Learning in the Wild: Coding Reddit for Learning and Practice

Kumar, Priya; Gruzd, Anatoliy ; Haythornthwaite, Caroline; Gilbert, Sarah; Esteve Del Valle, Marc; Paulin, Drew

Published in:
Proceedings of the 51st Hawaii International Conference on System Sciences

Learning in the Wild: Coding Reddit for Learning and Practice

Priya Kumar
Social Media Lab,
Ryerson University
p1kumar@ryerson.ca

Anatoliy Gruzd
Ted Rogers School of Management,
Ryerson University
gruzd@ryerson.ca

Caroline Haythornthwaite
School of Information Studies,
Syracuse University
chaythor@syr.edu

Sarah Gilbert
The iSchool,
University of British Columbia
s.gilbert@ubc.ca

Marc Esteve del Valle
Department of Media Studies and
Journalism, University of Groningen
M.Esteve.Del.Valle@rug.nl

Drew Paulin
School of Information,
University of California, Berkeley
drew.paulin@ischool.berkeley.edu

Abstract

This paper introduces a 'learning in the wild' coding schema, an approach developed to support learning analytics researchers interested in understanding the different types of discourse, exploratory talk, and conversational dialogue happening on social media. The research examines how learner-participants ('Redditors') are leveraging subreddit communities to facilitate self-directed informal learning practices on the social networking site. The coding schema is tested and applied across four 'Ask' subreddit communities ('AskHistorians', 'Ask_Politics', 'askscience', 'AskAcademia'). The research brings attention to how knowledge, ideas, and resources are being shared and supported outside the confines of traditional education and professional environments.

1. Introduction

There are many ways to use the Internet for learning. There are resources such as Wikipedia pages, YouTube videos, online news, electronic books, and open access journals. There are interactive learning opportunities, such as open courses and online degrees. And then there are the wilds of open online discussions on sites such as Digg, Snapzu, Stacksity, Voat, and Reddit [38]. These social media sites offer arenas for discussion that are contributed to, led, and moderated by members of the site. Discussions can be for play, social interaction, and curiosity; but they are also for learning. This is learning that is not occurring through instructor-led courses; it is not based on a pre-defined syllabus; and there is no mandate to cover essential texts and ideas. Yet, they are sites for learning where questions are asked and answers provided, where crowds of participants comment, correct, and argue about answers, and where those who answer make the effort to present information in informed, accessible ways, with citation to sources and further resources. No one manages this learning; no one earns a university degree or a workplace promotion from this kind of teaching or learning (at least not directly). It is thus that we call this 'learning in the wild' (with due acknowledgement of Hutchins' *Cognition in the Wild*). It is informal and non-formal learning that is taking place outside traditional educational environments, with what is asked about, answered, and learned at the discretion and direction of those who ask and answer. It is crowdsourced learning, but not of curricula or courses, but in conversation-sized pieces, based on crowdsourcing interest in answering just-in-time questions.

This paper reports on the development of a coding schema for content analysis of informal learning on social media derived by examining the kinds of learning happening on Reddit, and results on the kinds and distribution of learning practices found in four 'Ask'[1] subreddit communities. The research contributes to our understanding of online conversations in support of learning, and to content analysis for online social learning practices.

1.1. Trends in open, online learning

Our research group has been working for a number of years studying the practices of learning online, primarily in open online classes, and observing and researching the trends toward more learner-centered participation. Among

[1] 'Ask' subreddits use a Q&A style format where questions are posed and answered by Reddit users. They vary by scope, audience, and topic.

URI: http://hdl.handle.net/10125/50131
ISBN: 978-0-9981331-1-9

the trends is the way open, online participatory practices merge with learning practices in online settings. An early trend in online education, that appeared even before the more widespread recognition of participatory media, is the adoption of collaborative learning. The wide adoption of collaborative learning as a pedagogical choice emerged in part as a response to the demands of a 24/7 classroom of simultaneous, asynchronous, online discussion, and but also in recognition that social learning is a practice that sustains adult learners beyond course contexts and translates to practices that are found in adult life and work communities [21, 33].

