering Ethics Committee (ERGO/FPSE/17160).

### 4.2 Findings

#### 4.2.1 Sample analysis

Sampling of three batches of 20 comments for the test rating and 24 batches of 60 comments for the main rating process resulted in a total of 1,500 comments equally distributed among the three MOOCs. However, random sampling resulted in an unequal distribution of comments from each week of each MOOC’s duration. The Social Science and Humanities MOOC sample contained as
many comments from week 1 as the following weeks of each course (three weeks for the Social Science MOOC and five for the Humanities MOOC), while the Natural Science MOOC sample contained a more evenly distributed sample, selected from four of the six weeks of its duration (Figure 17).

Figure 17: Weekly distribution of sample comments.

Combining the test rating of 60 comments by all seven raters, with the 1,440 comments rated by two analysis methods, results in a total of 3300 ratings. Rating by both frameworks shows a normal distribution of scores, with Bloom scores skewed towards 0 rated comments (Figures 18 and 19). By exploring results by each MOOC (Figures 20 to 28), the high number of 0 ratings using the Bloom rubric can be ascribed to the prevalence of comments rated as ‘off-topic’ in the Social Science and Natural Science MOOCs.

Figure 18: Total scores using Community of Inquiry: Cognitive Presence rubric.
Analysis of the occurrence of diverse, discipline-specific language in the MOOC2015 corpus presents mixed results, with approximate alignment to Tse and Hyland (2009) discernible in the incidence of boosters and engagement markers (Table 10). However, the relatively high occurrence of hedges in both ‘hard’ and ‘soft’ MOOCs is more closely aligned with findings in the pilot study,
which suggests a possible ‘MOOC effect’ involving an increased tendency for MOOC learners to adopt tentative language.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedge (tentative + discrepancy)</td>
<td>5.00%</td>
<td>5.63%</td>
<td>5.47%</td>
<td>2.24%</td>
<td>1.67%</td>
</tr>
<tr>
<td>Boosters (certainty)</td>
<td>2.38%</td>
<td>1.5%</td>
<td>1.25%</td>
<td>1.69%</td>
<td>1.35%</td>
</tr>
<tr>
<td>Engagement markers (you + we + qmark)</td>
<td>2.9%</td>
<td>2.36%</td>
<td>2.28%</td>
<td>3.19%</td>
<td>1.94%</td>
</tr>
</tbody>
</table>

Table 10: Percentage occurrence of academic rhetorical features in MOOC2015 corpus compared with MOOC2014 corpus, and Tse and Hyland 2009, mapped to LIWC categories.

While using LIWC software to analyse comments may not capture the nuance of a more in-depth linguistic analysis, the accordance of the findings of this study with disciplinary studies literature is noticeable. Although the higher occurrence of engagement markers (i.e. questions) in the ‘soft’ MOOC comment sample may be influenced by the sample’s weighting towards the first week of these MOOCs (where learners are encouraged to write about themselves and ask questions), the association of other rhetorical features (i.e. reader pronouns and boosters) with the ‘soft’ MOOC comment sample suggests that discipline-specific rhetoric has been captured.

### 4.2.2 Research questions

Table 11 shows results generated when pedagogical analysis methods are compared with each other, with learner interaction (‘likes’), and with a wide range of linguistic analysis metrics. Cells have been shaded to show the seven strongest, significant correlations: word count, words per sentence, causation, differentiation, negation, cognitive process words, and first-person pronouns. Due to the number of significance tests (a total of 103) p values have been noted at <0.05, <0.01 and <0.001.

### 4.2.3 RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forum comments? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?

As the key test of objectivity in content analysis research, establishing the degree to which raters agree is vital. Unfortunately many studies either report percent agreement or report discussion leading to full agreement (De Wever et al., 2006). Krippendorff argues that this approach is of little use as it tests only the reliability of the individual raters, rather than the method (Krippendorff, 2018).
There is no settled method of testing agreement. However, as I treat CoI and Bloom levels as ordinal values, I adopt the Intraclass Correlation Coefficient (ICC) method, which is regarded as appropriate for this type of data (Hallgren, 2012).

### Table 11: Correlation and prediction results.
(dof = degrees of freedom, ns: not significant; *: p<0.05, **: p<0.01, ***: p<0.001).

<table>
<thead>
<tr>
<th>Category/Framework</th>
<th>All comments (dof=1499)</th>
<th>Aggregated comments (dof=149)</th>
<th>≥50 word comments (dof=606)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bloom</td>
<td>CoI</td>
<td>Bloom</td>
</tr>
<tr>
<td>Methods</td>
<td>r=0.909 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likes</td>
<td>r=0.237 ***</td>
<td>r=0.243 ***</td>
<td>r=0.263 ***</td>
</tr>
<tr>
<td>Positive correlations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word count</td>
<td>r=0.687 ***</td>
<td>r=0.704 ***</td>
<td>r=0.759 ***</td>
</tr>
<tr>
<td>Cause</td>
<td>r=0.125 ***</td>
<td>r=0.101 ***</td>
<td>r=0.573 ***</td>
</tr>
<tr>
<td>Diff.</td>
<td>r=0.220 ***</td>
<td>r=0.195 ***</td>
<td>r=0.443 ***</td>
</tr>
<tr>
<td>Negation</td>
<td>r=0.122 ***</td>
<td>r=0.110 ***</td>
<td>r=0.458 ***</td>
</tr>
<tr>
<td>Cogproc</td>
<td>r=0.125 ***</td>
<td>r=0.101 ***</td>
<td>r=0.397 ***</td>
</tr>
<tr>
<td>WPS</td>
<td>r=0.382 ***</td>
<td>r=0.389 ***</td>
<td>r=0.0430 ***</td>
</tr>
<tr>
<td>Aux verbs</td>
<td>r=0.104 ***</td>
<td>r=0.092 ***</td>
<td>r=0.371 ***</td>
</tr>
<tr>
<td>Power</td>
<td>r=0.222 ***</td>
<td>r=0.224 ***</td>
<td>r=0.369 ***</td>
</tr>
<tr>
<td>Sidr</td>
<td>r=0.145 ***</td>
<td>r=0.143 ***</td>
<td>r=0.197 *</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>r=0.271 ***</td>
<td>r=0.275 ***</td>
<td>r=0.280 ***</td>
</tr>
<tr>
<td>Negemo</td>
<td>r=0.112 ***</td>
<td>r=0.116 ***</td>
<td>r=0.449 ***</td>
</tr>
<tr>
<td>Prepositions</td>
<td>r=0.161 ***</td>
<td>r=0.169 ***</td>
<td>r=0.025 ns</td>
</tr>
<tr>
<td>Negative correlations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronouns</td>
<td>r=0.283 ***</td>
<td>r=0.268 ***</td>
<td>r=0.372 ***</td>
</tr>
<tr>
<td>1st per. sing.</td>
<td>r=0.321 ***</td>
<td>r=0.317 ***</td>
<td>r=0.353 ***</td>
</tr>
<tr>
<td>Affiliation</td>
<td>r=0.119 ***</td>
<td>r=0.117 ***</td>
<td>r=0.387 ***</td>
</tr>
<tr>
<td>Posemo</td>
<td>r=0.304 ***</td>
<td>r=0.327 ***</td>
<td>r=0.362 ***</td>
</tr>
<tr>
<td>Emotion</td>
<td>r=0.167 ***</td>
<td>r=0.166 ***</td>
<td>r=0.381 ***</td>
</tr>
</tbody>
</table>

Therefore, to establish the reliability of the pedagogical analysis methods used in this study, intraclass correlation coefficients were calculated between pairs of raters (two-way mixed, absolute agreement) and provided inter-rater reliability average scores of 0.831 for Bloom and 0.818 for CoI. According to Cicchetti (1994), this suggests excellent agreement between raters and indicates that the pedagogical frameworks were interpreted and applied similarly across raters.

When classifying the strength or correlations, this study follows Evans (1996), who classifies r values of less than 0.20 as very weak, 0.20 to 0.39 as weak, 0.40 to 0.59 as moderate, 0.60 to 0.79 as strong, and 0.80 or greater as very strong correlations.

When comparing pedagogical scores derived from the two analytical methods there is a very strong correlation of 0.909 (p<0.001), suggesting close association between Bloom’s levels of learning and
CoI’s measures of meaningful and productive discourse. While they describe critical thinking in different ways, they seem to be relatively consistent in measuring its presence and strength.

4.2.4 RQ 2: Are linguistic content analysis measures (such as LIWC) significant indicators of levels of critical thinking in MOOC discussion forum comments, as identified through pedagogical content analysis?

This study sought to establish which word count indicators had significant correlations with levels of critical thinking using LIWC2015 software. Two moderate to strong indicators were evident in aggregated comments: word count (Fig. 26, highest correlation: $r = 0.759$, $p<0.001$) and first-person singular pronouns (Fig. 27, $r = -0.533$, $p<0.001$). Moderate indicators in aggregated comments that were also very weak indicators in ≥50 word comments were observed, including: causal words (Fig. 28, e.g. because, effect, hence: $r = 0.573$, $p<0.001$), differentiation (e.g. but, apart, else: $r = 0.443$, $p<0.001$), negation words (e.g. no, not, never: $r = 0.458$, $p<0.001$), words per sentence (wps: $r = 0.43$, $p<0.001$), negative emotion words (e.g. afraid, envious, ugly: $r=0.449$, $p<0.001$), and cognitive process words (e.g. decide, idea, solve: $r = 0.397$, $p<0.001$). Weak indicators in all approaches included: pronouns (e.g. I, them, itself: $r = -0.372$, $p<0.001$), auxiliary verbs (e.g. am, will, have: $r = 0.376$, $p<0.001$), and power (e.g. control, protect, warn: $r = 0.369$, $p<0.001$). Finally, word types with weak, or very weak, correlations across all approaches to corpus analysis included: positive emotion words (e.g. agreed, fair, kind: $r=0.39$, $p>0.001$), affiliation words (e.g. belong, our, share: $r=-0.387$, $p>0.001$), emotional tone (combined summary of posemo and negemo: $r=-0.382$, $p>0.001$), conjunctions (e.g. and, also, although: $r = 0.28$, $p<0.001$), words with six letters or more (sixltr: $r = 0.2$, $p<0.05$), and prepositions (e.g. above, between, unless: $r=0.169$, $p>0.001$).

These results show statistically significant, positive correlation, between several word count indicators and levels of critical thinking, confirming results of studies which associate high word counts with thoughtful, ‘exploratory’ exchanges in formal CSCL environments, and first-person person singular pronouns (‘I-talk’) with a negative effect associated with higher levels of critical thinking, as well as other indicators, which are discussed below.
4.2.3 RQ 3: to what extent do typical measures of attention to learning (such as social interactions) indicate levels of critical thinking in MOOC discussion forum comments, as identified through pedagogical content analysis?

Correlations between the pedagogical content analysis methods and measures commonly used to measure social interaction such as ‘likes’ (Qi et al., 2012), and sentiment analysis were also explored. While ‘likes’ gave positive, significant results across all approaches to analysis, the correlation was weak (Table 11 – maximum r = 0.298, p<0.001). In terms of sentiment, typical measures include positive and negative emotion (negemo), and emotional tone. While negemo provided significant, positive, moderate associations with both frameworks in aggregated...
comments, this category produced non-significant results in ≥50 word comments. Results for other sentiment analysis were negatively, and weakly correlated across all approaches.

Having found ‘likes’ not to be a significant indicator of levels of critical thinking in the pilot study (as described in Chapter 3) it was surprising to observe the contrary in this more in-depth study. However, these results suggest that, while statistically significant, user ratings and sentiment provide only a weak indication of critical thinking in MOOC comments.

4.3 Conclusions

4.3.1 RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forum comments? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?

4.3.1.1 Inter-rater reliability

In their wide-ranging review of content analysis methods De Wever et al. identify six interrelated criteria for assessing their effectiveness; instruments should be: “accurate, precise, objective, reliable, replicable, and valid” (De Wever et al., 2006, p. 8). Central to these criteria are the theoretical basis of the instrument, the unit under analysis (i.e. the comment as a whole, or in part), and the extent to which they can be replicated across a variety of settings; from an individual rater agreeing with themselves, then two or more raters reaching agreement, to the reliable use by many different groups of researchers (Rourke et al., 2001). The content analysis methods used in this study were applied by a small group of seven raters, who applied the analysis criteria to individual, whole comments. The high level of agreement in this study (ICC: Bloom = 0.831 and CoI = 0.818) suggests that these methods may be successfully applied in other settings, and provide the foundation of automated rating systems that conform to commonly held values regarding levels of critical thinking.

Investigating the behaviour of analysis methods in different contexts is facilitated by exploring the comment data using three sampling techniques. Looking at all 1,500 comments allows us to make inferences about general word count and interaction categories, individual ≥50 word comments facilitates LIWC word category analysis at an individual contributor level, and aggregations of all comments provides a general view of contributor’s comment behaviour. These approaches are useful in different contexts. For example, understanding language dynamics at an individual level is important for analysing the behaviour of specific contributors, and an aggregated approach can provide an overview of how activity within the course is progressing.
4.3.1.2 Pedagogical content analysis methods

As in the pilot study, the pedagogical content analysis methods used in this study were also very strongly correlated. Some educational psychologists argue that individual and distributed cognition are two distinct, interrelated processes (Moore and Rocklin, 1998). As each method emphasises these different aspects of cognition (Bloom – individual, CoI – distributed), this suggests that, in this study, there is a strong connection between individual levels of critical thinking and how this develops through discussion. This outcome may result from aspects of learning design that are unique to the FutureLearn platform.

For example, while online learning environments are not always synonymous with improved depth of thinking (Slagter van Tryon and Bishop, 2009), there is some evidence that providing learners with timely and detailed prompts can lead to higher levels of thinking (Darabi et al., 2011). By placing discussion forums within the context of each activity and providing educator and mentor support (León et al., 2015), the FutureLearn platform encourages sharing and situated debate (Ferguson and Sharples, 2014), with the explicit intent of building communities of inquiry and inspiring higher levels of critical thinking.

In addition, the two instruments may be measuring very similar behaviours related to the depth and intensity with which people write about what they are thinking. If we agree that there is an approximate connection between complexity of writing and depth of critical thinking, it is reasonable to assume that someone who has applied greater attention to their learning, and wishes to share this with others, will use more elaborate arguments (‘Create’ in Bloom), or attempt to summarise debates (‘Resolution’ in CoI), and suggests that comments evidencing these types of focus will tend to be ranked in a similar manner. Although the instruments based on those frameworks are sensitive to different aspects of learning, results show consistency in measuring the presence and strength of critical thinking, which suggests their interchangeability in quantifying these properties in this setting.

As discussion forum research indicates that lower levels of critical thinking are typically prevalent in formal CSCL environments, the relatively high occurrence of comments coded as Integration and Resolution in CoI, as well as the three highest levels of Bloom, is also notable. There are a number of possible reasons for this. It may simply have occurred as an artefact of the random sampling method used in this study, a prevalence for the raters to be generous in their evaluations of comments, or it may suggest success in a platform design that explicitly aims to encourage learners to engage in higher level comments.

While it is possible that random sampling may have produced a higher proportion of high level comments, this is unlikely given the size of the sample and diversity of sources. Further, although raters were instructed to apply the highest rating appropriate for each comment, during the instruction workshops there was no indication that they were over-generous in their assessments.
This leads to the conclusion that undertaking research into the effect of learning design on learners' engagement with critical thinking may be a fruitful undertaking.

4.3.2 **RQ 2: are linguistic content analysis measures (such as LIWC) significant indicators of levels of critical thinking in MOOC discussion forum comments, as identified through pedagogical content analysis?**

An important aim of this study was to determine predictors that closely align with cognitive processes in CSCL, and the literature indicates that LIWC is an accurate tool for measuring significant aspects of language use in this setting.

The most relevant outcome from regression analysis of comparisons of outputs from LIWC2015 and the content analysis instruments is the clear, statistically significant, positive correlation between word count and levels of critical thinking, which confirms findings of studies which associate high word counts with thoughtful, ‘exploratory’ exchanges in formal CSCL environments (Ferguson and Buckingham Shum, 2011; Joksimovic et al., 2014). In addition, results for first-person singular pronouns (‘I-talk’) are also supported in the literature, in showing moderate, significant results across all profiles, with a negative effect associated with higher levels of critical thinking (Robinson, Navea and Ickes, 2013).

Findings for causal and cognitive process words all provide moderate positive correlations with both pedagogical content analysis methods across aggregated comments, and agree with the findings of studies of language use in formal education as well as informal settings (e.g. Pennebaker, Mayne and Francis, 1997; Creswell et al., 2007; Ferguson and Buckingham Shum, 2011). While power words provided weak associations across all approaches, their correlation with higher levels of critical thinking in this study suggests self-confidence in expressing opinions (Winter et al., 1998).

With regard to emotional tone and positive sentiment words, the statistically significant, though weak, negative correlation between these categories and learning objects with high scores is another noteworthy outcome of this study. Some studies suggest that a correlation with positive sentiment indicates a loss of focus from key tasks (Leshed et al., 2007), which, by showing higher incidence of these categories associated with lower levels of cognitive engagement, this study appears to agree with. However, the positive association of positive sentiment words with higher levels of cognitive engagement in longer comments (containing 50 or more words) in this study implies agreement with other studies that suggest that higher levels of positive emotion equates with a greater focus on group cohesion, and use of prior knowledge, rather than systematic analysis (Schwarz, 2000).

Results for negative emotion were also mixed, with significant, moderate positive correlation with critical thinking in aggregated comments, but no significant results in ≥50 word comments. While this category is associated with higher levels of critical thinking in the literature (Hertel et al., 2000)
this study suggests that this, along with positive sentiment, may not be a reliable measure in all samples.