As open online initiatives drive collaborative, participatory, crowdsourced forms of learning, they also depend on greater learner autonomy and responsibility. Today's learners grapple with self-directed learning [17], making sense of their own experience through connecting and creating their own learning ecologies [31]. Open online courses, including MOOCs with 'massive' enrolments, attract such learners, people who are not necessarily aiming for a certificate at the end of their experience but who are taking the opportunity and initiative to be self-directed learners [42]. These learners maintain their own responsibility for reading source material, engaging with fellow learners, and completing learning assignments. At times, they help the whole learning process by acting as explainers for others, synthesizers of material, citation providers [21], and active evaluators of others' work [36]. Personal information management and personal learning management become prominent for these learners, as they pick, choose, and consolidate the use of particular social media and forums for their learning practices and their learning portfolios (e.g., essays and reflections posted as blogs). The learning perspective of connectivism, and connected knowledge [40] comes to the fore when individuals make sense of their information environments by connecting resources, and actively constructing "learner generated contexts" that support individual and group knowledge [31].

The growth in online learning in educational settings, and the prospect of massive enrolments in single courses such as MOOCs drives a third trend. This is the trend toward more automated measurement and evaluation of online conversations as learning analytics develops as a field. Learning analytics is most commonly understood as "the measurement, collection, analysis and reporting of data about learners and their context, for purposes of understanding and optimizing learning and the environments in which it occurs" [40]. Within this field, there is a growing area of research on social learning analytics. Research approaches such as conversation analysis, natural language processing and social network analysis are brought into play to gain an understanding of online social learning processes [4, 16]. Work in social learning analytics is in its formative stages, and is challenged by the multiple ways to approach open, online learning. Considerations can include intra-group relations, technology choices and affordances, virtual versus face-to-face interaction, and online conversational practice, as well as how people learn.

Social learning as originally conceived holds that learning occurs through observation of behaviors, including others' reactions to those behaviors; the learner (e.g., a child), chooses to imitate or not the behavior according to the reactions observed. For adults, apprenticeships provide a framework for this kind of learning by observing and doing [29], with master craftsman modelling appropriate practice. In open, online environments similar learning processes are going on as individuals lurk and observe before posting, as they observe inappropriate behavior sanctioned.

In formal education, it may be assumed that it is the teacher's practice that is being observed and imitated. But formal learning also entails formal structures – the teacher as the only voice to be heard; the physical room structures that put teachers at the front; the right answers for examinations and the right approaches for assignments. These constrain who talks to whom, and thus who and what is observable to model. Such is not the case for open, online learning. Even in educational settings, the norms of turn-taking conversation are transformed by asynchronicity and the reduced cues of online, computer-mediated environments. Social learning analytics expands the view of social learning to include consideration of social networks that reveal how learning opportunities occur or do not occur according to the structure of the networks of people, ideas and resources in which individuals are embedded [20, 23]. Work in computational social learning analytics and social network analysis is beginning to be used to help visualize the otherwise invisible teacher-student and student-student online interactions in distance learning programs [16, 20].

As well as changing who interacts with whom, new media has altered the way we communicate and interact. Technologies provide different kinds of features – asynchronicity, anonymity, text and pictures – that affect what can be communicated and how. In open online forums, distinctions between teachers (producers) and learners (consumers) are blurring [13]. New literacies are emerging that respond to the changed nature of conversational practice in online settings and online learning [22].

Buckingham Shum and Ferguson [4] also expand the

ideas of social learning by deliberately addressing supposedly off-topic conversations and bringing attention to the full set of interactions that impinge on learning behaviors: "the focus of social learning analytics is on processes in which learners are not solitary, and are not necessarily doing work to be marked, but are engaged in social activity, either interacting directly with others (for example, messaging, friending or following), or using platforms in which their activity traces will be experienced by others (for example, publishing, searching, tagging or rating)" [4, p. 5].