4.3.3 RQ 3: PS is closely correlated with typical interaction measures of attention to learning.

The significant, positive association of ‘likes’ in this study with high PS was unexpected, as this metric has been reported as ambiguous and unreliable, and outcomes from the pilot study produced insignificant results for this metric. While results from this study, and good research practice, suggests caution when making inferences from Web paradata and ambiguous phenomena like individual behaviour or cognition (especially with correlation values of less than $r = 0.3$), the significant result across all comments and aggregated comments suggests that a process of “aggregated trustworthiness” is possibly at work (Jessen and Jørgensen, 2012). In this setting, a sufficient number of MOOC forum contributors may be using the ‘like’ button as an indicator of the trustworthiness and expertise of certain posts, rather than using it to signify agreement or ironic disagreement. Although I was unable to uncover any empirical evidence in the literature to support this particular argument, Facebook ‘likes’ have been used to predict intelligence levels (Kosinski, Stillwell and Graepel, 2013), and some researchers have found that the ‘like’ feature can moderately stimulate learner motivation (Shih, 2011), suggesting this may be a fruitful area for further research.

Finally, words containing six or more letters returned weak correlations, and words per sentence, negation, auxiliary verbs, conjunctions and prepositions returned mixed and insignificant results in the $\geq 50$ word comment sample. Although the use of long words and long sentences have been associated with higher levels of critical thinking, the literature reports mixed findings for this category. In this study, negation, auxiliary verbs, and conjunctions produced moderate or weak results in the aggregated comment sample, but analysis of the $\geq 50$ word comment sample revealed no significance for these features in individual CoI coded comments, with conjunctions and auxiliary verbs producing very weak correlations in Bloom-rated comments. These inconclusive results suggest that using these categories as a sole indicator of critical thinking is not advised, but that there may be a place for these categories supporting analysis that include other significant features.

Together with the unexpected significant result for likes, the low correlation values or lack of significance of prepositions was also unanticipated. When aggregating all three MOOCs, findings from this study do not appear to agree with the large number of studies that have found statistically significant positive associations between prepositions and attention to reflective behaviour or increase cognitive load. However, exploratory analysis of results filtered by MOOC revealed insignificant results for prepositions in the business-related MOOC, with significant moderately correlated results for this word type in the other two. This may be explained by the lower incidence
of ‘off-topic’ comments in the latter MOOC samples, which further suggests that aspects of language analysis are highly context-dependent.

4.4 Summary

This study set out to answer three key questions:

RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forums? In particular, can different people consistently apply them, and do different frameworks identify the same types of learning activity?

Converting informal MOOC comments into comparable scores based on multiple pedagogical analysis methods is a significant research challenge. In this study, a group of seven raters achieved a high degree of reliability using both pedagogical analysis methods (ICC: Bloom = 0.831 and CoI = 0.818), which enables us to have some confidence in the generalizability of these methods in future studies. Building on previous research in formal CSCL environments I established close associations between two distinct methods applied to informal settings which contradict previous findings (e.g. Chan et al., 2002), suggesting the value of further investigation of critical thinking evaluation in MOOCs.

While the coding schemes used for pedagogical content analysis have been developed to assess different aspects of learning (individual and distributed cognition), when applied in this context, and correlated against language categories, sentiment and ‘likes’, there appears to be very little difference in how they measure levels of critical thinking (r=0.909, p>0.001).

RQ2: Are linguistic content analysis measures (such as LIWC) significant indicators of levels of critical thinking in MOOC discussion forum comments, as identified through pedagogical content analysis?

Confirming previous research (e.g. Ferguson and Buckingham Shum, 2011; Vosecky, Leung and Ng, 2012; Creswell et al., 2007), through LIWC2015 analysis I identified word count and first-person singular pronouns as convincing indicators of depth of learning (wc: r = 0.759, p<0.001; pp: r=-0.533, p<0.001), with causal words, differentiation, negation words, words per sentence, negative emotion words, and cognitive process words providing moderate results (causal: r = 0.573, p<0.001, diff.: r = 0.443, p<0.001, negate: r = 0.458, p<0.001, wps: r = 0.43, p<0.001, negemo: r=0.499, p<0.001, cogproc: r = 0.397, p<0.001). Other word categories provided mixed results within the two sampling methods used, suggesting a supporting role for these categories in future studies.

RQ3: To what extent do typical measures of attention to learning (such as social interactions) indicate levels of critical thinking in MOOC discussion forum comments, as identified through pedagogical content analysis?
While producing significant results, and confirming previous work suggesting ‘likes’ prompt engagement with higher levels of learning (Darabi et al., 2011), both measures of sentiment and ‘likes’ were weakly correlated with measures of critical thinking (posemo: $r=-0.390$, $p<0.001$; negemo: $r=0.449$, $p<0.001$; like: $r=0.298$, $p<0.001$). Therefore, as with weakly correlated word types, this suggests secondary roles for these measures in future research.

Henri (1992) suggests that the object of analysing education-based CMC interactions is to improve “the efficacy of interaction with students” (p. 117). Despite progress in codifying content analysis methods, and the development of automated Natural Language Processing techniques, the absence of effective tools means the process of coding remains arduous and time consuming (Mu et al., 2012). For instructors, the timely identification of learners in need of pedagogical support is as relevant now as it was when Henri addressed the issue 25 years ago, and the strong correlations with LIWC2015-based proxies for pedagogical activity and the coding schemes used for pedagogical content analysis in this study suggest significant promise for automated tools.

This chapter set out to build on the pilot study presented in the preceding chapter, and identified distinctive features of learner comments suggestive of critical thinking. Both studies involved using coding schemes used for pedagogical content analysis to manually rate comments posted on MOOC discussion forums from which Pedagogical Scores were derived, and correlations with linguistic and interaction features were identified. The next chapter reports on a study where I use these findings to answer my fourth research question – RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in CSCL discussion forum comments? This computational approach has the potential to go beyond the linear regressions presented in this chapter, combining multiple metrics in order to predict pedagogical scores, and facilitates the automatic rating of comments.
Chapter 5: Automated Evaluation of MOOC Comments

This chapter reports on the first part of the final phase of my three-stage study, the aim of which is to answer my fourth research question:

*RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments?*

Machine learning (ML) affords a substantial approach to automating analysis. The technique applies appropriate search methods to produce computational approaches that automatically improve with experience (Mitchell, 1997). Research in Learning Analytics and Educational Data Mining suggests that ML techniques are effective in automatically identifying pedagogical activity (Kizilcec, Piech and Schneider, 2013; Wen, Yang and Rosé, 2014b; Dowell et al., 2017; Kovanovic et al., 2017), and the aim of answering this research question is to develop an algorithm that can successfully measure levels of critical thinking in MOOC comment data.

The purpose of this automated evaluation is to provide support to mentors in managing the flow of comments at the massive scale evident in MOOCs, so that they may identify interesting comments and improve their interaction with learners.

As identified in the preceding chapters, the pedagogical frameworks used to rate comments, Community of Inquiry: Cognitive Presence (CoI) and Bloom, are very closely correlated, and this strongly indicates their interchangeability regarding the measurement of critical thinking in MOOC comments. In order to train an algorithm to automatically rate comments I used comment data that had been rated using the CoI model, as it is the simplest model of the two, having fewer categories than Bloom. However, because of its close association with the Bloom model used in my study, I propose to integrate both approaches. Rather than use CoI’s five classifications to signify the distinct Cognitive Presence categories of triggering, exploration, integration, or resolution (plus off-topic), because of this close association, it is more appropriate to use the categories as an ordinal scale to indicate levels of critical thinking. Thus, the CoI model is transposed to ‘critical thinking values’ as follows: 0 (off-topic) = low, 1 (triggering) = modest, 2 (exploration) = average, 3 (integration) = good, 4 (resolution) = high.
5.1 Method

5.1.1 Machine Learning

ML program design requires four main elements: the training set (e.g. a collection of comments); the target function (e.g. correctly rated comments); the learned function (e.g. pedagogical and linguistic content analysis methods); the classification method (e.g. Naïve Bayes, J48, Random Forest) (Mitchell, 1997). Having gathered the first three elements in the earlier stages of my research, the next phase involved selecting the classification technique and building the classifier algorithm, testing them on datasets that have been labelled by human judges, applying them to unlabelled data, and evaluating the results.

There are three main approaches to machine learning: supervised learning requires a set of classified data from which to learn; unsupervised learning finds hidden structures within unlabelled data; and reinforcement learning learns from interactions with a changing environment. As the dataset used in this study has been previously rated, this facilitates a supervised learning approach, and the literature review has identified Random Forest as an effective classification method for this type of data (Caruana and Niculescu-Mizil, 2006; Fernández-Delgado et al., 2014). To develop an algorithmic model for predicting levels of critical thinking in MOOC comments adopting the Random Forest classifier, I used WEKA (ver. 3.9.1), a widely-used machine learning ‘workbench’ application (Witten, Frank and Hall, 2011).

5.1.2 Random Forest classification

Random Forest classification is produced by the construction of a large number of decision trees during the data training phase, with the resulting algorithm being applied to a test dataset. Decision trees typically facilitate judgements about data which enable categorisation (e.g. if a particular data property exists, then it belongs to a certain category, if not, then another category – Figure 29), but is susceptible to overfitting training sets. Random Forest corrects for this by using a collection of decision trees and averaging outputs (Figure 30). This enables ensemble learning, whereby each tree evaluates a different random sub-sample of the data and contributes each evaluation to a voting scheme from which predicted classification is decided (Breiman, 2001). In addition, the decision tree method is randomised by choosing a range of potentially effective options, randomly selecting among them, and creating different trees with each pass. This process is based on the concept that while each tree may be a weak learner, when combined they deliver a better performance (Witten, 2013).
Figure 29: Decision tree showing possible categorisation judgements based on Word Count feature properties. (C0: low, C1: modest, C2: average, C3: low, C4: high).

Figure 30: Random Forest creates an ensemble of decision trees.

5.1.3 Data processing

As with all ML methods, the Random Forest approach may be compromised by the lack of sufficient data in the training set. There are no strict rules for estimating the size of a training set, although the aim should be to achieve a low error rate from test outputs. In this study, time and cost constraints allowed coding of just 1,500 cases. Because of the unreliable nature of LIWC analysis of text containing fewer than 50 words, these instances were removed from the rated comment dataset (MOOC2015 corpus) to leave 600 rated comments containing 50 words or more (n=600 dataset).
In order to evaluate classifiers derived from the machine learning process, data is typically split into two independent sets; a training set (to train a model algorithm) and a testing set (on which to validate the model). A review of machine learning literature indicates that there is no hard-and-fast rule on how to split a dataset in preparation for training and testing. Mitchell (1997) suggests splitting the data by a ratio of 66:33. However, in their learning analytics research Kovanovic et al. (2016) split their data 75:25, and others suggest training/testing splits of 70:30, 60:40, or 50:50 (Quora Inc., 2016). In their comparison of two models used to predict bankruptcy, Odom and Sharda (1990) compared results from splitting their small dataset (n=129) 50:50, 80:20, and 90:10, and found the 50:50 data set produced the best results overall. Further, Polat and Günes (2007) found a 50:50 split of their breast cancer diagnosis dataset (n=683) obtained higher classification accuracy than splits of 70:30 or 80:20. As a 50:50 appeared more likely to provide the most accurate results, I split my n=600 dataset to provide a training and testing dataset containing 300 instances each.

While the n=600 dataset I used in this study is less than half the amount required for training estimated by Warrender, Forrest and Pearlmutter (1999), who suggest 800 instances as a reasonable rough guide for training purposes, it is greater than the 20 – 140 sample sizes required to train ‘good classifiers’ identified by Beleites et al. (2013). This indicates that the n=600 dataset, while small, contains sufficient instances with which to train a classifier.

5.1.3.1 Balancing classes

As is common in text classification, the n=600 dataset contained more of some classes of comments than others, with middle range classifications (‘average’ and ‘good’) more prevalent than ‘high’ and lower classifications. This ‘class imbalance’ can lead to overfitting of the model, and inaccurate results (Chawla, Japkowicz and Drive, 2004). One way to overcome this problem is to use a resampling method to produce a dataset with equal numbers of instances in each class. Typical methods involve generating synthetic samples from randomly selected instances of minority classes, and under-sampling instances from the majority classes.

A systematic approach commonly used to generate synthetic samples is the Synthetic Minority Over-sampling Technique (SMOTE). Instead of making copies of samples from the minority classes, the SMOTE algorithm works by creating new, synthetic samples. It creates new instances by selecting two or more similar instances (using a distance measure) and building them one attribute at a time by a random amount, within the difference to the neighbouring instances (Chawla et al., 2002). By applying this method to minority classes, and under-sampling majority classes, balanced training and test sets were produced, as detailed in section 5.2 below.
5.1.4 Validation

There are a number of ways to assess how a machine learning algorithm will generalise to an independent data set. Typically, with large amounts of data, the model is validated using a percentage split of the data, but with more modest sized datasets (like the n=600 dataset), cross-validation is the preferred method. Essentially, cross-validation is a systematic method of evaluating algorithmic models using different partitions of a dataset, which keeps the variance in that evaluation as low as possible. It works by dividing the dataset into a number of segments (e.g. 10) each containing an equal number of instances. It then takes nine of these segments for training and uses the last piece for testing. It repeats this process 10 times ('folds') using a different segment for testing each time, and averages the results from each fold. In WEKA, this approach is further improved by using stratified cross-validation, which selects an equal representation of class instances in each fold and outputs a low-variance model (Witten, Frank and Hall, 2011).

5.2 Procedure and classification

Designing the best algorithmic model for predicting levels of critical thinking in MOOC comments entailed a number of processing stages.

5.2.1 Process dataset

In preparation for the qualitative stage of my research, comments containing fewer than 50 words were removed from the rated comment dataset (MOOC2015 corpus) to leave 600 rated comments containing 50 words or more. All LIWC attributes that returned statistically insignificant results in the large-scale study were also removed (leaving 17 attributes – Table 10), and average CoI ratings (i.e. critical thinking values) were rounded to whole integers.

5.2.2 Split dataset

The n=600 dataset was put into random order and split in half; one half for training and the other for testing. This resulted in two sets containing 300 instances each, with similar but not identical, numbers of instances in each class.

5.2.3 SMOTE training set.

Because of unbalanced classes in the data which may result in overfitting, training and testing sets were balanced using the SMOTE method (Figures 31 – 32).

5.2.4 Apply classification method

Using the default settings in WEKA, the Random Forest classifier was applied to the training set. Average CoI scores were converted from numerical to nominal data, and, applying 10-fold cross-
validation, the dataset was classified using the critical thinking scores and the 17 significant attributes.

5.2.5 Select best model

In order to optimise the performance of the model derived from the training set, the attributes which produced the worst reliability metrics were removed (Table 10).

5.2.1.1 Validate

To test the validity of optimised model further, it was applied to the test set, and Cohen’s kappa and Intraclass Correlation Coefficient results were calculated.

5.3 Findings

Following the procedure outline above, once the n=600 dataset had been processed and split, the training and testing sets were balanced, and the chosen classifier was optimised and validated against the test set.

5.3.1 SMOTE

Because of imbalance in classification categories present in the training sets, SMOTE was applied. Using the SMOTE filter in WEKA to add new instances to minority classes, and reducing the number of instances in majority classes, resulted in new, balanced training datasets, containing an equal number of instances in each class. Most of the higher range classes in these new datasets contained comment instances that had been rated by humans, and the other classes included synthetic data that had been derived from human-rated instances. Figures 31 and 32 show the results of applying SMOTE to the training and testing sets.
5.3.2 Building the best classifier

The Random Forest classifier was applied to the training set, and evaluated against the critical thinking attribute using 10-fold cross-validation. All 17 of the other attributes were used in the first
instance, and validation metrics derived from this process were assessed. To achieve the best validation results for each training set, the attributes which produced the worst reliability metrics were removed and tested, and by a process of iteration the optimal set of attributes were selected. It was found that removing the likes, negation, prepositions and pronouns attributes, and using just 13 linguistic attributes produced the best model (Table 12).

<table>
<thead>
<tr>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likes</td>
</tr>
<tr>
<td>Word count</td>
</tr>
<tr>
<td>Causation</td>
</tr>
<tr>
<td>Differentiation</td>
</tr>
<tr>
<td>(Negation)</td>
</tr>
<tr>
<td>Cognitive process</td>
</tr>
<tr>
<td>Words per sentence</td>
</tr>
<tr>
<td>Auxiliary verbs</td>
</tr>
<tr>
<td>Power words</td>
</tr>
<tr>
<td>Six letters or more</td>
</tr>
<tr>
<td>Conjunctions</td>
</tr>
<tr>
<td>Negative emotion</td>
</tr>
<tr>
<td>(Prepositions)</td>
</tr>
<tr>
<td>(Pronouns)</td>
</tr>
<tr>
<td>First person singular</td>
</tr>
<tr>
<td>Affiliation words</td>
</tr>
<tr>
<td>Positive emotion</td>
</tr>
</tbody>
</table>

Table 12: Attributes used to build the classifier. The attributes in brackets were removed to build the final model.

### 5.3.3 Validation

Table 13 shows reliability results generated when the Random Forest classifier is applied to training and testing sets. The initial classification using all 17 attributes on the training set produced Cohen’s κ of 0.575. This was improved by removing attributes that produced lower reliability outputs, and the best model achieved with just 13 attributes (κ = 0.6209). The model was then applied to the testing set, and achieved a Cohen’s κ of 0.2083 and an Intraclass Correlation Coefficient score of 0.695 (highlighted in Table 13).
<table>
<thead>
<tr>
<th>Classification</th>
<th>Cohen’s k</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest (training set, 17 attributes)</td>
<td>0.575</td>
<td></td>
</tr>
<tr>
<td>Best model (training set, 13 attributes)</td>
<td>0.6209</td>
<td></td>
</tr>
<tr>
<td>Best model (testing set n= 300)</td>
<td>0.2083</td>
<td>0.695</td>
</tr>
</tbody>
</table>

Table 13: Reliability results (results for test set are highlighted).

A confusion matrix derived from the classified test set was produced (Table 14). This table presents the true class in rows and the predicted class in columns, with each entry representing the number of instances predicted for each class (accurate predictions are bolded). The confusion matrix derived from the classified test set shows poor prediction accuracy for comments rated as having ‘low’ levels of critical thinking, where just nine of the sixty comments have been accurately predicted (15%). However, improvement further up the scale is evident, as predicted scores are more closely aligned with actual scores (e.g. 36 out of 60 comments [60%] rated ‘high’ are accuracy predicted), or are within one level of the actual rating (e.g. 56 out of 60 comments [93%] rated as ‘good’ have been predicted within one level of accuracy).

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted level of critical thinking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Low</td>
<td>9</td>
</tr>
<tr>
<td>Modest</td>
<td>6</td>
</tr>
<tr>
<td>Average</td>
<td>2</td>
</tr>
<tr>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 14: Confusion matrix for the test set (n=300).

5.4 Conclusion

5.4.1 RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments?

The purpose of this stage of my study is to develop an algorithm that measures levels of critical thinking in MOOC comments. The algorithm uses the Random Forest classifier with attributes derived from linguistic features that are associated with levels of critical thinking in comments rated by humans. Measuring success in this area can be undertaken in two ways: using established reliability measurements that compare actual ratings with the algorithms predictions, and seeking human responses to these predictions. In this chapter I focus on the former approach, before evaluating the latter in the next chapter.
5.4.1.1 Reliability

While achieving a kappa of 0.6209 when applied to the training dataset, suggesting ‘substantial’ agreement (Landis and Koch, 1977, p. 165), the model performed less well when applied to the test set, achieving just 0.2083. This outcome compares unfavourably with results achieved in similar work (for example Kovanovic et al. (2016) achieved a kappa score of 0.63 in their research using cognitive presence categories), but may be considered ‘fair’ agreement according to Landis and Koch’s criteria.

Although widely used in the learning analytics domain, the kappa statistic is more appropriate for evaluating pair-wise agreement in nominal scales, as it does not account for the closeness of incorrect to correct predictions observed in multi-item ordinal scales. Since my critical thinking scale is ordinal, and perfect agreement is not critical, the Intraclass Correlation Coefficient value of 0.695 achieved by the algorithm may be considered a more suitable measure of its reliability. Cicchetti (1994) asserts that ICC values of between 0.60 and 0.74 are indicators of ‘good’ agreement, which within terms of my thesis, may be sufficiently close to provide potentially useful ratings if applied to new unlabelled data.

As I have argued earlier, complete accuracy in this setting is not vital. For example, if a comment is misclassified as ‘good’ rather than ‘high’ it is unlikely to have any adverse effects in this setting. Rather the rating serves as an indication that, on balance, a comment is likely to contain something of interest to a mentor who, for example, may be seeking learners who are engaging with a topic at a deeper level. I therefore considered the reliability measurements obtained for the algorithm to be good enough for use in the next stage of my study.

5.4.2 Further machine learning

As good ICC values were achieved for this algorithm, I continued with my study, applied the machine learning algorithm to a ‘live’ MOOC, and sought responses from MOOC mentors in semi-structured interviews. This case study is presented in Chapter 6.

Following this case study, I reviewed my approach to the machine learning task described above, and reflected on the decisions I had taken. There were a number of factors outside of my control which affected the decisions I took. Most important among them was the timing of the MOOC I had selected as the best prospect to test the algorithm. In order to begin my case study as close to the start of the MOOC as possible, I had to limit time spent on building the algorithm, and missed opportunities to test different approaches and optimise design.

On reflection, I concluded the key factors influencing my approach: using a reduced dataset ($n=600$) containing only comments with 50 or more words, and choosing to split the data 50:50, may have had a significant impact on the performance of the algorithm.
An accepted rule of thumb is that more data usually provides better results than designing more sophisticated algorithms (Rajaraman, 2008), and ML researchers often compare the effects of using different training to testing ratios (e.g. Odom and Sharda (1990), and Polat and Günes (2007)). Therefore, in order to explore the ML process, once I had completed the ‘live’ MOOC case study, I adopted a new approach using all of the MOOC2015 dataset (n=1500), and a 70:30 split with balanced classes.

5.4.2.1 Results for the new approach

Table 15 shows reliability results generated when the Random Forest classifier is applied to training and testing sets derived from a 70:30 split of the entire MOOC2015 dataset (n=1500). The initial classification using all 17 attributes on the training set produced Cohen’s k of 0.5464 (lower than the 0.575 achieved using the reduced dataset with a 50:50 split). This was improved by removing one attribute that produced lower reliability outputs (negation), and achieved a kappa of 0.5548 (again, lower than the 0.6209 realised earlier). The model was then applied to the testing set, and achieved a Cohen’s k of 0.4417 and an ICC score of 0.85 (highlighted in Table 15) – a significantly better result than produced earlier (k = 0.2083, ICC = 0.695).

<table>
<thead>
<tr>
<th>Classification</th>
<th>Cohen’s k</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest (training set, 17 attributes)</td>
<td>0.5464</td>
<td></td>
</tr>
<tr>
<td>Best model (training set, 16 attributes)</td>
<td>0.5548</td>
<td></td>
</tr>
<tr>
<td>Best model (test set n=450)</td>
<td>0.4417</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 15: Reliability results for n=1500/70:30 dataset.

The confusion matrix resulting from the classified test set (Table 16) shows improvement on the previous model, with prediction accuracy considerably higher than the previous model for both ‘low’ and ‘high’ rated comments (70% and 73% respectively for ‘low’ and ‘high’ rated comments, compared to 15% and 60% in the earlier model). Although not as great, improvement on the earlier model is also evident in the middle of the scale, with prediction accuracies of 36% for ‘modest’, 46% for ‘average’, and 50% for ‘good’ ratings, compared with 35%, 42% and 32% respectively. Predictions that are within one level of the actual rating are also improved in the new model (e.g. 81 out of 90 comments [90%] rated as ‘average’ have been predicted to within one level of accuracy, compared to 85% using the earlier model).
### Predicted level of critical thinking

<table>
<thead>
<tr>
<th>Actual</th>
<th>Low</th>
<th>Modest</th>
<th>Average</th>
<th>Good</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>64</td>
<td>13</td>
<td>8</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Modest</td>
<td>29</td>
<td>33</td>
<td>22</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Average</td>
<td>5</td>
<td>21</td>
<td>41</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>Good</td>
<td>0</td>
<td>7</td>
<td>13</td>
<td>45</td>
<td>25</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>22</td>
<td>66</td>
</tr>
</tbody>
</table>

Table 16: Confusion Matrix n=1500/70:30 test set.

### 5.5 Summary

This chapter set out to answer my fourth research question:

*RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments?*

Building a text classifier that can reliably measure critical thinking in MOOC comments is a substantial research task. Although based on pedagogical ratings that achieved strong associations with linguistic and interaction features present in MOOC comments, and using a highly-regarded method with which to build a new classifier, the reliability of the resultant algorithm proved to be uncertain, and not as effective as similar classifiers developed elsewhere. In addition, further research demonstrates that different approaches to managing data can produce substantially different outcomes. The implications of these outcomes are discussed in Chapter 6.

Nevertheless, because the critical thinking rating scale is ordinal, and perfect accuracy is not critical in this setting, the reliability results achieved with the earlier classifier (ICC = 0.695) were considered good enough (following Cicchetti (1994)) to continue testing its success in a ‘live’ MOOC environment, using the algorithm to classify unlabelled comments, rather than comparing predictions with previously rated comments. In the following chapter I report on a qualitative case study where I use the first classifier described in this chapter to automatically rate comments in this setting. Results from this approach are shared with relevant MOOC moderators, and semi-structured interviews seeking evaluations of the usefulness of this automated classification are analysed and discussed.
Chapter 6: ‘Live’ MOOC Case Study

This chapter reports on the second and concluding part of the final phase of my three-stage study, the aim of which is to answer my fourth research question:

**RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments?**

Expanding on the work undertaken previously, this chapter tackles the thorny issue of what happens when an algorithm leaves the ML lab and enters the real world. By applying the model algorithm presented in the previous chapter to an on-going MOOC, and seeking evaluations of the outcomes of the algorithm from course educators and mentors (i.e. moderators), my aim is to discover whether they find these outputs generally useful, and, specifically, if the automatic classifications align with their evaluations of learners’ levels of critical thinking.

In Chapter 2, I discussed the theoretical foundation of the role of educators and mentors as being within the frame of ‘teaching presence’ as described in the CoI model (Garrison, Anderson and Archer, 1999). I also discussed evidence that moderators experienced overwhelming difficulties handling the large volume and diversity of comments they are required to manage. But what reflections do MOOC moderators’ have on their own practice?

### 6.1 MOOC moderators’ reflections on their practice

In their survey of online and MOOC teaching in the US, Jaschi and Lederman report that 80% of the more than 2,700 teaching staff questioned agreed that it is “very important” that an online course: “provides meaningful interaction between students and instructors” (2014, p. 11). Numerous studies also support the view that meaningful interactive has a positive impact on student satisfaction (Thurmond et al., 2002; Walker and Kelly, 2007; Espasa and Meneses, 2010; Croxton, 2014). Furthermore, both the theory supporting teaching presence, and observation of educators and mentors practice, support the opinion that MOOC moderation involves more than simply encouraging learners to discuss course subject matter, but also to challenge and direct learners towards higher levels of learning. But, as Bayne et al. argue: “There are many ways to get it right online. ‘Best practice’ neglects context.” (2016). While aiming for effective teaching presence through meaningful interaction may be a desirable goal in face-to-face or campus-based online courses, the MOOC context entails many complex challenges for educators and mentors that are not found in small-scale settings (Ross et al., 2014). Research into the role of MOOC mentors indicates that large-scale teaching leads mentors to feel not only overwhelmed by the volume and diversity of comments, but also challenged in their ability to provide rapid and rigorous answers to complex encounters with learners (León et al., 2015).
However, gaining a fuller understanding of moderators’ experiences of MOOC teaching is hampered by a paucity of studies in this area, with their perspectives on teaching at MOOC-scale limited to blog posts, and a small number of qualitative studies. To date, research has predominately focused on learners’ experience, and institutional issues, and “the lack of published research on MOOC facilitators’ experience and practices leaves a significant gap in the literature” (Liyanagunawardena, Adams and Williams, 2013, p. 217). Although the literature suggests that reflection on the teaching process is critical to successful learning outcomes (Biggs and Tang, 2011), this deficiency continues. In a recent systematic literature review of the topic, Deng and Benckendorff (2017) found that of the 48 papers reviewed, only four involved interviews with instructors. Of these four, one explored face-to-face facilitation (Chen and Chen, 2015), and another reports on educators’ lack of online teaching experience, uncertainty on the purpose of MOOCs, and difficulties in engaging with the “volume of anonymous students” (Evans and Myrick, 2015, p. 305) – an observation shared by other studies (Gerber, 2014; Haavind and Sistek-Chandler, 2015). In their qualitative interview-based survey of eight MOOC instructors, Haavind and Sistek-Chandler also suggest that software-based approaches to assess interaction present opportunities to lighten the instructors load (2015 cited in Blackmon, 2016, p. 99).

León-Urrutia, Fielding and White's qualitative study on postgraduate students' attitudes to their experience of MOOC mentoring, reports study participants’ improved skills in a number of areas (confidence in their subject-knowledge through having their expertise challenged, communicating to diverse audiences, and managing their online identity), but also highlights the perceived “unrealistic” expectations placed upon them (2016, p. 7). For example, one participant reports: “[e]ven just reading the contents takes an hour, more than an hour”, and another found themselves “spending half a day reading ten papers to answer one question”. This supports the results of earlier work that identifies choosing which comments to respond to, and uncertainty in responding to learners queries and challenges, as key concerns for educators and mentors (León et al., 2015).

However, beyond reflections on the large volumes of comments typical of many MOOCs and suggested methods on managing this phenomenon, there is little research into MOOC moderators’ interactional practice within comment forums. FutureLearn’s Blog reports an educators’ reflections on forum moderation and advises caution against in-depth engagement because of the unforeseen effect an intervention may have on other learners:

“As a MOOC educator, if you do step into the discussion, it can have a huge ripple effect among the thousands of learners. So we tried to make very general, reassuring comments and not step in too much” (FutureLearn, 2017b).

Attention to the facilitation aspect of teaching presence is also advocated in an online video aimed at instructing new MOOC educators:
“Teacher presence is a very important part of the socialisation of students into online learning, and it’s not that you are therefore dominating and telling students what to learn, it’s that you’re playing the role of ‘guide-on-the-side’, the person who’s there to help the students along but not to become the one they rely upon” (College of Fine Arts: The University of New South Wales, 2011; 1m 30s).

But other educators give greater emphasis to the direction aspect of teaching presence. Ross and Bayne’s review of UK MOOC pedagogy reports one educator’s interview response: “[e]ducation is all about interaction between teacher and student, and if you think you can just put up a bunch of resources on the Web and tell students to just get on with it, you might as well write a book and they can buy it or borrow it from the library. It’s not a course. A course is about having a taught experience” (2014, p. 45).

Discovering MOOC moderator’s attitudes to providing pedagogical support in comment forums is central to answering random forest 4, which seeks to determine if ML algorithms can be trained to successfully measure levels of critical thinking in MOOC discussion forum comments. By triangulating case study participants’ attitudes to their pedagogical activity within comment forums, with their observations on the accuracy of the automatic rating, and the quantitative data analysed in the previous chapter and the research literature, it is hoped to gain a fuller picture of the issue than would be gained from quantitative analysis alone.

6.2 Method

In order to gain insight into how useful my automatic rating method might be to potential users, it is essential to understand how the ratings are used and understood by educators and mentors. A great deal of current research focuses on learners and institutional issues, but, in speaking directly with MOOC moderators about their experiences with comment forums in MOOCs we can begin to understand the role forums play in an educator’s or mentor’s pedagogy, and how automatic rating may be of use to them.

Data collection in this case study seeks answers to a number of questions: what are MOOC educators’ and mentors’ attitudes to encouraging critical thinking; what characteristics of learner comments do they look for when choosing comments to respond to; and what are their attitudes to the automatic method used to rate comments?

These questions are clearly associated with moderators’ personal experiences of working on a MOOC, and in this study I adopt Interpretative Phenomenological Analysis (IPA) as an appropriate research method for this stage of my research; primarily because of its emphasis on
gathering participants’ “personal perception or account of an object or event” (Smith and Osborn, 2015, p. 53).

IPA is a form of phenomenological method that is intended to understand the lived experiences of participants, and the context in which their experience takes place. Within the perspective of my research, this involves gathering and interpreting interview evidence from moderators engaged in interacting with learners via a MOOC comment forum. The goal of the research was not intended to be generalisable to all moderators and settings, but to gain knowledge about MOOC comment forum interaction and the usefulness of my automated rating method through the rich detail described by interview participants. A lack of generalisability to all circumstances is not important to this exploratory study, as it provides the research community with potential directions for further enquiry and improved knowledge of the phenomena.

Consistent with Smith and Osborn (2015), while using a semi-structured interview approach, I attempt to be impartial during data collection, and let the participants’ experiences develop on their own terms. IPA employs a dual approach to analysis which I adopt in the interpretation of participants’ answers: empathetic and questioning. This means that my experience in this field allows me to both understand the point of view of the participant, while also interpreting their responses critically.

In keeping with other IPA studies, this case study uses a small, purposively selected sample. I chose to interview three educators and three mentors working on FutureLearn’s Digital Accessibility MOOC run in early 2017 in the belief that occurrences of the phenomenon (pedagogical use of MOOC comment forums) was most likely to be found within this group. The most important factor was that participants are actively engaged in the phenomenon, and are “experiential experts” (Smith and Osborn, 2015, p. 59). In this research, each interview is compared with the other and compared and critically evaluated against the literature review. This involves a process of triangulation – using alternative methods to check and collaborate findings (Denscombe, 2010).

Participant selection was achieved opportunistically. Via a professional contact, I was permitted to approach 12 MOOC educators and mentors employed on the Digital Accessibility MOOC, from whom six volunteered to take part. These participants were based in six universities each based in six different European countries. Each participant was responsible for moderating forum comments during different weeks of the MOOC’s five-week duration. With the exception of one mentor, all had previous teaching experience, including student assessment, and had mediated at least one MOOC in the past. Two of the educators had taught in online environments for more than 10 years.

Forum comments from each week were automatically analysed using the Random Forest algorithm described in Chapter 5. This was derived from a ML process that applied linguistic and interaction features associated with levels of critical thinking in MOOC comments, as rated by humans.
With the exception of one mentor (who was interviewed face-to-face), each moderator was interviewed via Skype, and audio recorded. The initial plan was to interview all moderators twice for approximately 30 minutes each time. Interview 1 was to be undertaken before participants were given an analysis of all comments for their week, with Interview 2 taking place after having received the analysis (see examples of full interview transcripts in Appendix D and F). However, two mentors dropped out of the project after undertaking their first interview, which meant that just four of the remaining moderators were interviewed twice.

Prior to the first interview all participants were sent an information sheet (Appendix A), which set out the reasons for the study, and the first list of questions (Appendix B). Once the first interview was completed with each participant, automatic analysis of the comments for the week they were responsible for moderating was undertaken. An information sheet containing a description of the automatic method, graphs showing interaction for each step undertaken in the week in question (e.g. Figure 33), illustrative comments related to each step, a spreadsheet containing all of the weeks’ comments (containing 50 or more words) plus automatic ratings for each comment, and a second list of interview questions were then sent to each participant (Appendix C). The supplied graphs enabled educators to immediately determine the timing and level of engagement in each step and throughout the week. For example, Figure 33 clearly displays a greater number of higher rated comments being made by learners at the start of the week compared with later.