But, it is not just subject learning that happens in online settings. Conversations about topics and practices as exhibited by online postings and reactions contribute to both individual learning and group practice. Thus, we add to social learning the need to learn about the social – the rules, norms and practices of the local environment. Learners entering online conversations join or create new communities of practice, where rules and norms are defined and reinforced. Research on both virtual communities and group behavior show that the task of learning how to be a member of such a community or group can be a major hurdle to participation [10, 21, 29]. The need for such learning is evident even in the terms used for new users – newbies, apprentices, lurkers, legitimate peripheral participants – and for more advanced users – experts, wizards, gurus. Sanctioning those who do not follow the rules is common in online forums, keeping participants in line about appropriate language, topic, expertise, and genre of posting, and allowing newbies to observe the consequences of not following the rules.

General trends in education, career growth, and the pace of change in knowledge all point to the need for learning that is both lifelong and lifewide [25]. Learning has always taken place outside educational institutions, but the development of open, online forums provides the opportunity to study this kind of learning 'in the wild'. Thus, mindful of the growing importance of open, online learning for career and personal needs, and the range and types of learning occurring in online learning communities and groups, we set out to explore how learning unfolds in open, online environments, operating outside educational institutions.

The setting we chose to start with is Reddit, which we explore with in-depth analysis of conversational learning practice in four subreddits. The major contribution of the work so far is our coding schema. This entailed a multi-stage process of development that addressed both the kinds of considerations we were aware deserved attention in online learning conversations, and the need for a parsimonious schema that could be applied first by independent human coders and later for automated text analysis. This paper presents our coding schema and how it was developed. This includes the iterative process of code refinement, where members of our team piloted and pre-tested the schema across four Reddit subreddits: 'AskAcademia', 'Ask_Politics', 'askscience', and 'AskHistorians'. We then present evaluation results, in which our final coding schema was applied to a larger sample of comments from 'AskHistorians' by three independent coders.

Overall, our aim is to contribute an empirically rigorous understanding of the way exploratory dialogue, behaviors and talk unfold in tandem with learning processes, with the aim of understanding the nature of learning practices in open, online social networking sites. By detailing our process of refinement and validation, we also invite other scholars to apply our coding schema to their research across other social media online learning environments (e.g. Twitter, Facebook, LinkedIn).

1.2. Coding "learning"

Previous research on coding learning has focused on addressing formal settings (e.g. conferences, educational courses, teams) or more open online interaction. Techniques and computational tools have been applied to a single case or to specific online phenomenon, with the aim of understanding learning processes and improving practices. For example, studies have used quantitative predictive modelling to show how knowledge is constructed, disseminated and validated in open online settings [8, 26]; and automated dialogue assessment tools to improve participatory collaboration in virtual classrooms, academic communities and communities of practice [35, 41].

While keeping previous work in mind, in developing our coding schema, we followed on Buckingham Shum and Ferguson in their work of identifying elements of *exploratory dialogue* in a manner suitable for machine learning [11]. Exploratory dialogue is one of three kinds of talk identified by Mercer in a study of classroom talk:

> "Exploratory talk, in which partners engage critically but constructively with each other's ideas. Statements and suggestions are offered for joint consideration. These may be challenged and counter-challenged, but challenges are justified and alternative hypotheses are offered. Partners all actively participate and opinions are sought and considered before decisions are jointly made. Compared with the other two types, in Exploratory talk knowledge is made more publicly accountable and reasoning is more visible in the talk" [32, p.146].

Like Ferguson and her colleagues, we build on Mercer's exploratory talk because it represents the kind of

constructive, collaborative interaction that reflects adult, collaborative learning and is likely to advance both individual and group knowledge. We expect this kind of talk to support informal learning because online textual discussions involve active processes of co-reasoning, constant negotiation, and knowledge, idea or resource sharing [11]. In terms of individual learning, we make the assumption that if we find exploratory talk, we expect learning to have occurred. However, we stress here that our aim is to understand online processes in the service of learning and we are not addressing individual learning outcomes.

While our focus is on exploratory talk, the other two forms of talk identified by Mercer may also have relevance: "Disputational talk, which is characterised by disagreement and individualised decision making"; and "Cumulative talk, in which speakers build positively but uncritically on what the others have said" [32]. Disputational talk may affect the way learning proceeds, shutting down interaction, and excluding participation. Cumulative talk may serve to reinforce an idea, or it may signal social agreement. Thus, while we focused on exploratory talk, in developing the coding schema, we kept in mind these other forms of talk.