![Figure 33: Automatic ratings for Step 1, Week 1.](image)

Each dot represents a comment containing 50 or more words.
Following Smith and Osborn (2015), the transcribed interviews were coded for emerging themes, then grouped into categories, and finally re-evaluated to determine the scope of answers within each category.

Ethical clearance for this study was granted by the University of Southampton’s Faculty of Physical Science and Engineering Ethics Committee (ERGO/FPSE/25220).

6.3 Findings

The case study focuses on six MOOC moderators located at universities in Europe, who have recently been engaged in moderating a MOOC comment forum. Through a process of reading, and re-reading the interview transcripts, and noting patterns in topics and issues discussed during the interviews, I identified eight key themes. Table 17 shows how five of these themes map to the interviews with the six participants’ first interviews.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Participant</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learners' motivation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using the comment forum</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Moderators' engagement strategies</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Moderators' response triggers</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identifying critical thinking</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 17: Identified themes in first interviews.

6.3.1 First interviews

6.3.1.1 Learners’ motivation

Participants recognised that learners’ intrinsic motivation and their belief in the importance of the course to their development was as a significant factor in the level of their engagement in the MOOC:

I think it is motivation, I think if they’ve got time and I think if they – it strikes a chord with them. So, if it’s relevant to them. (B1)

You really have to be, have the interest or really self-discipline yourself to follow the course, or follow all activities and do all the activities. I think many people drop out, although they
really wanted to but, yes, they don't have time, right? So, you have to find time. I think this kind of engagement is very important. \(D1\)

A willingness to engage with the subject at a deep level, and invest effort in exploring associated topics was also acknowledged as a key attribute for successful learners:

The ideal learner is one who has an interest in the topic, in the sense that it is wide and not too narrow, and also on the other hand, who is ready to invest time into looking a little bit beyond what's in the course material, and do some own analysis – and some own extra look at maybe Wikipedia, or somewhere else. \(A1\)

But one participant also acknowledged the significance of the role of educators in understanding learners’ expectations, ensuring that course content was relevant and did not deter learner engagement:

If you didn’t pitch it at the right level I’d say they’re gone…as soon as you mention HTML or those kind of more, even slightly more technical terms you know you run into difficulty. \(B1\)

As the course was free for anyone to join, and learners had a diversity of backgrounds and prior experiences, another participant recognised the importance of support from fellow learners in developing a learning community:

And then also sometimes you have peer pressure, and sometimes you feel that you are part of the community, does help to get you going. Because sometimes it can be very lonely to follow a course without having, you know, peer support. \(D1\)

**6.3.1.2 Using the comment forum**

Several participants recognised the central importance of the comment forum in helping learners develop a community, and engage with the topic, as well as assist moderators in identifying specific areas of interest:

It’s the comments themselves that are the most important things, I think, in terms of being able to help the person engage, not just with other people on the course, but actually with their educator as well…That’s a really good indicator when people are really starting to engage with a comment, is the fact that you can see lots of replies and people replying to replies, if that makes sense. So, when I see that, I know that there’s something really interesting if people actually bother to take the time to reply to something. \(C1\)

However, all participants acknowledged that the volume of comments presented insurmountable problems for forum moderators wishing to monitor all communications:
I only have about an hour in which I'm able to make comments, so I can't necessarily read through every single comment. (C1)

I think that's a technical problem in the sense that navigating through the comments means to read a lot of text. And because the texts are short, they take a lot of time in order to be reflective yourself. (A1)

The comments are very hard to read because there's no organisation, no categorisation, there's a page, command page for every part, section but there's no category, there's no organisation. So, I have to go through all the comments to answer them and give some, and it is very hard to look at the comments, some comments in the area because sometimes there's hundreds of comments and you have to go through them all. (F1)

In this MOOC, all moderators were based in European time zones, and some reported their ability to follow and respond to comments from other parts of the world was compromised:

Because of the structure of the MOOC, I think there should be an immediate feedback but I cannot ensure this because I have other things to do. [Laughs]. And since people are active at any time, they come from all over the world, I cannot be present all the time. (A1)

So, for example you see, at nine o'clock you'll get the Brazilian ones, or something of the Australian comments, and then later on you get the English comments, and maybe the American ones, then later on again...I think you have to devote a lot of time to keeping up with the commentary like that. Otherwise what you tend to do is, you know, 'there's 100 comments, I'll go through the first 20 and see what they are saying'. (B1)

One participant suggested that MOOCs were a potential retrograde pedagogical development that inhibits student-centred teaching and learning:

So, we understand that, you know, many years, in 30 years, people talk about individualised learning, personalised feedback, right? And now you come to MOOC and you have lost this. Because it's almost impossible to do it, to give individual feedback in time. (D1)

### 6.3.1.3 Moderators engagement strategies

Despite issues related to the volume of comments, all participants recognised the importance of their visibility in the forums, and adopted questioning techniques to stimulate debate, and encourage deeper thinking:

I use the comments to engage learners by asking questions, asking them to share their experience, asking them to search for information and to test especially the technology we are showing in the course and give their feedback. (F1)
I think [in] the discussions you can ask leading questions to deliberately provoke discussion and to assess. So, for example you might put down a particular guideline or an approach and say, ‘But will this work if you took it outside, or if it was happening in a noisy room, for example a conference hall?’, or something like that. So, you are looking at different ways to get them to think deeper which I suppose is another aspect of critical thinking. (B1)

In the extract above, the participant perceived that setting leading questions can motivate higher level thinking, but in the extract below, they also suggest that moderators’ responses are controlled by concern for maintaining learner engagement.

There’s a steep exponentially decreasing curve of drop-off. So, what can you give them to take away? That would be the primary concern to make sure that those, right, if they leave after week one, what will they get? If they leave after week two, what will they get? And so on. So, it is, I think we are lucky in that we managed to retain a fair few, but I think when dealing with MOOCs, I think engagement and drop off is probably primary concern ahead of anything else. (B1)

Several moderators also expressed concern for the lack of the lively pedagogical discourse they experienced in face to face teaching:

In classroom you could, you know, intentionally direct them to encourage them to think differently, right? So, and then you can comment on them, and then you can ask the other students, ‘what do you think’, right? So, you have the class dynamics. Well, online you can ask a question, and then it is, it is often. I mean, I have done, I think I’ve used at least two hours every day following the MOOC weeks. And I read all the comments. And I follow the comments. So, I can ask a question, for example, that encouraged them to think differently or think critically. But usually in MOOC you don’t get such close follow-up. Then this project is normally. You are lost I would say. (D1)

This is different to a lecture where you have maybe one to three questions in a row and there the discussion probably helps a lot of people, and other people jump in and continue to discuss. While in the MOOC, there’s no such like, no such let’s say, ‘discussion’, going on in a sense. (A1)

Someone wrote: ‘digital accessibility is important because it allows someone with a disability access to the digital world, and more importantly, independence while doing so’. And then I responded, well, ‘Digital accessibility only creates access to the digital world, or does it also impact things or activities that are normally not digital?’ Because, it’s very easy to think of digital as just having access to the computer and Internet…but for some people…it impacts much more than just the digital world. So, that was kind of what I was trying to get a learner to realise that, but there was no response then. (E1)
6.3.1.4 Moderators’ response triggers

In addition to asking questions to prompt debate, all participants reported focusing on learners’ questions, when looking for comments to respond to:

I tend to find I’m actually answering where other learners have actually replied to a comment. I see that, I tend to focus on that first, and then basically just scan through the comments as quickly as I can and anything, that, sort of, stands out, that has key words in it, or things like that, or where I see a question mark, so there are people actually asking questions. Those are the sorts of things I’ll be focussing on and trying to look for and if I see that, I think I can answer. (C1)

In the extract above, the participant also reported searching for topic specific words in addition to questions from learners. In the extracts below, other participants explain how they also look for evidence that a learner has engaged in new thinking:

If the question is based on their own thoughts, right, then they have thought about the thing, and also have done some research. And then on top of that, they ask a question. Usually this question reflects their own understanding and their own research. So, I like those type of questions, because it makes me think about it. (D1)

I am actually looking for these comments which are reflective in a sense, and maybe come up with extra information, but also maybe ask questions…I remember last time when the course ran first, I had a person who was really commenting in order to challenge also others, and basically get some feedback…I thought if I replied to these people then others who read and don't think they can…add anything more, they can benefit from this. (A1)

Several participants also reported seeking to correct learners’ misunderstandings:

People wonder if something ‘like that’ exists, or think this would be nice to have, and then I point out a few things: ‘Yes, this already existed as a solution, etc.’ Sometimes, just noticing when they are flat out wrong. Also provide pointers to information, I’ve corrected that. Yeah, two important categories. (E1)

Another participant observed peer-support for learners where learners’ help one another:

I saw one particular learner who said, basically, ‘I understand X, but I’m struggling with the Y, can somebody explain that?’, and that’s what they did. And that’s where I see people really move on. (C1)

In addition to seeking out learners’ questions and misunderstandings, this participant asserted that learners also need to be prompted to apply their new learning:
If you’re wanting learners to really start to think critically, to evaluate or judge things for themselves, you can’t expect them to just do it in a vacuum. You’ve got to give them some kind of guidance or pointers. You don’t want to give them the answer on a silver platter. You want to get them to think about what they’re doing, but you can’t expect that unless…you’re taking existing concepts then getting them to apply them, and they come up with their own version. (G1)

6.3.1.5 Identifying critical thinking

Participants were also asked to consider how they identify in-depth engagement in learners’ comments. Several cited evidence of relevant reflection on personal experience, or provision of new pertinent examples, as important markers.

One of the typical scenarios [is] In My Workplace – ‘I have not done this’, or ‘I have done this’, and then they come up with additional thoughts that go beyond what’s in the course material. (A1)

When a learner can provide an example of their own, and they can then relate what they’ve learned to an example of their own, that’s when I can see very clearly that they’ve understood something properly…You’ve seen how they’ve taken it forward. It’s not just a theoretical thing, they’ve actually done something with it. That’s how I can conceive they’ve taken it on board. (C1)

One participant highlighted learner discussions that took an analytical approach to specific applications:

Also you get kind of partial SWOT-type analysis of the strengths and weaknesses of certain things. So, some will say for example a particular technology is good and others will say, ‘well I had difficulty with it’, or ‘there's too much of a learning curve’. (B1)

As in the extract above, the following participant explains how comments that demonstrate application of knowledge are important:

Simply knowing about different disabilities in our case and the variety of the requirements is not something that you only reflect on, but you really have to go through this and really, after the course, you should be able to enumerate this, and be able then to apply this knowledge in case you encounter a type of disability not covered in the course. (A1)

One participant spoke about learners who critique the educators’ arguments, putting forward alternative hypotheses and interpretations based on their knowledge and experience:

One would be, being provocative to some degree. That they are arguing against. That they offering counter argument. That they are looking at it from a different perspective. That
they have, I suppose, absorbed the lesson and added something to it would indicate it as well. So, a good example of that, that cropped up quite a bit in the MOOC and will crop up again in week five is the way that they would localise the general principles. (B1)

Some participants also suggested learners engage in critical thinking when they bring their perspectives, and new evidence to debates within the forum:

- When I read the comment I look for: 1) they have thought about the issue themselves, and they have their own opinions about things; and 2) they have new thoughts that nobody else has written or commented on. And sometimes they have, they post links that I didn’t even, I am not aware of. So that shows that they have done some research themselves. (D1)

- So, it is, one is, you know, getting involved in the debate. I didn’t see it on this round but on the last round for example there was a big healthy debate about whether serif based fonts are unhelpful for people with dyslexia or not. You know, people were bringing back-up arguments and statistics, and everything like that, which I thought was interesting because they were challenging the perceived norm. (B1)

However, participants’ opinions on the value of encouraging critical thinking are divided. One had not considered critical thinking:

- I don’t think I have thought about critical thinking to be honest. (D1)

Another expressed the view that gaining proficiency over the content was more important:

- In a course like this where it is largely, I suppose, partly informative, partly to relevance, and partly designed about quick takeaway, I think it might be a secondary concern in some degrees. (B1)

Whereas several participants perceived the value of monitoring learners’ critical thinking:

- As somebody who is providing feedback, I think it is important to find those who are really important that come up with something that hasn’t been covered and where my example also is worth reading for others. (A1)

- I think it to be absolutely fundamental. I think it would be really critical to do that. (C1)

- Particularly in MOOC, if you know your project, for example, can give the teacher or educator some possibilities to have a quick overview of how the students are doing, and if you can come this far and to give the educator some advice, so that they can actually – based on this – and give feedback to the students. It will be a very, very good addition to the MOOC pedagogy I would say. (D1)
6.3.2 Second interviews

After the first interview, participants were sent the results of the automatic rating of learner comments in the week each participant was responsible for monitoring, including an explanation of how the method was developed (Appendix E). Two mentors were unable to continue with the project, and the opinions of remaining four participants were sought regarding the automatic analysis they had been sent. During analysis of the transcripts (see Appendix F for an example of a full transcript) three themes emerged, Table 18 shows how they map to the interviews with the four participants’ second interviews.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Participant</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning of rating levels</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Accuracy of automatic rating</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Usefulness of automatic rating</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 18: Identified themes in second interviews.

6.3.2.1 Meaning of rating levels

One of the key themes identified in the interview transcripts related to uncertainty regarding the meaning of the 5 critical thinking levels. One participant observed some ambiguity in the meaning of the levels and was uncertain if the categories were part of an ordinal or nominal scale:

I’m not sure if I understood the 0–4 rating. Does 4 denote higher critical thinking? Or is 1 to 4 different levels or types of critical thinking? (C2)

Another participant suggested there was a lack of clarity on distinctions between rating levels, and suggested amalgamating categories to produce a simpler scale:

At first I was a bit confused, because the approach that you have to provide scores is very brief, and doesn’t really explain where, basically, the decision is between a three and a four, or a one and a two…The three, the two, the one are difficult. So maybe it is better to bring them together to one category. (A2)

The accompanying explanatory document clarified the meaning of a 0-rated comment as ‘off-topic’, yet some participants believed the rating was vague, and often inaccurately applied:

Zero is ambiguous, I believe. (D2)
I think all the zeros I could see seemed to be pretty much on-topic. Or they were talking around the subject. For me that didn’t really work so well. But what it seemed to be picking up was more I-talk. So, people talking about themselves or their family. It was more personal stuff. (B2)

However, this belief was not universal among all participants. One found 0 ratings to accurately reflect the content of the comments to which it was applied:

Q: The aim of the analysis is to show different levels of critical thinking and the zero rating is supposed to be off-topic. I mean, when you were looking at the analysis that I did, did that actually mean anything, did that zero actually accord with off-topic comments?

A: As far as I could tell yes, it did and it was. I was actually, to be honest I was quite surprised how few there were overall which, which I thought was really, really helpful. (C2)

6.3.2.2 Accuracy of automatic rating

Despite these concerns all participants expressed the view that the overall accuracy of the automatic ratings appeared reasonable. One participant observed that the ratings provided a useful overview of how the MOOC was progressing:

I think actually it was about right; it was enough to kind of get a feel for what was going on, and particularly for the visualisation. To actually see it was excellent, it was really helpful. (C2)

Another participant found the ratings to be generally indicative of the level of critical thinking she had observed during her weeks’ moderation:

In general I agree most part but I think there are more...I will say there are more comments that represent critical thinking that are not in here. (D2)

When asked directly to rate the accuracy of the automatic ratings, this participant gave a relatively high score:

I would say roughly 80%. (D2)

In addition to a spreadsheet containing all comments and ratings for each week the participants carried out moderation work, the participants were sent an information sheet containing example
comments with their associated automatic ratings. While most participants provided general observations on the accuracy of these ratings, participant A provided an unprompted, comment-by-comment interpretation of each rating.

Of the 67 comments reviewed by Participant A, just 10 were identified as inaccurately rated. Of these 10 he suggested eight required up-rating, and two down rating. Several comments were singled out as ‘well chosen’, but underrated. For example, Participant A suggested one comment that was rated three, should have been rated four:

It was one of those examples where I thought, oh, we are so lucky to find these people in this course, so in my mind it is actually a four. (A2)

This participant also reported that the automatic analysis had highlighted a comment he had previously missed, and suggested that it should also receive a higher rating:

I didn’t see this when I was going through the comments that this guy wanted to have more information about measuring the correctness of mental maps. I would have liked to refer to some applications. Actually, it’s what I would like to see from a student, yes: ‘Give me more information’. So, a two is certainly not adequate here. (A2)

Regarding another comment that refers to a discussion on a ‘Walkers Guide’, Participant A suggested that this should have received a lower rating, in part because the comment author had misunderstood the learning context:

The person who wrote – who was given a four – criticises the accuracy of maps, but in terms of, OK, ‘British Ordnance Survey is the best’. Yes, but this is not bringing to the point. That’s really about people, more information and benefit from it, information and maps. So, in my mind it is a three. (A2)

While this participant does not provide alternative ratings for all 10 comments that he believed were inaccurately rated, when his reflections are mapped and compared to the automatic ratings he was provided with, this results in an informal percent agreement measurement of 93% (Appendix G). However, it should be noted that this remarkably high level of agreement is derived from a tiny fraction of all comments in this MOOC, and, while participant A confirmed his agreement with many of the ratings, he also suggested that care is required when interpreting machine-derived scores:
The zeros and the fours are quite, quite useful. Not always, not always accurate, but otherwise I was surprised to see that these examples are OK, and therefore – with a certain reservation: ‘Be careful, this is automatic’. (A2)

Care when interpreting automated feedback was also highlighted by another participant, who explained that rule-based methods may contain in-built biases:

For me it’s a recommender system, right? If you look at this, you recommend some of the comments to you, right? So, all recommender system has a general problem is the narrowness of it, right? So, if you only have certain rules, and you only selected those that fit in those rules, right? Then you will miss the diversity. Of course, all such systems have the same problems, right? And whether it’s harmful depends on if you wanted to use this with regard of the students critical thinking – I will think a bit more about it, right? So, yeah. It doesn’t give you a complete picture of the critical thinking of each individual learner. I will say that, if you use this, purely use this, to judge and that will give you a bias picture. (D2)

While suggesting that the ratings were useful, another participant suggested that an approach that used key words would provide closer scores:

It’s a bit like hitting it with a blunt instrument by nature. And I think the next generation of this, where maybe you’re looking at new phrases and stuff, that might be something. But I think for what it is itself, it’s a finer approach and fine methodology and so on. So, I think it’s very good. (B2)

Participant B also suggested that the number of replies in a thread may also indicate deeper levels of engagement:

I think looking at the length of the threads would be an interesting metric to put in there. Because I think that the length of the threads is a crude indicator of deeper exploration. (B2)

6.3.2.3 Usefulness of automatic rating

All participants stated that they believed this type of automatic feedback would be useful in enabling moderators as well as learners to find in-depth comments. One participant pointed out that feedback should be used as an indicator, but not the final arbiter:

To me it’s kind of a suggestion. I understand what the system gave me it’s something that I can consider. But the decision is mine, right? So, I would make decisions, whether I wanted to look at the others beyond what was given. Or I will look at some of them and not all of them. Because I feel that the decision should be mine. (D2)
Several participants considered the measurement may be a useful indicator of learner engagement, that could be useful to educators and course designers who are observing which parts of a MOOC attracts more in-depth attention, as well as learners seeking what may be described as discussion ‘hot spots’ – comments within the forum that other learners appear to consider more important than others:

I found that it kind of focused on critical thinking, which I hadn’t really thought about. And what would be interesting would be to – when you’re designing the course – to be thinking in those terms. I suppose when you’re designing a course, you’re looking at the getting material across and doing the job, or whatever, but you’re not thinking critical thinking as a way of engagement…what would be interesting for the people who are doing that to maybe reframe your critical thinking metrics as engagement metrics. And I think that could be useful. People are thinking critically, it could be that they’re more engaged with the topic. (B2)

Well, if it was educators then I think yes, it would give a very useful gauge as to how effective the MOOC material was that they designed, I think that would be very helpful to them…Basically, it's just given like a priority, helping to prioritise what they were looking at. (C2)

6.4 Discussion

As presented in the findings section above, moderating MOOC comment forums offers a number of challenges for educators and mentors. While learners’ critical engagement in forums is considered important, and is acknowledged by participants as useful in order to respond to learners, and comprehend the effectiveness of their pedagogy, it also carries certain implications of overall importance to RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments? Measuring the algorithms’ success in this case study requires reflection on three overarching themes: moderators use of comment forums, how they identify critical thinking, and their attitudes towards the automatic ratings of levels of critical thinking. In this section I explore how responses from participants in this study compare with findings in the literature in order to provide a broader context for these themes.

6.4.1 Using the comment forum

The role of the forum was seen as central to supporting learners and engaging them at deeper levels in the content. Learners’ intrinsic motivation was identified by some participants as essential for them to engage with MOOC content at a deep level, and moderators’ support had an important part to play in encouraging this. This accords with Cormier and Siemens’ (2010) assertions regarding the vital role of critical discourse in the development of knowledge within MOOCs.
All moderators used questions to stimulate deeper thinking, several reported correcting learners’ misunderstandings, and one reported providing links to further information as their key activities within comment forums. As described earlier, these activities of prompting dialogue, correcting misinterpretations, and suggesting reading, are at the core of the two main features of the teaching presence model: facilitation and direction (Anderson et al., 2001; Garrison, 2007). However, this study did not uncover any other form of teaching presence. For example, there were no reports of participants simply acknowledging comments, attempting to seek consensus, or summarising debates.

A possible explanation for this may be one given by some participants, and reported elsewhere: that they felt constrained by concerns that they may inadvertently discourage learners through excessive interventions (FutureLearn, 2017b). Notwithstanding this observation, participants not reporting engagement in all areas of facilitation and direction does not mean that they did not actually engage, or consider it worthwhile. Participants reported engagement appears to confirm agreement with Ross and Bayne's (2014) contention that a MOOC should be a taught experience, rather than “a bunch of resources on the Web” (p. 45).

However, all participants acknowledged that they were only seeing a fraction of the comments in the forum, which limited their attention to teaching presence and compromised their ability to provide personalised learning, a concern reported extensively in similar studies (Gerber, 2014; Evans and Myrick, 2015; Haavind and Sistek-Chandler, 2015).

6.4.2 Identifying critical thinking

When reviewing and responding to comments, moderators reported that they focused on learners’ questions, misunderstandings of the course content, and also indications of learners’ novel reflections on the content. Despite some participants asserting the primacy of the ‘informative’ and knowledge transfer objectives of the MOOC, all acknowledged the desirability of learners reflecting on, and demonstrating, their understanding within the comment forum. Participants identified learners’ critical engagement with the MOOC, and adopted pedagogical terms associated with Bloom’s Taxonomy as well as Community of Inquiry: Cognitive Presence to describe their experiences.

In terms of Bloom, participants recognised learners’ understanding when they were observed to have applied their knowledge in different contexts, undertaken informal analysis, and carried out evaluations of the different technologies under discussion. Similarly, participants identified learners’ comments that related to behaviours associated with Cognitive Presence stages. For example, learners’ questions may be classified as triggering events, personal stories as exploration, and the introducing new ideas as integration. While none identified comments associated with higher levels of thinking the Bloom (e.g. synthesis and create), or Cognitive Presence (resolution), this is
consistent with the literature where these categories are rarely found in online discussion forums (Kovanovic et al., 2017).

6.4.3 Accuracy of automatic rating

While acknowledging some errors, all participants agreed that the automatic rating they were provided with were reasonably accurate, with two participants estimating high levels of agreement. In addition, the tendency for participants to suggest alternative ratings for no more than one category different from the computed rating, reinforces the ordinal, rather than nominal nature of these levels. This further affords a degree of confidence that the automatic rating method described in Chapter 5 provides dependable feedback.

Participants reflection in their initial interviews indicates that their performance, and what they look for in learners’ comments, closely aligns with pedagogic theory articulated in teaching presence, cognitive presence, and Bloom’s Taxonomy. Having established this association, their further reflections on the accuracy of automatic rating in the second interviews adds credence to the assertion that the ratings are indeed reasonably accurate, and a useful indication of the level of critical discourse.

However, some caution should be attached to this assertion. While one participant expressed satisfaction with the accuracy of ‘off-topic’ (zero rated) comments he was presented with, several participants raised concerns that they were more closely associated with first person narratives than ‘off-topic’ interjections. An explanation for this may stem from the exclusion of comments containing less than 50 words. Because these brief comments are considered ‘unsafe’, they were omitted from the automatic rating process, which meant that participants were only shown longer zero rated comments. The omission of shorter comments may have skewed the automatic feedback, so that a higher ratio of inaccurately rated comments was presented to participants, than if all comments were included. In addition, my review of these comments also agrees with participants that raised the issue – many zero rated comments were not off-topic, but simply contained a high percentage of personal pronouns.

At the other end of the scale, the absence of participants reporting learners’ comments that associated with CoI’s resolution phase, or higher levels of Bloom’s Taxonomy, suggests that the rating system does not fully align with either method, and places further uncertainty on the meaning of comments rated as 4 (the highest score available). There are two main explanations for this potential vagueness. Participants may simply not have recalled the highest-level comments they encountered, or may not have considered pedagogical theory during their interviews.

Uncertainty surrounding the meaning of different levels within the scale suggests the value of future research. This should explore and closely evaluate the automatic rating system, and the manner in which its levels associate with the content of comments.
As reported in other studies (Haavind and Sistek-Chandler, 2015 cited Blackmon, 2016), participants also acknowledged the value of some form of assistance with monitoring learners’ comments. However, this finding combined with uncertainty on the meaning of the automatic ratings suggests that participants’ acknowledgement of the accuracy of these rating may have been partially influenced by a desire to receive any form of potentially useful feedback. Moderators working within the FutureLearn platform do not normally have access to analysis tools beyond what is available to learners (e.g. the interface provides feedback on comments with high numbers of ‘likes’), and research into psychology in the workplace indicates that “any feedback is more effective than none” (Gruneberg and Oborne, 1981, p. 2). This introduces the possibility of bias in the assessment of the automatic scores, and a willingness to disregard some inaccuracies. While this issue cannot be completely disregarded, the IPA method adopted in this study treats participants as expert witnesses of their experience, and my belief is that they provided impartial responses to my enquiries.

One participant raised the issue of the possibility of in-built bias in the rating method, and the potential of this method to neglect possible outliers. This thesis, along with the many Automatic Essay Scoring (AES) studies that have been undertaken over the past 60 years, strongly suggests that people who have attained higher levels of education (e.g. including raters employed to rate forum comments) tend to give preference to texts that contain high word counts, long words, and greater use of ‘academic’ language. Therefore, it is almost inevitable that an automatic rating method that filters and rates according to these general parameters will result in high levels of approval from people who are trained to support learners.

Conventional approaches to assessment often have the unintended effect of diminishing learner confidence in their abilities (Grenfell, 2012), and concerns regarding digital exclusion, and the unsuitability of mainstream measures of literacy have been raised by many educational researchers. Garcia and Pearson (1991) assert that mainstream measures of literacy distort the assessment of students from diverse linguistic, cultural and/or economic backgrounds, and suggest measures prioritising evaluations that take these influences into account. More recent research into linguistically-diverse MOOCs by Reilly et al. (2016) concludes that AES tools show lower validity when scoring non-native English speakers, and recommend that MOOC designers take into account diversity in English literacy skills.

In a global learner environment of MOOCs, these concerns need to be tackled to ensure that the needs of learners from diverse backgrounds are addressed. Following Garcia and Pearson, further research should be undertaken into measures that improve teacher and administrator awareness of diversity in language and culture; refine assessment criteria so that it more closely reflects learners’ actual performance; and improve large scale tests to make them more authentic.
Nevertheless, it is worth noting that participants reported their awareness of the imperfect nature of the automatic ratings, and that they should be treated as suggestions, rather than final decisions on a learners’ critical thinking level. This accords with data science literature that advises caution on the use of algorithmic decision making (Derman and Wilmott, 2009; O’Neil, 2016), and Garcia and Pearson’s suggestion that educators should assume a realistic view of tests, which recognises the transitory nature of their reliability.

Participants also suggested refinements to the rating method, including searching for topic-related key words in comments, and counts of the number of replies to a comment in the algorithm’s calculations. These approaches have been successfully adopted in learning analytics research (e.g. Hsiao and Awasthi, 2015; D’Mello et al., 2008; Yang, Wen and Rose, 2014; Waters et al., 2015), and are useful suggestions for further work, but beyond the scope of this thesis.

6.4.3.1 Reflections on a different approach

In Chapter 5 I suggested an alternative approach to the machine learning process I originally adopted to produce the algorithm applied to comments evaluated in this MOOC case study. Rather than the 50:50 training and testing data split I implemented on a reduced dataset (those comments containing 50 or more words), on reflection, a 70:30 split of all comment data was thought to have greater potential for improved accuracy. When the model algorithm produced from this alternative approach was applied to its test set, a Cohen’s $k$ of 0.4417 and an Intraclass Correlation Coefficient score of 0.85 was achieved – a significant improvement on the original approach ($k = 0.2083$, ICC = 0.695).

The outcome from this exercise suggests that using the algorithm derived from a 70:30 split of all data would produce different results from the classifier actually used in the ‘live’ MOOC case study. Re-evaluating all of the comment data from the MOOC case study is beyond the scope of my thesis, but, applying the new algorithm to the first week’s comment data from the MOOC provides an impression of what may have occurred had the alternative approach been adopted.

Comparing predicted ratings using the original algorithm with those produced using the new algorithm gives a $k$ of 0.2609 and ICC of 0.643, suggesting some differences in agreement. Exploring the different ratings achieved with the new model in detail reveals that the new algorithm has a tendency to rate comments either the same (83/183 – 45%) or higher (73/183 – 40%) than the earlier version. Most of the new ratings were either the same or within one level of the previous rating (157/183 – 86%).

While the outcomes from the new model algorithm are different from the earlier version, it is uncertain if applying the new algorithm to the data used in the ‘live’ MOOC case study would have produced significantly different responses from case study participants. Certainly, the one participant to provide detailed feedback on some of the automated ratings indicated a greater
tendency to ‘up-rate’ comments than ‘down-rate’ them, but within the terms of this study it was impossible to determine if the new algorithm would have provided significantly improved accuracy in practice. However, the improved mathematical accuracy calculated for this model strongly suggests its adoption in further studies.

6.5 Limitations and potential issues

The value of research is generally perceived to reside in the representative nature of its sample and its success at supporting broad inferences. Because of its small scale, research projects which adopt the IPA method cannot normally make claim to the empirical generalisability of its findings. However, in this study I appeal to its theoretical generalisability, whereby moderators in similar situations can learn from the experiences of the participants in this study (Smith and Osborn, 2015).

The theoretical generalisability of this type of study is also open to question, particularly when interviews are used as the primary means of data collection. Firstly, the accounts produced from open-ended questioning may bear little relationship to what participants may say and do in their everyday lives. In this study, the participants selected themselves from within the sample I had chosen for convenience. They were selected on a ‘first come, first chosen’ basis and cannot be viewed as representative of MOOC moderators in general. In addition, participants may have seen my position as a postgraduate researcher as being privileged (although I had no involvement in the production or running of the MOOC, or in the running of the MOOC platform) and provided answers they thought were expected, rather than accounts of their actual experience.

The interviews for this study were collected ‘out of context’ from the reported activity, but it is the MOOC comment forum context which governs how the participants work as moderators. There is also a reported tendency to see participants as “disparate individuals” who have no social interaction with each other (Bryman, 1988, p. 39) which overlooks the context in which the reported processes are learnt and developed. Further, all interviews are subject to the benefit of hindsight where participants may feel encouraged to create narratives and superimpose meaning and structure onto what may have been poorly thought out and disorganised processes.

All this implies that this study can have nothing to say about the actual experience of participants or MOOC moderators in general as they go about their work. However, within the boundaries of IPA methodology, I view participants’ narratives as expressions of plausible stories that describe the activities in culturally understandable terms. They may not be ‘true’ as in the presentation of factual, unbiased information, but they represent ‘cultural stories’ that make the actions of the participants explicable to those who may not understand. Richardson suggests that:

“Participation in a culture includes participation in the narratives of that culture, a general understanding of the stock of meanings and their relationships to each other.” (Richardson, 1990, p. 24)
Within this understanding, I assume the authenticity of the participants’ related experience, and consider that the participants trusted the good intentions of the study and were assured by the guarantee of anonymity.

By applying an organised and controlled approach to interpretation, and by clearly, precisely and openly describing the data collection and interpretation techniques used in this study, I aim to remove doubts regarding the meticulous nature of this research. By comparing my literature review with the ‘real world’ experience of the MOOC moderators who took part in this study we gain a fuller understanding of the issues surrounding learning within MOOCs, and so are able to make a useful contribution to the field.

6.6 Summary

This small-scale case study set out to substantiate the findings presented in the previous chapter, and add depth to the resolution of RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments? In the previous chapter I confirmed the moderate accuracy of the model algorithm, based on quantitative data. In this chapter I set out to reinforce this confirmation through the analysis of a qualitative study. Specifically, I sought to discover if participants mentoring an on-going MOOC found the outputs of the model algorithm generally useful, and if automatic classifications of comments presented to them aligned with their assessments of learners’ depth of thinking. The analysis of the work of MOOC moderators, and evaluation of participants’ responses described in this chapter suggests that, with some caveats, the ratings were generally useful to participants, and aligned with their evaluations of learners’ levels of critical thinking. The caveats identified by participants were: uncertainty regarding the individual meaning of the rating levels; possible bias in favour of any feedback, regardless of accuracy; and in-built algorithmic bias that favours conventional attitudes to literacy. Nevertheless, participants also recognised the temporary and advisory nature of the automatic ratings.

The IPA method employed by this study has revealed a number of relevant issues and can tell us a great deal about MOOC moderators’ attitudes regarding the automatically rated comments they were presented with. It has proved to be a useful approach in gaining a rich, in-depth picture of their experience of an important area of their work. I have acknowledged the limitations of this study but argue that difficulties in generalisability can be overcome by its relatability to other similar studies.

Conclusions can now be made from all the studies reported in this thesis, and recommendations for further research will be made in the following chapter.
Chapter 7: Overall discussion and conclusions

This chapter includes the final discussions and conclusions based on the data collected and analysed in the preceding chapters. To review the work that has been undertaken, a summary of each study is presented. Findings are evaluated and integrated to gain a more thorough understanding of the usefulness of automated analysis of MOOC learners’ comments for those involved in developing, designing, and running MOOCs. A discussion of these results follows, before limitations and future work opportunities are presented, and final conclusions made.

7.1.1 Results Review

This thesis presents a number of studies exploring the reliability of pedagogical content analysis, associations between the outcomes of this analysis and linguistic and interactional indicators of critical thinking, and the validity and usefulness of automated analysis for those involved in running MOOCs, especially MOOC moderators. Following the pilot study, which applied four different coding schemes used for pedagogical content analysis to learners’ comments in MOOCs and demonstrated statistically significant associations between levels of critical thinking identified in comments and linguistic categories, a larger scale investigation was carried out using two content analysis methods and multiple human raters in order to validate these findings and produce training and test data sets for use in building a machine learning classifier. A case study of MOOC moderators was undertaken to gain improved understanding of the usefulness of the classifier, and different perspectives were analysed: moderators’ attitudes to their role, their views on the importance of identifying levels of critical thinking in learners’ comments, and the accuracy and overall usefulness of the automated ratings.