Our aim is to develop a general coding schema that will hold across different informal learning settings. However, at first instance, we defined and refined our coding by working with several Reddit communities (subreddits) particularly oriented to asking and learning about different spheres of knowledge. The next section describes the Reddit setting and the subreddits we worked with.

2. Reddit

Reddit is an online news sharing site that is commonly referred to as 'the front page of the Internet' for the way it presents headlines and how crowd-based online voting raises the profile of news or other items to a front page equivalent. By its own account, "Reddit bridges communities and individuals with ideas, latest digital trends, and breaking news" [37]. Reddit has become increasingly popular since its launch in 2005, and now maintains a relative stronghold as the go-to, self-organized community site for people interested in current affairs, social commentary and Internet subcultures. As of April 11, 2017, Reddit ranks 17th in terms of total global traffic, and 4th in the U.S. where over half of its total users reside [1]. Anyone with an Internet connection can become a member of Reddit (a Redditor) and, with little or no formal training, use the site to share information and resources across a plethora of niche communities known as subreddits.

A key aspect of Reddit is that contributions are anonymous, leading to potential transgressions; however, development of rules and norms, also known as reddiquette, make it possible for the platform to function [30]. Subreddit communities are moderated and content is user-generated, affording users the opportunity to comment anonymously, browse, and stay updated on a multiplicity of subjects at their discretion. Behavior modelling is shown through norms and practices that reward appropriate behaviors consistent with site-wide Reddit culture, and with distinct subreddit subcultures [2]. Redditors can upvote or downvote others' posts or comments (a score known as 'karma'), affecting the order in which posts and comments are displayed on the page: upvoted posts and comments rise to the top while downvoted posts and comments go to the bottom [12].

We felt that Reddit would be an ideal site for examining learning practices because participation engages self-motivated learners, occurs outside traditional professional settings (e.g., academic research, university lecture halls, workplaces), combines perspectives from experts and non-experts alike [34], and covers topics chosen, promoted and responded to according to the contribution and direction of members. A focus on exploratory learner dialogue fits well with Reddit because the platform maintains a user-generated participatory online culture through its informal, openly accessible, group-based subreddit communities. Moreover, there is variety in the different subreddits that can highlight different community learning norms and dialogue; not all subreddits are alike, and each community maintains its own subject expertise, thematic focus and social norms that may or may not be conducive to collaborative online learning processes. Depending on the subreddit the kind of dialogue can be transactional and functional in nature (i.e. sharing specific resources, strict Q&A, offering advice); in other cases, posts may be more conversational or argumentative, leading to ever-revolving debates between members that expand overtime and never really 'end' in a strict sense. This range of practices was kept in mind in creating the schema.

As will be shown below, following extensive development we put forth a 'learning in the wild' coding schema to understand and assess the different types of discourse, exploratory talks and overall nature of learner conversations happening on Reddit. Our team applied the final schema to four subreddit communities – 'AskAcademia', 'Ask_Politics', 'askscience' and 'AskHistorians'. Three independent coders were used to test and evaluate the utility of our final coding schema. For this validation process, we chose the 'AskHistorians'

subreddit, to see whether our schema was able to reliably capture the nuances, social cues and linguistic markers that we argue play a role in facilitating exploratory dialogue and informal learning processes online.

'Askscience' was created 8 years ago and is a default subreddit, meaning that users are automatically subscribed upon creating an account and must choose to opt out if they do not wish posts to appear on their front page. As of writing, 'askscience' has 14,191,675 subscribers. 'AskHistorians' and 'Ask_Politics' were created 5 years ago; as of writing the former has 604,531 subscribers and the latter 24,887. 'AskAcademia' has 29,026 subscribers, is 6 years old. These subreddits offer multiple avenues for comparison, both in terms of coding schema refinement and the diversity of informal learning processes, exploratory talk and group conversations that take place on Reddit.