The overall study set out to answer four key research questions:

RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forum comments? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?

RQ2: Are linguistic content analysis measures significant indicators of levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?

RQ3: To what extent do typical measures of attention to learning (such as social interactions) indicate levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?
RQ4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments?

An outline of the overall research design for my thesis is shown in Figure 34 below.

![Figure 34: Overall research design.](image)

### 7.1.1 Pilot and large scale studies

An initial pilot study contributed to understanding the limits of the use of ‘likes’ as indicators of on-topic engagement, and confirmed associations between learners’ language use and levels of critical thinking. It strongly supported the value of further content analysis using methods based on Bloom’s Taxonomy and Community of Inquiry, different datasets from a variety of courses, covering diverse subjects, with contributions from different participants, and analysed by multiple raters, in order to establish widely applicable techniques.

This analysis was undertaken in a large-scale study, which confirmed some of the findings from the pilot, and also identified distinctive features of learner comments suggestive of critical thinking. Both studies used pedagogical content analysis methods to manually rate comments posted on MOOC discussion forums from which levels of critical thinking were derived. Strong correlations with linguistic and interaction features were identified, as were close correlations between the content analysis methods used.

### 7.1.2 Building machine learning classifier

A ML method was adopted to build an algorithm that would automatically rate comments based on predicted levels of critical thinking. A rating scale derived from a pedagogical content analysis method used in the large-scale study (CoI: Cognitive Presence) was adopted to more closely reflect...
the ordinal nature of the rating method (indicating five stages, from low to high levels of critical thinking). Labelled data derived from manual pedagogical ratings applied in the large-scale study was used to build a new classifier adopting a highly-regarded ML method (Random Forest). Applying the new classifier to previously rated comments in a test data set produced a good estimated level of reliability (ICC = 0.695).

Further investigation undertaken after the MOOC case study demonstrated that different approaches to algorithm training produced a substantially different outcome, achieving a higher estimated reliability value (ICC = 0.85). However, because perfect accuracy in this setting is not required, and the critical thinking scale developed from the content analysis methods used in the large-scale study acts as an ordinal list, the reliability results achieved with the earlier classifier were considered good enough to test in a ‘live’ MOOC setting.

7.1.3 MOOC case study

The final study was undertaken to establish the consistency of the classifier when applied to rating unlabelled comments, and evaluated by MOOC moderators. I sought to discover if participants mentoring an ongoing MOOC found the outputs of the model algorithm generally useful, and if automatic classifications of comments presented to them aligned with their professional assessments of learners’ levels of critical thinking.

Analysis of MOOC moderation, and evaluation of participants’ responses suggested that, with some qualifications, the ratings were generally useful, and aligned with their appraisals of learners’ levels of critical thinking. While participants acknowledged the temporary and advisory nature of the automatic ratings, some caveats were identified: uncertainty regarding the individual meaning of the rating levels; possible bias in favour of any feedback, regardless of accuracy; and in-built algorithmic bias that favours conventional attitudes to literacy.

7.2 Research questions

My thesis sets out to answer the four key research questions outlined below.

7.2.1 RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable when applied to MOOC discussion forum comments? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?

Translating MOOC comments into comparable scores based on multiple pedagogical frameworks is a substantial research activity. In the pilot study and large scale experiment, a single coder (in the former) and a group of seven coders (in the latter) achieved a high degree of reliability using
different pedagogical analysis methods (e.g. $r = 0.909, p < 0.001$), which afforded confidence in the generalisability of these methods in future studies. While the coding schemes used for pedagogical content analysis have different theoretical foundations, and have been developed to evaluate different aspects of learning, when applied to MOOC forum comments, and correlated against language categories, sentiment and 'likes', there was a high level of similarity in the measurement of levels of critical thinking. This strongly indicates that very similar levels of critical thinking were identified in these studies using different methods. This further suggests a high degree of interchangeability between these methods in this setting.

7.2.2 RQ2: Are linguistic content analysis measures significant indicators of critical thinking in MOOC discussion forum comments, as identified by pedagogical content analysis?

Word count and first-person singular pronouns were identified as convincing indicators of critical thinking when correlated with pedagogical ratings, with causal words ($r = 0.573, p < 0.001$), power ($r = 0.369, p < 0.001$) and all pronouns ($r = -0.372, p < 0.001$) providing moderate results. Other word categories (differentiation, negation, cognitive process, words per sentence, auxiliary verbs, power, words containing six letters or more, conjunctions, negative and positive emotion, prepositions, and affiliation) provided mixed results, suggesting a supporting role for these categories in building a critical thinking classifier.

7.2.3 RQ3: To what extent do typical measures of attention to learning (such as social interactions) indicate levels of critical thinking in MOOC discussion forum comments, as identified through pedagogical content analysis?

In the pilot study, the LIWC measure of positive sentiment was found to have a strong correlation to levels of critical thinking, but no significant associations with 'likes' were discovered. However, the large-scale experiment producing significant results for both measures of sentiment and 'likes', although they were weakly correlated with critical thinking measures (maximum: $r = 0.298$, $p < 0.001$). This suggested subordinate roles for these measures in building a critical thinking classifier. Ultimately, both negative and positive sentiment attributes were proven to be useful in this process, but the 'like' attribute was not.

7.2.4 RQ 4: Can machine learning algorithms be trained to successfully measure levels of critical thinking in MOOC discussion forum comments?

The estimated reliability of the ML algorithm produced in the classifier building phase of my research provided a ‘good’ ICC value of 0.695 (Cicchetti, 1994). This was considered to be sufficiently reliable to be applied a ‘live’ MOOC environment, where MOOC moderators provided expert evaluations of the usefulness of this automated classification.
MOOC participants mentoring an on-going MOOC found the outputs of the model algorithm generally useful, and reasonably well aligned with their assessments of learners’ levels of critical thinking. Of the four participants, one estimated that 80% of the automatic ratings appeared accurate, and another’s in-depth feedback on 67 rated comments suggested 93% agreement with the classifier. This outstanding and unexpected result from a very simple algorithmic model containing just 14 attributes, suggests a positive answer to RQ4 – algorithms can indeed be trained to successfully measure levels of critical thinking in this setting.

However, some qualifications to this outcome were identified regarding the meaning of the rating method, attitudes to feedback, and bias, which are discussed in the next section.

7.3 Discussion

My thesis is firmly grounded in the multidisciplinary field of Web Science, combining as it does two distinct disciplines: research practices common in Computer Science (Machine Learning) and investigations of Educational theory and practice (e-Learning pedagogy). This blend of disciplines has many similarities with Learning Analytics research, and this connection is reflected throughout.

With its emphasis on developing data-driven decision making and interventions in teaching and learning, LA is having a huge impact on pedagogy. The widespread adoption of MOOCs by higher education institutions is not only making high quality learning opportunities available to hundreds of thousands of learners worldwide, it is also having a disruptive effect on teaching practice. Through the adoption of LA methods to explore learner interaction in comment forums, my thesis focuses on a small part of this disruption (the problems experienced by MOOC moderators when managing huge numbers of learners’ comments), adds a measure of explanatory power to the predictive accuracy of statistical and machine analysis, and offers some solutions that have potential to assist those involved in planning, running, and learning in MOOCs.

7.4 Developing an automated analysis for online learning

The overwhelming number of MOOC learner comments is an issue of concern for those developing large-scale learning environments. MOOC moderators have a vital role in orchestrating the learning experience, encouraging learners to take part, and stimulating critical discussion of study topics. But successfully managing the large numbers of comments typical of MOOC comment forums is challenging, and often impossible.

While much research has been undertaken that explores critical discourse in formal online learning environments, very little has focused on the informal MOOC context. This thesis set out to study this setting and explore the potential for developing an effective automatic method that might assist MOOC educators and mentors in their decision-making processes.
My investigation of levels of critical thinking in MOOC discussion forums employed multiple coding schemes used for pedagogical content analysis as a means to rate levels of critical thinking in learners’ comments. By doing this I aimed to identify reliable, pedagogically sound approaches to rating comments that could be used to build an automatic classifier and which would be readily understood by MOOC moderators, and to discover similarities in how raters evaluate comments using these approaches, so that a synthesised, automatic rating method could be developed that was both meaningful and effective.

Using robust methods employing multiple raters, I found that the coding schemes used for pedagogical content analysis could be reliably applied to MOOC learners’ comments, and that there were strong correlations between all methods used. This clearly indicated close similarities between each method’s measurement of levels of critical thinking, supporting my contention that, in the context of measuring learners’ depth of engagement with a topic, use of these methods was essentially interchangeable. Further, I identified convincing linguistic and interactional associations with these content analysis methods that facilitated the development of a machine learning algorithm that could use linguistic analysis to automatically rate levels of critical thinking in comments.

Testing the machine learning algorithm achieved good reliability when applied to a prepared test data set, and when applied to unlabelled comments made by learners in a live MOOC, and evaluated by MOOC moderators, the model algorithm was considered to have provided useful, actionable feedback. In addition, the tendency for participants to suggest alternative ratings within one level further reinforces my contention that these levels should be treated as ordinal rather than nominal, and therefore using ICC as a measure of reliability is more suitable in this setting than Cohen’s kappa.

Through their descriptions of what they looked for in learners’ comments when deciding to intervene, participants in the case study also identified key pedagogical concepts that are intrinsic to the content analysis methods used to build the classifier, thus confirming the importance of identifying and measuring levels of critical thinking to their work.

In Chapter 2 I outlined concerns about the use of LA as a means to intervene and guide learners, including uncertainty regarding the validity of automatic methods, and ethical concerns regarding bias and privacy issues.

### 7.4.1 Validity of automatic methods

While no data model can completely or accurately model reality (Baskin, 2014), it is important that models use data that are fundamentally related to the subject area; that there is a clear and relevant relationship between the data and what it is being used to represent, and that it is not simply used as a convenient proxy due to difficulty collecting more appropriate data. In my thesis, I have used
learners’ written contributions to comment forums as a means to measure levels of critical thinking. The use of this form of data as a proxy for thinking is well-established, in language analysis, disciplinary studies, and in learning analytics. In addition, I have adopted robust measures to establish the validity of my approach: by using several content analysis methods, multiple raters, by adopting established measures of reliability, and by testing the resultant automatic classifier in live environment, and seeking the views of potential users.

In their research into the use of algorithms in medicine, Cabitza, Ciucci and Rasoini, (2017) argue that as uncertainty is a fundamental trait to diagnosing illness, the development of ML as an aid to clinical decision making needs to acknowledge this uncertainty, and use automated feedback with caution. The same level of uncertainty can be said to exist in education. For example, although in my large-scale study (Chapter 4) agreement between raters was observed to be relatively high, there were some areas of substantial disagreements between assessments in a number of instances.

As reported in my MOOC case study, some participants observed that there is a strong case for treating machine derived value judgements not as definitive evaluations, but as an adjunct to the decision-making process. Following Cabitza, Ciucci and Rasoini, I recommend that when developing automated decision-making systems for education, it is essential that relevant stakeholders (including MOOC moderators) are closely involved. Further, time and effort should be invested in evaluating these systems, both prior to deployment and for the duration of their use, and the effects of automated decisions on moderator and learner outcomes require continued, on-going appraisal.

### 7.4.2 Ethical considerations

Personal privacy may be infringed if personal data is shared in a way that goes against the reasonable expectations of data subjects. In LA, data processing has two distinct implications for personal privacy which have the potential to breach personal privacy (Hildebrandt, 2017). In first order learning analytics, inferences regarding identifiable learners are made which may be acted upon. In this setting, if institutional safeguards are not sufficiently strong, personal data may be retrieved by people who do not have a legitimate need to access.

In terms of my study, using personal data to identify comments made by individual students to enable moderators to intervene more effectively requires built-in safeguards to ensure that data is not processed or shared in a way that violates learners’ reasonable expectations. It is important that personal data is not shared out of context, and that only those who can justify access are given privileged access.

Less stringent safeguards in second level learning analytics are required when exploring patterns derived from pseudonymised data (i.e. data which does not identify specific individuals). This enables the identification of patterns in data that link attributes, behaviour, context and other features to
measures of performance; explores what kind of attributes correspond to what kind of performance; and make predictions on which interventions may improve future learner performance in different circumstances.

However, unintended privacy violations may result from second order learning analytics. As has been demonstrated in my research, linguistic analysis can reveal levels of critical thinking and academic discipline, and research elsewhere has shown that it may also identify other personal attributes (e.g. ethnic background, religious affiliation, economic status, or geographic location). Despite the use of pseudonymised data in second order learning analytics, there remains the possibility of re-identifying data subjects through the combination of a variety of non-personal data sources. This implies that before operationalising the proposed method, in-built safeguards are required which, like those proposed for first-order analytics, prevent the sharing of data out of context.

7.5 Limitations and Future Work

The question of what constitutes a fair algorithm has been raised in recent years (Koene, Webb and Patel, 2017), and the use of human raters in this research, who while not trained to make unfair or prejudiced choices in their rating of comments, may have been influenced by conscious or unconscious partiality which favoured certain types of writing above others. By using these choices to build the ML model, the resultant algorithm can be said to effectively replicate these biases. Once operationalised and applied at scale, the algorithm carries these biases into decisions and recommendations that may, erroneously, be considered as objective, conclusive reports on the ability of learners.

However, as well-educated students, experienced in communication styles favoured in higher education, the raters in my research had many similarities with educators and mentors whose job it is to evaluate learners’ comments. Therefore, the automated recommendation model derived from raters’ educated opinions can be said to mimic the behaviour of human moderators. By making the development of the model explicit, moderators may be directed to read an automatic recommendation as guide to be interpreted, rather than a definitive statement.

Using an interpretive phenomenon analysis approach in the final stage of this research also introduces a form of bias. As has been noted, researchers cognitive bias may subconsciously influence case study participants (Rosenthal, 1976). In my research, I was surprised to observe very strong correlations between my algorithms’ output, and participants’ evaluations of them. While these appears to be genuine responses, a more rigorous approach adopting double blind research strategies is recommended for future studies.

The model I developed in this research is based on methods that are widely known in education, however, the case study revealed some concern among participants regarding the meaning of the
rating system. To address this, further work is required to both investigate and improve methods of explaining how the model operates.

While my research demonstrates the potential for improving moderators experience, the declared outcome of LA is to use research findings to improve learning. The lack of reliable evidence demonstrating this has been identified (Ferguson and Clow, 2017), and a key focus of future research should be to ascertain the impact automated comment evaluations have on moderator behaviour, and the consequences of this for learners.

### 7.6 Final Remarks and Conclusions

The usefulness of automated feedback for moderators, and the significance of automatically evaluated comments for learners, is a complex issue. Because the large amount of data generated by interactions within MOOCs is acting as a significant ‘push factor’ in the development of methods that use data to improve the experience of online learning, there is a danger that analysis is limited to easily collected data. Using data typically generated by Web interactions (e.g. views per page, time on page, link clicks, and number of comments) as proxies for learning, may not accurately reflect the critical activity undertaken by learners. On the other hand, by adopting pedagogical methods to analyse learner comments, and building an automatic method to evaluate these interactions based on established theory and practice, my research maintains a close relationship with learning activities, and provides feedback which may ultimately be used to improve outcomes for learners.

### 7.7 Contribution

My thesis presents a number of contributions to current research, as well as practitioner and policy debates in areas that affect the development and use of automated analysis methods in Computer-Supported Collaborative Learning.

My examination and evaluation of Content Analysis methods, my novel approach to comment analysis, and the new datasets created, contribute to current deliberations in Learning Analytics and CSCL research. Through the use of robust methods, this thesis contributed to establishing the validity of automatically measuring levels of critical thinking in informal learner comments. My study confirms previous research identifying word count, first-person singular pronouns, causal words, differentiation, negation words, words per sentence, negative emotion words, and cognitive process words as moderate to strong indicators of levels of critical thinking. In addition, the close correlation of ratings using different coding schemes used for pedagogical content analysis strongly suggests their interchangeability in indicating levels of critical thinking in informal learning environments.
In my user study, where practitioners reflect on their needs when running MOOCs, my pedagogical theory-based automated analysis contributes to on-going debates on the application of automated feedback in learning. My study confirms previous research which identified the volume and velocity of learners’ comments as a significant challenge to MOOC moderators. For moderators, struggling to monitor and respond to learners’ comments, the timely identification of those in need of pedagogical support is essential. By identifying strong correlations with LIWC2015-based proxies for pedagogical activity and the coding schemes used for pedagogical content analysis, and demonstrating the potential for automatically rating comments in terms of levels of critical thinking, this thesis provides practitioners and policy makers with a coherent method to evaluate learner engagement.

Finally, through the adoption of well-known pedagogical analysis methods and using robust approaches to identify correlations between these analyses and the linguistic features of comments, I have taken substantial steps to ensure that the automated decisions described in this thesis are intelligible to potential users. By making a substantial contribution to debates regarding the transparent development of algorithmic decision-making methods, this thesis provides policy makers with a readily understandable and explainable processes.
Appendix A

Participant Information Sheet

<table>
<thead>
<tr>
<th>Ethics reference number: ERGO/FPSE/25220</th>
<th>Version: 2.0</th>
<th>Date: 27/01/2017</th>
</tr>
</thead>
</table>

Study Title: Does automatic classification of MOOC comments align with educators' evaluations of learners' levels of critical thinking?

Investigator: Tim O’Riordan

Please read this information carefully before deciding to take part in this research. If you are happy to participate you will be asked to sign a consent form. Your participation is completely voluntary.

What is the research about?
This is a PhD student project which aims to inform understandings of how learning analytics and automatic analysis of comments made in MOOC discussion forums, may inform online teaching practice and learning design. The study will involve semi-structured interviews in which participants will discuss their views on the topic in a friendly environment.
At the end of the study, you will receive an email attachment containing the study findings and see how your data was used.

Why have I been chosen?
You have been approached because you have been identified as a relevant participant in the MOOC development and implementation process in your university.

What will happen to me if I take part?
You will participate in two interviews with the investigator, which will take about 60 minutes in total. The first interview will relate to your attitudes to and understandings of monitoring learner activity in MOOCs in general and indicators of critical thinking in particular. If you wish, you may share with the investigator any documentation related to your academic/teaching practice which are relevant to the interview discussions.

Are there any benefits in my taking part?
It is expected that the study will add to current knowledge about learning analytics which could (for example) contribute to university policy formulation on the subject.

Are there any risks involved?
There are no particular risks associated with your participation.

Will my data be confidential?
Your data will be held on a password protected computer, and used only in accordance with the Data Protection Act (1998). In addition, the data will be anonymised by separating identifying data. Your data will be linked to your consent form by an anonymised code. Your data will solely be processed/analysed/edited by the researcher and/or their supervisor(s). Audio recordings will be
Appendix B

Monitoring MOOC Learners

Title: Interview Schedule for MOOC Moderators 1
From: (tba)  
Date: 13/02/2017

Start:
Investigator distributes consent and information forms and invites participant to read both before signing. The participant will be asked if they give consent for an audio recording to be made.

Questions:
- Can you tell me a little about your previous experience of:
  - Teaching
  - Online teaching/learning
  - MOOCs (Massive open online courses)
- What (if anything) did you find particularly surprising or unexpected about the MOOC this week?
- What do you consider to be the key attributes of a successful MOOC learner?
- How do you encourage learners to engage in critical thinking?
- What are your thoughts on the importance of monitoring learners critical thinking?
  - How important is monitoring learners’ level of critical thinking compared with monitoring other forms of engagement?
  - What features of learner comments indicate engagement with critical thinking?
  - What platform features assist with monitoring learners’ critical thinking?
  - What was the most useful platform feature that assisted you in monitoring learners critical thinking?
- Do you have any further comments?

Tim O’Riordan
Direct tel: +44 (0)735209156
Appendix C

Monitoring MOOC Learners

Title: Interview Schedule for MOOC Moderators 2
From: (tba) Date: 13/03/2017

24 hours before interview:
Participants will be sent a report detailing analysis of comments posted during the period they moderated the MOOC, and asked to reflect on the relationship of this analysis with their experience.

Start:
The participant will be asked if they give consent for an audio recording to be made.
The investigator will provide a summary of the previous interview including the key outcomes reported by the participant.

Questions:
- Are you satisfied that the investigators’ summary of your previous interview is accurate?
  - Are there any omissions or errors?
  - What, if anything, would you like to change or add?
- What are your thoughts on the comment analysis provided?
  - What, if any, aspects of the analysis do you find confused or unclear?
  - How accurately does the analysis reflect your experience of the week’s activity?
  - Does the analysis reveal any aspects of the discussion forum that were overlooked during your period of moderation?
  - What aspects of the analysis do you find useful?
  - What aspects were not useful?
  - How could the analysis be improved?
  - How could the visualisation of the analysis be improved?
- What (if anything) did you find particularly surprising or unexpected about the comment analysis?
- Do you have any further comments?

Tim O’Riordan
Direct tel: +44 (0)735209156
Appendix D

Transcript of first interview with Participant A
27 February 2017

Q1: First of all, would you mind telling me a little bit about your previous experience in teaching?
A1: Well I am a professor since now 20 years, more than 20 years and teaching computer science. The subject that I am involved in, in FutureLearn is teaching and accessibility. My other areas are human computer interaction, media informatics, so I would say quite a large area within computer science that I am covering. I am teaching lectures in front of 300 students but also 15 students, so in terms of size how many students I see. Normally it's really frontal lecturing, so standing in front of class and explaining things. I like to teach and ask questions so some students find that worth a remark, warning others, he's asking, otherwise I think I am not really the very best lecturer in that sense that maybe more work could help better the students. But as always, there's too much to be done. What I like very much and what I've been involved in is the use of e-learning platforms. I started with Blackboard in 2003 and had Moodle all the time until basically this last term where we finally moved one course to the system called OPAL which is here in the local federal state, it's actually the, let's say recommended e-learning platform that which I always avoided because of accessibility issues.

Q2: In the area of MOOCs, how much experience have you had in the Massive Online ...?
A2: Oh, only basically when we started the MOOCA project, one and a half years ago roughly. I started to look into this. I had heard about this in conferences, the contacts wanted to know a bit how it works and that's when the project really came in very handy to see how it works. So, the only course I really joined as a lecturer was last October/November and now this week and last week basically. As a let's say learner, I did one course in September last year.

Q3: In this, the past week where you've been working as an educator, as a mentor, did you find anything particularly surprising or unexpected?
A3: Yes, there were some people fitting too well into the scheme I would say. One person says well my husband is colour blind and I know this guy who is colour blind, it's very, very unusual that people are aware of this and can describe this relatively precisely. So that was unusual in that sense. You know one person really confirmed and basically said he's blind which was also never the case last time. Normally people don't say a lot about this if they are handicapped themselves. That was my understanding. So, this was a bit unusual, yes.

Q4: What is success in terms of MOOCs? What would be your ideal learner, if there is one?
A4: Let me describe a little bit what happened last October and November. Basically, I had asked my own students to join the MOOC and asked them to submit their quiz results to me separately since I could not get direct access to this information. So, I understood that it was quite, let’s say interesting and challenging also for the students to follow the MOOC. So, it was not as I would have guessed it’s obvious how it works and what type of, what’s the proper answer to the questions. So, I think it is, even on the level of the computer science students that I normally teach, a new piece of content that they are learning so that worked, it was not, let’s say too difficult, they all managed. And that brings me basically back to the question. In the course you see people who seem to be completely unaware of this topic and come from a side where they have some, I would say, house wife type of interests. Okay this accessibility sounds like you can do something for handicapped people when basically they realised the second term, digital means really about computer science and trying to understand how computers make inaccessible parts of things that you can do. So that’s basically the lower end that you meet these people for whom the course is a bit too demanding over the course, in the end I suppose. And then there are the others who come with a focus that is very in depth where they don’t see the big picture and only look into a more technical side. For them it is probably very difficult to motivate them to do the whole course. So, the ideal learner is one who has an interest in the topic in the sense that it is wide and not too narrow, and also on the other hand who is ready to invest time into looking a little bit beyond what’s in the course material and do some own analysis and some own extra look at maybe Wikipedia or somewhere else. So, I think this would be the perfect learner who is ready to do that.

Q5: How would you go about encouraging learners to do that, to engage in critical thinking?

A5: Yes, I am actually looking for these comments which are reflective in a sense and maybe come up with extra information but also maybe ask questions. So, I remember last time when the course ran first, I had a person who was really commenting in order to challenge also others and basically get some feedback. So, I thought if I replied to these people then others who read and don’t think they can do, cannot add anything more, they can benefit from this. So, I think the comments that come up with, let’s say, not in depth question but just reflective questions in the sense that they ask more about how that can be and how it is linked to something else, these are the best to show that they have learned something. On the other hand, one should not forget that there’s also a lot of factual knowledge. Simply knowing about different disabilities in our case and the variety of the requirements is not something that you only reflect on but you really have to go through this and really after the course you should be able to enumerate this and be able then to apply this knowledge in case you encounter a type of disability not covered in the course. Again, that means reflective thinking but in the sense of understanding the needs of people with a disability means also to understand what has been done for some people already and then be able to extrapolate from this and come up with a
way to say okay this is different but not everything is different, I should follow such an approach as this has proven to be useful in the past.

**Q6: What are your thoughts on the importance of actually monitoring learners’ critical thinking?**

**A6:** Yes, good question. Because of the structure of the MOOC, I think there should be an immediate feedback but I cannot ensure this because I have other things to do. [laughs] And since people are active at any time, they come from all over the world, I cannot be present all the time. So, in order to get the communication going, I think only a small window when people look it up and follow a kind of conversation. So, I haven’t seen a lot of chains of comments on a particular comment. Only basically maybe five to ten comments further where you find another reflection on a comment, or what is universally said, a repetition takes place and somebody says exactly the same what has been said before. Because they are not reading all the 50 comments that were sent before. So, I think that’s a technical problem in the sense that navigating through the comments means to read a lot of text and because the texts are short, they take a lot of time in order to be reflective yourself and come up with something that you combine. This is different to a lecture where you have maybe one to three questions in a row and there the discussion probably helps a lot of people and other people jump in and continue to discuss. While in the MOOC, there’s no such like, no such let’s say discussion going on in a sense. So as somebody who is providing feedback, I think it is important to find those who are really important that come up with something that hasn’t been covered and where my example also is worth reading for others. So, if I don’t see a good chance not only to echo and parrot basically, play the parrot, I’d better be quiet and look for better comments where I can make a stronger comment.

**A7: How important would you rate monitoring learners’ levels of critical thinking compared with monitoring other forms of engagement? I am thinking like “liking” comments or the kind of social aspects of the comments that you see?**

**Q7:** Yes, the “likes” are a confirmation in that sense and in terms of analysts, they are much easier to analyse, yes simply count them and see a lot of people were considering this comment and, but not many actually could come up with some additional comment on it. So yes, the reflections that some people make are not immediately the same as simply a comment. The point is really how can you measure this in a clear way that this has been a new thought or went into depth. [Coughs] Sorry.

**A8: What features of learner comments indicate engagement with critical thinking?**

**Q8:** One of the typical scenarios in my workplace, I have not done this or I have done this, and then they come up with additional thoughts that go beyond what’s in the course material. So that is the easiest thing, if they reflect on what they do in work. Then the next is other people they are aware of and some really manage to reflect
on the content itself without relying on another example. So these are the three categories that I find.

What does not happen is that educators [laughs] discuss together. This is something somehow it does not work. I am not sure why but when my colleague from Austria who is with me this week and was last week, no this week it's another colleague but last week then I read his comments and I maybe only once ever have commented in addition to that. So, in terms of one learner and two educators which is possible, theoretically, practically it doesn’t happen. I am not sure why.

A9: What would you consider to be the kind of main attributes of successful MOOC design?

Q9: Yes, that’s a good question because we are now moving some additional courses onto a Moodle platform and help to find out if the videos are the, let's say the selling point. When we prepared for our course, we made a lot of videos because it looked like this is attractive to learners just to get it like in TV and get some information that could be written in a page also. But also, can be explained by standing in front of a whiteboard and speaking about it and showing some slides. Often the videos gave the opportunity to talk to practitioners, to disable people themselves so it created the higher motivation for the people than just reading text without seeing the person. So, I think that's the selling point for us that we could show many more ways how disabled people work with computers and so far I don’t know of a course who did that before. I think that's why my students also this time, because I do already some exams, look to have benefit from that and have a better overall picture, and have understood a bit more I think.

A10: What specific aspects of the Learning Design you think worked well in the MOOC you're working on at the moment?

Q10: Oh, well the one key element that we tried in our courses that we would use a concept called Personas or later we called them user stories in order to build up a template person that has a particular disability in order to give the participants a human person's name and be able to reflect on colour blindness, on blindness, on deafness and things like that. So that we were hoping this would give them not more motivation but help them to structure the landscape and the knowledge. We tested this, maybe you don't know this but we published even on this approach and I tested this basically in my own class. If my students, based on this can analyse other types of disabilities then more actively. That worked quite well. They wrote additional essays and this was analysed and looked like they have understood much more about what it means to work in that area and not only talk about blind people but really get an understanding in the needs of people with a disability and only then look into what are possible solutions. It is more the subject now that I've tried to describe the term Learning Design in that sense, it is, well five weeks and from that side the discussion that we had in our group who decide those five weeks were fruitful. There was no overlap to my surprise despite that the development was in different areas, in different universities. We avoided overlapping because we made a
clear plan in the beginning and therefore there was no need for correction. Basically only just a few technical errors that had happened where links were missing and things like that. But otherwise I was surprised that the design worked even in a distributed scenario.

Q11: Are there any aspects you think could be improved in Learning Design in this particular MOOC that you think could be improved?

A11: Well the planning was decental and we had a meeting, a face-to-face meeting and a second one in order to finalise it. Because every party of us used to teach the whole subject. We had to split it into as we had intended into a beginner's course in order to then come up with eight, let's say, additional courses that then would be designed by each of the responsible persons alone. So, I found that an interesting approach in order to gather a lot of new materials because in our area there's not yet an acknowledged curriculum, how to teach on digital accessibility. It is not existing, there's no professional organisation who has said this is what you have to cover. So that was very, very interesting to see what the other colleagues think and what their experiences showed them is necessary to be taught. I learnt from my Irish colleague really some interesting concepts and, yes maybe also from a few others how to present things. So, in terms of the Learning Design itself I think it's simply, yes make plans and verbalise your plans and find time to discuss and what I am looking forward to is then reflection on this. Can this really, this whole outline of the whole project that has not always basically led to FutureLearn courses and MOOCs but two separate MOOCs basically, be then brought into a more, let's say, planned for curriculum if other people want to teach on digital accessibility.

Q12: What features on the platform itself, the FutureLearn platform do you think assist with monitoring learners' critical thinking?

A12: Not many, I must confess. The comments, the forum, this is very useful and it works because there are enough people. If I think of a class maybe with three hundred students which I have normally and I ask them to use the forum then all I get is, oh you have forgotten this and what is the answer to this assignment? And can anybody help me? But never any content that is being discussed if I use a normal learning platform. It seems like they are all afraid of the professor who is also reading this possibly. While in the MOOC it looks like grown up people, everybody is the same, this happens much, in a much more democratic way I would say. Not saying that I would delete any message which is maybe unfavourable for me also, no I can't do that even in my learning platform. But somehow the fact that people comment and that you see the number means okay doesn't matter, that's not a big deal if I add my advice on top of that.

Q13: I say what's the most useful feature, and I think you've just answered that question which is the comments.

A13: In terms of content it is the videos which have to be designed and produced in a good way so it was worth to get some, let's say training, a little training when we were in Southampton. That helped a lot to make some useful videos while yes, I
would say the, their simple numbering, their counting of the comments and then the ability to comment, this is what people give the voice and which allows them to think about it, what can I add on top of this?

Q14: I mean as far as just as an additional question here, as far as the feedback that you get is there anything else that you would like to see that you’re not getting from this platform?

A14: As I’ve said if I want to integrate my own students I would like to see how they are doing. Because then there’s not double work for them to report on what they did in FutureLearn. While if I use my Moodle, I have immediately all the comments, all the uploads linked to the students’ account and therefore the quiz is all linked to the student’s account and all the reporting is based on this. But it puts more stress on the students so yes.

Q15: Are there any comments that you would like to make?

A15: What I felt was very, very good is the fact that FutureLearn looked into accessibility and also fulfil the promise as far as I can understand. So, I congratulate them, they are the only one that I am aware of. We looked for MOOCs that are accessible and really FutureLearn did a good job.

AUDIO ENDS
Appendix E

Digital Accessibility MOOC – Week 1 Analysis
6 - 13 February 2017

In a nutshell
I have developed a machine learning algorithm developed from human ratings of MOOC comments using methods derived from established pedagogical theory. The aim is to use these human ratings, along with linguistic analysis to automatically rate comments in terms of evidence of comment authors’ critical thinking.

In tests the algorithm proved to have an acceptable level of accuracy (Intraclass Correlation Coefficient = 0.768), which means that more often than not it generated the same rating as a human rater.

What data was analysed?
1882 comments related to Week 1 steps, made by learners between 6-13 February.

How has the data been cleaned?
Educators and mentors’ comments have been identified and removed.
Author ID’s have been removed.
Comments containing 49 or fewer words have been removed, leaving a total of 814 longer comments (≥50 words).

What is this analysis based on?
The analysis is based on two forms of analysis: Pedagogical Content Analysis and Linguistic Analysis.

Pedagogical Content Analysis:
In my earlier study, I employed 7 human raters working independently to evaluate 1500 MOOC comments using two established pedagogical content analysis methods: Community of Inquiry: Cognitive Presence (CoI) and Bloom’s Taxonomy of the Cognitive Domain. Very close associations were identified between both rating methods, suggesting that they were measuring very similar cognitive activity – namely critical thinking.

In normal use CoI is used by human raters to classify four stages through which a discussion develops – from ‘Triggering’ through to “Resolution”, and usually includes a 0 rating for ‘off topic’ comments. In the earlier research project, the close association between Bloom’s 6 levels of critical thinking with the 4 levels of CoI strongly suggested that CoI could successfully be used a general gauge of critical thinking rather than a specific measure of how a discussion evolves. So, while the automatic rating given to each comment is on a scale of 0 to 4, it does not claim to measure Cognitive Presence as described by the Community of Inquiry framework, but rather evidence of increasing levels of critical thinking. In this way CoI ratings act as a proxy for levels of critical thinking.
LIWC analysis:
Among computational approaches to language analysis, Linguistic Inquiry and Word Count (LIWC) was chosen as suitable for analysis of online discussion, and evaluation of cognitive processes. The application adopts a quantitative, word count approach that aims to reveal the psychological meaning of words taken out of context from their original settings. It searches within text files for over 2300 words or word stems, tracking stylistic aspects of language use classified into 82 dimensions (e.g. articles, prepositions, pronouns), psychological processes (e.g. positive and negative emotion categories, cognitive processes), and other categories.

In my earlier study, certain word types and linguistic characteristics were found to have close associations with the pedagogical content methods suggesting their usefulness in predicting levels of critical thinking. During the algorithm training process 13 LIWC attributes were identified as acceptably accurate identifiers: word count, words per sentence, words containing 6 or more letters, personal pronouns, auxiliary verbs, conjunctions, positive emotion words, negative emotion words, cognitive process words, causal words, differentiation words, and words associated the thematic apperception test categories: affiliation and power.