3. Development of the Coding Schema

The process of developing the coding schema comprised three stages. In all stages, the coders were researchers in the research team, each aware of the literature in this area, the kinds of learning processes that might occur, and the aims of the research. Coders included two doctoral students, one post-doctoral fellow, and three faculty holding university positions. One member of the research team, a post-doctoral fellow, was designated as the 'primary coder' with responsibility for managing the coding process and gathering input individually and collectively from coders. In general, the research team met weekly in a team Skype meeting and coding experiences were shared. The coders applied each version of the schema to subreddit datasets, and then engaged in discussion about pros and cons of particular codes, the range of activity that should be coded, and how codes should be refined. Each stage culminated with the definition of the next stage coding schema.

3.1. Stage 1: Exploratory dialogue and intra-group behavior

In Stage 1, we adopted Ferguson et al.'s cue phrases framework that includes the following seven categories (Table 1): 1. Critique; 2. Discussion of Resources; 3. Evaluations; 4. Explanations; 5. Explicit Reasoning; 6. Justifications; 7. Others' Perspectives [11]. Ferguson et al's cue phrases were developed and piloted in 2011 in a series of studies that added a qualitative layer to quantitative data through self-trained (automatic) detection and analysis of exploratory and non-exploratory dialogue [10, 43]. Because of the open nature of the Reddit environment, and its greater similarity to online group behavior and virtual community practices [14, 18, 19] our schema was extended with two additional categories addressing group behavior: 8. *Learning the Rules* was added to capture the dialogue acts and content submissions that we argue are particularly unique to Reddit, e.g., following subreddit norms and guidelines that explain how to be an effective contributor or member of the community; 9. *Socializing* was added to capture the human context (e.g. the expressions of gratitude, approval, confrontation or opposition) of Reddit conversations.

Table 1. Reddit codebook version 1

Code	Definition	Linguistic Dialogue Example
1. Critique	The comment suggests disagreement; something may be wrong, faulty or in need of correction/ revision/ reassessment.	'However', 'not sure', 'maybe', 'hmm not really', 'think it through', 'actually, not exactly'
2. Discussion of Resources	The comment references and provides details of additional outside resources (e.g: links to external websites, forums, books, articles) to support understanding or extend discussion.	'Have you read', 'more links', 'check this out', 'look at', 'read this'…BOTH online and offline resources
3. Evaluations	The comment appraises and assesses the merit, worth and/or significance of something.	'Likely', 'good point/example', 'could be', 'fair enough'
4. Explanations	The comment has a descriptive quality and undertakes a process of 'thinking it through' by explaining, brainstorming and justifying a position or idea.	'Means that', 'our goals', 'the aim is', 'meaning', 'it depends, for example'
5. Explicit Reasoning	The comment works out ideas in a logical manner, often reaching a conclusion or proving a point through example based inferences. This includes taking the same line of argument further through questions/objections.	'Next steps', 'relates to', 'that's why', 'then you would', conditional 'if X then Y', 'along these lines'
6. Justifications	The comment reasons/expresses/offers judgment in terms of something already known or found.	'I mean', 'we learned', 'we observed', 'based on'
7. Others' Perspectives	The comment extends discussion by putting forward additional/alternative views and positions, increasing the range of an idea.	'Agree', 'another way to look at it', scholar/public figure argument, 'their research focuses on', 'through this lens'
8. Learning the Rules	The comment references the Reddit platform and may remind users of the protocol/code of conduct for the particular subreddit.	'See/don't forget subreddit link', 'this post doesn't belong here', up-/downvote mentions, acknowledging OP redditors
9. Socializing	The comment follows an informal, small-talk and conversational-like structure between users.	'Thank you', 'much appreciated', gratitude, positive/negative informal conversations, sarcastic one-liners and

		jokes, personal attacks/criticisms 'you know nothing', 'you are dumb'
Codes 1-7 from Ferguson et al. exploratory dialogue cue phrases (2013); Codes 8-9 added.		

In the Stage 1 coding, we used DiscoverText, a cloud-based text-analysis software program [39] that allowed assigning multiple coders to the same dataset. The first cycle of coding was undertaken on a dataset of 1% of 2015 subreddit posts (excluding parent submissions) from each of 'Ask_Politics' (n=189), 'AskAcademia' (n=197) and 'askscience' (n=163). Each sample was coded by three coders, and was then assessed through Krippendorf's alpha, a conservative benchmark index commonly used to measure the validity and intercoder reliability in content analyses [6] and is well suited for projects that involve two or more coders and multiple coding categories [9].

In the first instance, Krippendorf's alpha statistics showed a relatively low agreement among coders ('Ask_Politics' 0.16, 'AskAcademia' 0.2 and 'askscience' 0.22). In this iteration, coders had three difficulties. The first was in distinguishing between cue phrases for *Explanation* versus *Explicit Reasoning*, and for *Discussion of Resources* versus *Others' Perspectives*, particularly for dialogue that could be described as information seeking and knowledge sharing. Second, coders expressed confusion when faced with dialogue in the form of questions, whether rhetorical, conversational, or seeking further clarification. And third, coders were unable to accurately capture Socializing, and distinguish between Socializing, Critique (negative commentary or disagreement) and Evaluation (positive commentary or agreement).

3.2. Stage 2: Reducing and refining codes

To try to resolve the inconsistencies and improve intercoder reliability statistics, Version 2 of the schema sought to capture more precisely the socializing, and resource and information elements of informal online learning (Table 2). We also removed *Justification*, and *Others' Perspectives* used in Version 1, because coders used both codes sparingly during the testing phase, thus suggesting little applicability for this context.

Version 2 also included a number of refinements of codes. For the second cycle of coding, we agreed that discussions surrounding resource and information elements were indeed a key feature of many of the online text-based discussions in the subreddit samples being studied. To capture this nuance, we added 6. *Information Seeking* as a category (i.e., general inquiry, or asking for help/clarification: 'tell me more', 'how do you', 'anyone know', 'any advice on'). Observation of the kinds of learning interactions found in Reddit dialogue, particularly in relation to the little used *Socialization* code, led to the introduction of codes *Critique* (negative/disagree), *Evaluation* (positive/agree) and *Explanation* (neutral). Our intention was to code socializing along a spectrum to capture the potentially 'good' and 'bad' feelings that may occur in tandem with online learning practices.

Table 2. Reddit codebook version 2

Code	Definition	Linguistic Dialogue Example
1. Critique	The comment suggests disagreement; something may be wrong, faulty or in need of correction/revision/reassessment . Formal/informal negative conversations, personal attacks, criticisms without explanation/discussion.	'However', 'not sure', 'maybe', 'hmm not really', 'what about', 'seems to me', 'actually, not exactly', 'you know nothing', 'you're dumb'
2. Discussion of Resources	The comment references and provides explicit details of additional outside resources (e.g: links to external websites, forums, books, articles) to support understanding or extend discussion.	'Have you read', 'more links', 'check this out', 'look at', 'read this'…BOTH online and offline resources
3. Evaluations	The comment appraises and assesses the merit, worth or significance of something. Formal/informal personal view or positive affirmation/expression of gratitude.	'Likely', 'good point/example', 'agree', 'could be', 'fair enough', 'thank you', 'much appreciated'
4. Explanations	The comment has a descriptive quality and undertakes a process of 'thinking it through' by explaining, brainstorming and justifying a position or idea.	'Meaning/means that', 'our goals', 'aim is', 'it depends, for example', 'that's why', 'another way to look at it', 'through this lens', 'I'd argue', 'same logic would apply'
5. Explicit Reasoning	The comment works out ideas in a logical manner, often reaching a conclusion or proving a point through example based inferences. This includes taking the same line of argument further through questions/objections.	'Next steps', 'relates to', 'then you would', conditional 'if X then Y', 'along these lines', 'maybe/maybe it's because'
6. Information Seeking	The comment asks a specific question, seeks clarification, posts a general inquiry, asks for help on a topic, issue or idea.	'Tell me more about', 'how do you', 'anyone know', 'any advice on how to'
7. Referencing Reddit	The comment references and cites the Reddit platform and may remind users of the protocol/code of conduct for the particular subreddit.	'See/don't forget subreddit link', 'this post doesn't belong here', up-/downvote mentions, acknowledging OP redditors

Since our research goal was to identify general patterns of learning, we examined multiple Reddit

communities in developing our coding schema. In our attempt to create a 'mutually exclusive' coding schema, we discovered that many single Reddit comments exhibited a number of different dialogue processes. Accordingly, we decided to allow up to three codes to be assigned per comment. Given these results, an increasing understanding of the elements of learning dialogue in the 'Ask' subreddits, and the need to arrive at a repeatable coding scheme, at the end of this stage we made the collective decision to revise and rewrite our codebook in its entirety, as described below.