LIWC does not provide usable results on comments containing less than 50 words. Therefore, these brief messages have been removed from the comment dataset.

For more information on my earlier study please see the accompanying unpublished paper.

What is provided?
In addition to the transcript of our first interview, and the unpublished paper, you have been provided a CSV file containing individual comments, their associated with LIWC analysis scores, and Pedagogical Scores (PS) automatically derived from Machine Learning analysis. You have also been provided with questions for you to consider for our next, follow-up interview. It is recommended that you read this document first before reviewing the others.

Below you will find graphs showing PS scores (Automatic Ratings) for comments containing 50 or more words for each step during the first week of the MOOC, along with a selection of comments associated with each step, and their automatic ratings. To see all the comments please refer to the CSV file.

Thank you.

Tim O’Riordan
15 March 2017
Appendix F

Transcript of second interview with Participant A

12 April 2017

Q1: Right, so, thanks for your help, for taking part and this follow up and final interview for this case study. I sent you the transcript for our first interview, could you let me know, are there any problems with that at all, any omissions, errors or anything you would like to add?

A1: No, so far it looks quite good I think, the transcriber has done a good job.

Q2: What are your thoughts on the comments and analysis that I provided?

A2: Yes, at first I was a bit confused because the approach that you have to provide scores is very brief and doesn’t really explain where basically, the decision is between a three and a four or a one and a two. So, it took me a little while to read through the actual comments, try to reflect on the scores and then it became a little bit clearer, but now and then I found I would have decided it in a different way, so maybe later we can go through them.

Q3: What did you find confusing or unclear, and it was that, the, the kind of, what, what does each rating mean and how do they...

A3: The meaning and actually whether as you wrote, the length of the text is so important or not and in the beginning I was quite sceptical about your decision to cut away everything that is below fifty words, because I couldn’t see them now, I must say, I did look it up, this decision is... Because some people in my mind, simply demonstrate, let’s say, following and way of thinking about the results by providing a few statements that can be quite revealing. I think there will be later a few examples, but those are above fifty words, so they are not really those sort of ones. I agree, many people who write very short phrases are probably not following a lot, but then if they say I, I cannot add more to the previous comments, I understand that they have read this and therefore I am not, I think they should be at least a one in my mind. They have followed and therefore they are in the course.

Q4: Going on to 2b, the accuracy of the analysis. I think you were looking at Week 3 and, how, how did your evaluation of that week, of how things were going with the learners’ engagement, their critical thinking during that week. How closely did the analysis that I provided accord with what your feelings of them are?

A4: So I have now made a little list of things I would have changed. Would you like to go with me through this list? Yes? So, let’s start in Week 3, Step 3.1. There is someone, yeah, that, overall, this relates to c) already but that we asked about peoples own experience with this, since this is a course about accessibility, this implies they have reflected already on their role as a possibly disabled person, which is a bit too demanding. So, most people simply said I’m commenting or I have some experience, but of course they were not asked to reflect on disabled persons need for some of their [unclear]. So those people are now flagged for this who have a disability and therefore have commented a lot because they could write a lot. So, in my mind, that is probably, some comments were shorter length, which should be actually analysed, but because we did not start the questions in this way, I think we should have changed this.
Appendix F

Q5: Yes, OK
A5: Yes. OK. In section, otherwise the rating is ok.

Q6: OK.
A6: 3.2. Its about Mary and rating one is an example, interesting comments and I think its pretty short, yes, but the examples are very brief. They are very correct so I would have raised this actually and see this as a two or even a three.

Then 3.3 there is a rating three which is already pretty long and so there must be a reason that the algorithm gave it a three. I found that the example is very well chosen and I am very glad that that person commented because it is something other people can benefit a lot because it is a first time report my colleague developed, and he is explaining what that colleague. It was one of those examples where I through, oh, we are so lucky to find these people in this course, so in my mind it is actually a four. Then their colour scheme problem which, now I am confusing this. No, lets simply skip and go onto the next one which is. Yeah, 3., yes, my notes were too brief.

3.4 is about other colour schemes and colours and here I was a bit surprised that rating one is an example where the user clearly uses the exact term “colour schemes” and therefore not only shows that he knows what is needed but that others can benefit from this so I think this is too low, in a sense, it is the right term that he has been using and he gives an example where he would say “other people can benefit from this”.

OK, so the zero, in this particular example is accurate. Somebody misunderstood and the four is not referring to colours.

Q7: OK.
A7: Which is surprising that the lower end has been detected so well.

Q8: I am glad that you spotted one that was right, so that’s good.
A8: No, no problem. Now, 3.5, 3.6 and 3.7 are all OK. So, I went through them and I think they are adequate. No problem, because I was also thinking how this should be reflected in the actual use and how to help the moderators and the other participants if they see a rating, because if, basically, you have, let’s say an input by a participant and a number of comments on this, the comments properly cannot be on the same level or rating, or below, it can be interesting to see a relationship or not because if people comment on some of the input of participants, then what does it mean in terms of their thinking. In my mind, it is an extra, but the scoring it doesn’t fit. So, it’s a bit difficult to reflect this. You could say sort it by, or filter, or things like that, they are really a large number of comments, but its probably important onto use this rating in order to improve the learning and also the participation.

In 3.8, there is a four on a topic for speech input. And then somebody has written a very long text, but actually its explaining his experiences from the past, not really applicable a lot to the current situation, so people would hear a lot about things that are not of interest any more. How the description is that I know it, yes, I have it also the same experience, but its not applicable to us anymore. They don’t use this approach, its been completely given up. So, it is not really a four in my mind. The zero in this case is pretty brief. But very short sentences say a lot of examples. How to use speech input. So, he describes this also by observation of his son, or her son, I don’t see this, but I think this should be a three. Its really, participant list a good of examples, to do with speech.
3.9, settings. Alright, creating settings, yeah. Three in my mind is better because its introducing a new type of disability. I would have given it a four because it is also referring to somebody’s experiences. “My elderly mother-in-law”. The two also I think is better, I would have three. Accidental [12.10]. We didn’t say about this a lot but it is actually what is needed so its accepted they are thinking they want to see. What can I do in order to help these people? Its not a four because this person doesn’t know what to do and didn’t look it up or found out. Certainly, it is not what we have covered in the course.

3.10, the three in my mind is only a two. Like it’s a problem, its really generic and not, doesn’t come to the point that we have to set up the features and actually learn about them. The zero here in that case is also much better because at first I was surprised. It’s a long piece of text and still it was given a zero and the descriptions are adequate really. It describes what is needed to and what you need to change. So, I would have given it a two or even a three.

In 3.11, there is a three. This maybe a bit off topic, but here the whole thing is about colour detection and then the person describes about flashing. Flashing has nothing to do with colours. The person doesn’t know that flashing is something that will be, or is discussed in accessibility in great extent because it affects people with epilepsy, so no problem. Certainly, flashing has nothing to do with colours, so it is off topic and therefore at least a one or maybe even a zero.

3.12, there is a two. Yeah, maybe a little bit too low. It’s about the colour vision demo and it’s about a washing machine. Its exactly what is needed, so if you take out your socks from the washing machine, want to dry it or put it on, how do you say it in English? I don’t know. You put your clothing on the line and you maybe want to do it in different way to make it easier when you collect it afterwards, so it’s a perfect example in my mind, so at least a three, or maybe even a four.

3.13, its about multi-media use of mobile phones. Yeah, I would have switched to a three or a four. Because three at the moment is not really coming to the point. The person refers to subtitles when we have already introduced in the weeks before that you would talk about captions. Maybe this is an English speaker or a non-native speaker, it confuses it, but it should have learnt that by then, we refer to captions when we mean what we prepare for, hearing impaired people, but then he talks about e-books and print so this certainly is not multi-media. So, a bit less ideal. Instead, the other person, three, refers to media content perfect and [16.01], is also perfect. Interesting for readers that there is a service that relies on video.

Then 3.14 here also I would say that those that have been rated low, the example three and example four. Rating two, ideal. I didn’t see this when I was going through the comments that this guy wanted to have more information about measuring the correctness of mental maps. I would have liked to refer to some applications. Actually, its what I would like to see from a student, yes? Give me more information. So, a two is certainly not adequate here.

3.15 is on walkers guide. The developer explains how he has created this so the person who wrote, who was given a four, criticises the accuracy of maps, but in terms of OK, British ordnance survey is the best. Yes, but this is not bringing to the point. That’s really about people, more information and benefit from it, information and maps. So, in my mind it is a three. While the person who was rated three simply gave lots of examples what they are
needed. In my mind, yes, a three is not too bad, but a four, I think is better because it's very informative what the guy wrote.

Then in 3.17, yeah, I would also have switched a little bit, the ratings so three is actually a four. It's about developers and how can we help developers or improve developers' work, so the person who was rated a three said something about accessibility auditing and quality insurance processes. These are the precise terms and I think everybody who were at this learned a lot from this comment. So, I think it is actually a four. Whilst the comment rated two, is very short, but personas is what we have been using throughout the course and the suggestion here is to use a standardised set of personas, so very precise comment. Very well placed in my mind, so again, other learners would benefit from this and in class this would really be a good place to start a good discussion. Personas, how complete can this be and things like that.

Finally, in 3.19. 3.18 is OK. No problem. 3.19, the reflection actually, I would change the scoring from two, which was. The example of it was rated two, to a three because again, this is someone who said I cannot simply repeat. This doesn't make sense, but actually, he repeats many of the features that we have covered, and therefore demonstrates that he has followed this and reflected on it in my mind. So, I would have raised this score a bit.

So, that was it.

Q9: OK.

A9: I hope this helps a bit.

Q10: It does, it does. I mean, as I say, thank you for. I mean, that's, that's a lot of detail there and I think, its kind of raised a question in my mind which of course is that thing that we are of the analysis appears to be identifying learners who have made a reflection, in some detail, on the subject, but have still got it wrong. They still don't even understand even though they may have reflected on it. Which of course, you know, it can happen and I just wondered, would you find that, you know, did you find it useful to see that people had reflected and still, and you know, had responded incorrectly, if you like, to the, to your questions?

A10: Very good question, because, well. Some feedback is better than no feedback. Definitely, and on the other hand, there is always the suspicion that in such a course, the variety of learners makes it very difficult to use as a form to reach a lot of readers, so if you are off topic, you better be quiet in my mind because it's a certain that you can consume in terms of comments and why are you wasting other people's reading time, by just saying, OK. I know and I am sorry. No, I think this should be avoided and I would have preferred, if I could say a zero should be maybe flagged out as something you don't want to read properly. And if its accurate.

Q11: OK, that's given me a bit of a thought there. Maybe there should be some uprating or downrating on, on my rating. You know what I mean. They are not my ratings, they are based on my team, analysed other comments, but. So, yes, that's interesting. OK. So, you've mentioned one, I think one comment where you, you missed someone's, someone's comment, so would you say that it revealed other aspects of the discussion forum that you were overlooked during your period of moderation?
Appendix F

A11: Yes, when the person wanted to have more information, clearly this was an academic who wanted to see more research in that area, so obviously for an introductory course, we have not dealt a lot with researches used, but just enough to see if there are some interest, then maybe we can bring them to a follow up course. It turned out in the end that we didn't produce follow up course on this specific week, but on other topics. So yes, I think for that person it is still OK, he knows a name, me, and can simply use [23.32] and look up, look up future publications. That's what a researcher would do anyway.

Q12: It became part of his personal learning network, yes. OK. So, what aspects of the analysis did you find useful?

A12: The zeros and the fours are quite, quite useful. Not always, not always accurate, but otherwise I was surprised to see that these examples are OK and therefore, with a certain reservation, be careful, this is automatically. I would see this reflected in the interface. I think I can say that, yes.

Q13: OK. And anything that you didn’t find useful at all?

A13: Yes, in the middle. The three, the two, the one are difficult. So maybe it is better to bring them together to one category.

Q14: Yes, that's something I have been thinking about or maybe just getting rid of numbers all together. Colour coding or something.

A14: Yes, colour coding but yeah, green and blue, it's a bit...

Q15: It can cause all sorts of other problems. Yeah. Yes. Anyway, I will have to think about that, maybe, yes, anyway. So, OK, would you, would you think that the analysis could be useful to other users? Like learners or administrators?

A15: I try to explain that some of the comments are reflections to previous comments and here the system doesn't allow to have a multi-level approach. So, some of the, some participants are reflecting to the reflections. Others are reflecting to the original ones and you can't find out what happens. So, you have to read all of this. I think your algorithm also cannot tell this. So, the answer would be no, in some cases. But if there is no such reflection on a contribution, then I think its useful if you reflect out the positive ones. So, the fours, I would like to see out. Good, readable, try to make use of it. The bad ones, I would be careful because the zeros can be good ones.

Q16: Yeah. OK, and do you think, I mean, you kind of touched on it a bit, maybe this analysis could be harmful to users?

A16: Harmful? Harmful? My English. What is harmful? I know what it means, yes, sort of, but normally I use harmful if you are really hurting somebody like. OK, if you flag a comment as a zero, then that person thinks twice before it would comment again. So, from that and again, the zeros are not to be flagged in my mind. From that side, for other uses. Not for the educators. So, if you say users maybe question is addressing both.

Q17: Yes.

A17: Both. So, if you said both, then yes, again, the answer should be yes, although I mean only the learners.

Q18: So, would you, I mean you've touched on this a bit earlier, I mean, how would you think this analysis could be improved?
A18: Yeah, technically, it's a bit difficult to give good advice here. The scoring at the moment has five levels and I was surprised that some of the, there are no examples of one in the zero and then I would have then started to look into shorter comments if there is no example for a one and a zero. Because for our discussion now, it would be important to have more comments on, or examples. But this is more meti-analysis than the answers, so if you couldn’t find some examples that are rated as zero and one, then maybe this is not the best approach. So, yes, I cannot give you a good example of how to improve it because I didn’t dive into the actual training. For example, I don’t know if you are trying the vocabulary per course material, per step, or how you come to actually, yeah, classification approach. I didn’t look into this that much. I wanted to spend more time with it, but I couldn’t find it.

Q19: Did you find anything particularly surprising or unexpected in the analysis?

A19: Yes, what was surprising for me was in the beginning, you wanted me to, to go basically through the comments and try to understand the scoring. That was a bit of a surprise. I didn’t know in the beginning that this would be happening in the follow up interview.

Q20: OK, right. Yes, sorry about that. Sorry, I didn’t explain it. So, just thinking about the design of the, I mean, what are your thoughts of the effects of the analysis on your reflection of the actual design of how the mood went?

A20: The motivation at the beginning was built on the assumption that users use their own mobile phone and that worked, but the very first discussion did not address this. We started too early in my mind by asking people about their apps, and this is not needed because they are on the course on accessibility. They are eager to find out about accessibility issues, so this was a bit in my mind for that particular week. In terms of the quiz, and you couldn’t analyse this clearly because this is not what your algorithm does.

Q21: So, no comments.

A21: Yeah, the quiz that we have provided for the participant. This was really quite useful in my mind to assimilate the discussion afterwards and the reflection was benefitting from this. So, from that side, I think we found a good approach to bring people to a point where they in the end they say they have learnt a lot and have them demonstrate this also.

Q22: OK, so I suppose the thing, you know, that you can see from step to step that there’s, you know, there are a high number of what you might say a higher rated comments in each step and I just wondered if, if you, if it was related, if you were expecting those, that response to each step as you went through, or was there anything that you thought well actually, I would have expected more you know, in depth responses to this step and it didn’t, it didn’t actually occur. Did you see anything like that at all or?

A22: The input was not well reflected and basically its our fault because we didn’t teach a lot about it, so the participants didn’t have a great chance to understand what is going on really when blind people rely on braille. Just explain, brail is an umbrella term that sends various notations within a certain language, so there is Grade 1 and Grade 2 in the UK and in Germany we have, but on top of this you have zero, we have mathematical brail, you have musical brail and we haven’t covered this at all, so if the reader now, if a participant now goes in and says my phone can handle brail, this not the right message. That we wanted to
give because certainly they can’t enter mathematical brail, nor contractor brail, so yeah. That was a bit of a problem I noticed afterwards.

Q23: OK, so is that something you would change as a result of this?
A23: Yes, yes.

Q24: The comment analysis or from your own, is this coming from your own reflection?
A24: From the comments analysis, yes.

Q25: Would you like to see this kind of analysis provided as standard feedback?
A25: Yes, I can say, yes.

Q26: And visualisation, do you think that this could be improved?
A26: Oh yes, definitely. I would use the standard star system, colourise the stars because people are more used to this, that you have a kind of recommendation system. So, they can recognise this more easily and don’t have to think is a four good, or bad?

Q27: Do you have any other comments at all?
A27 No, thank you so far for being able to participate and sorry for my delay but it was a bit of a hectic time in the last week so.

Q28: Yeah, like I said, I haven’t been well either so that has impacted on my chasing people up so anyway. I am going to stop recording now. So, thank you.

AUDIO ENDS
Appendix G

Automated ratings for example comments supplied to Participant A, compared with Participant A’s own ratings provided in the second interview

<table>
<thead>
<tr>
<th>Step number</th>
<th>Auto rating</th>
<th>A2 rating</th>
<th>Step number</th>
<th>Auto rating</th>
<th>A2 rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>4</td>
<td>4</td>
<td>3.10</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3.2</td>
<td>4</td>
<td>4</td>
<td>3.11</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3.3</td>
<td>4</td>
<td>4</td>
<td>3.12</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3.4</td>
<td>4</td>
<td>4</td>
<td>3.13</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>3.14</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.6</td>
<td>4</td>
<td>4</td>
<td>3.15</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3.7</td>
<td>4</td>
<td>3</td>
<td>3.16</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.8</td>
<td>4</td>
<td>4</td>
<td>3.17</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td></td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>3.9</td>
<td>4</td>
<td>4</td>
<td>3.18</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
References


References


References


References


References


References


References