3.3. Stage 3: Fully revised codebook

Version 3 (our final version) of our coding schema is a significant departure from Ferguson et. al's [11] coding used in the previous two stages. In this third cycle of refinement, we simplified the categories to facilitate coders' use of the codes, standardize multi-coder agreement, and address more specifically the types of exploratory learning dialogue that we were observing on Reddit. The revised schema includes three explicit explanation categories (*Disagreement*, *Agreement*, *Neutral*), two socializing categories (*Negative, Positive*), two types of information exchange (*Information Seeking*, *Providing Resources*), and one category of learning subreddit norms (*Subreddit Rules and Norms*) (see Table 3). Version 3 of the codebook captures two trends observed in reading Reddit posts: the positive expressions and supportive dialogue and information provision that pull participants toward each other and foster topic-specific discussions, and the more negative exchanges that monitor and sanction behavior, silence participants, and can stifle online learner dialogue.

Results of our coding test for Version 3 showed a more acceptable level of agreement (Krippendorf's alpha) between coders: 'Ask_Politics' 0.52, 'AskAcademia' 0.64 and 'askscience' 0.67. In preparation for our validation processes, we also tested the final version of the coding schema with 'AskHistorians' 2015 subreddit sample (n=267) and recorded an alpha of 0.57. While these values are considered to be of moderate agreement, they are much stronger than in Version 1 of our coding schema. Along these lines, we note that Ferguson et al.'s [11] binary classification (exploratory or non-exploratory dialogue) recorded an inter-annotator agreement score of 0.597, which they understood as having 'moderate agreement', and thus reliable enough to train an automated classifier. In designing our study on exploratory learning dialogue, we anticipated that adding multiple coders (3) and codebook categories (8) to our methodology could potentially decrease or produce lower levels of intercoder agreement [5, 27, 28]. At this stage, we decided to test the validity of our coding schema with independent coders on a larger, more recent dataset (2016 'AskHistorians' subreddit sample).

Table 3. Reddit codebook version 3 (FINAL)

Code	Definition	Linguistic Dialogue Example
1. Explanation with Disagreement	Expresses a NEGATIVE take on the content of the previous comment by adding new ideas or facts to discussion thread.	'But', 'I disagree', 'not sure', 'not exactly' with explanation/ judgment/ reasoning/ etc.
2. Explanation with Agreement	Expresses a POSITIVE take on the content of the previous posts by adding new ideas or facts to discussion thread.	'Indeed', 'also', 'I agree', with explanation/ judgment/ reasoning/ etc.
3. Explanation with Neutral Presentation	Expresses a NEUTRAL explanation/judgment/reasoning/etc. with neither negative nor positive reference to the content of the previous comments, nor necessarily any reference to previous comments.	Comments with non-judgmental language. Advice, brainstorming and first hand experiences are framed neutrally. 'I can understand', 'interesting', 'depends on…' or statement responses.
4. Socializing with Negative Intent	Socializing that expresses negative affect through tone, words, insults, expletives intended as abusive.	'no', 'you're an idiot', 'this has been explained multiple times'
5. Socializing with Positive Intent	Socializing that expresses positive affect tone, words, praise, humor, irony intended in a positive way.	'thanks', 'great feedback', 'you're correct'
6. Information Seeking	Comments asking questions or soliciting opinions, resources, etc. ('Does anyone know …?' 'How does this work?'). This does not include questions answered rhetorically within the comment, e.g., if a question is asked and answered.	'First you have to think what happens if …?' and then you can see what happens', 'does anyone know', 'can anyone explain'
7. Providing Resources	Comments that include direct reference to a URL, book, article, etc.; comments that call upon a well-known theory or the name of a well-known figure.	Link to resource copied (book, URL, article, audio/video file). Referencing theory/theorists, scholar or public work (Einstein, Newton, Freud).
8. Subreddit Rules and Norms	Comments on topics such as what is the appropriate sub-reddit for a particular discussion, what language is appropriate to use, how to back up claims by using resources, etc.	'See/don't forget subreddit link', 'this post doesn't belong here', upvote/downvote mentions, acknowledging OP redditors, and bots.

4. Schema testing and evaluation process

The sample of comments used for the schema evaluation were obtained by first randomly arranging all threads from the collected 2016 'AskHistorians' subreddit

data. In total, there were 142,279 comments in response to 41,214 submissions (threads). However, because the data was collected retroactively some of the original comments were deleted either by the authors or the moderators. After removing the 'deleted' comments, the remaining number of comments were 122,670. We then took the first 1% of comments n=1,227 which constituted our sample for evaluation. The sample comments were then manually coded by three independent coders. Prior to undertaking the coding, each coder completed a schema tutorial training-module.

Table 4. Coding results*

	ask_Politics	askAcademia	askscience	askHistorians	askHistorians
Year	2015	2015	2015	2015	2016
Sample Size	190	198	164	267	1,227
1.Explanation with Disagreement	91 (48%)	21 (11%)	16 (10%)	34 (13%	71 (6%)
2.Explanation with Agreement	11 (6%)	20 (10%)	10 (6%)	4 (1%)	45 (4%)
3.Explanation with Neutral Presentation	45 (24%)	102 (52%)	100 (61%)	67 (25%)	592 (48%)
4.Socializing with Negative Intent	37 (19%)	5 (3%)	0 (0%)	0 (0%)	4 (0%)
5.Socializing with Positive Intent	2 (1%)	44 (22%)	19 (12%)	31 (12%)	204 (17%)
6.Information Seeking	22 (12%)	13 (7%)	23 (14%)	29 (11%)	274 (22%)
7.Providing Resources	20 (11%)	13 (7%)	33 (20%)	64 (24%)	260 (21%)
8.Subreddit Rules and Norms	3 (2%)	6 (3%)	2 (1%)	0 (0%)	66 (5%)

*Note: For the 2015 'training' datasets, the counts represent an agreement between two or more independent coders. Comments where two or more coders did not agree were not counted or included. For the 2016 validation dataset, the counts represent an agreement between two or more independent coders. Percentages may be higher than 100% when coders have assigned multiple (maximum three) codes per comment.

Results from the three independent coders showed a marked improvement in Krippendorff's alpha: 'AskHistorians' 0.76 (79% agreement). We regard this alpha level to be acceptable, when considering that we allowed multiple codes (maximum 3) per comment. For exploratory studies like ours, alpha levels between 0.67 and 0.80 are considered reliable enough to draw out and develop cautionary conclusions [27, 28].

The 2016 'AskHistorians' distribution of results shows that this subreddit can be viewed as a positive, communicative and knowledge-rich learning environment. Coding trends reveal a higher proportion of neutral explanations, positive socializing, information seeking and resource sharing behavior (see Table 4). For comparison, we include in Table 4 the results of the research team's coding of the 1% samples from the 2015 'AskHistorians', 'Ask_Politics', 'AskAcademia' and 'askscience', which demonstrates how the coding schema capture learning processes and conversations across different subreddits.

5. Discussion

In sum, the results show the proposed coding schema can capture subtle nuances in the way people converse across different subreddits. Distribution results from the 2015 and 2016 'AskHistorians' subreddit show that online conversations and social learning processes connect people, ideas and resources. The 'AskHistorians' community rules and norms emphasize external content/sources and academic-level answers, which may explain why these learning behaviors are observed. Similarly, we note that the 'askscience' rules and norms also encourage Redditors to remain civil, avoid speculation and to answer questions with reputable sources, which could explain why subreddit dialogue is more functional in nature. Distribution results from the 2015 'askscience' subreddit also highlight a more resource-rich, transactional, and neutral Q&A learning environment.

In both of the above cases, we found the subreddit community to promote collaborative and participatory dialogue, which help encourage self-directed learning practices. The 2015 'ask_Politics' distribution results conversely show a greater proportion of comments with negative socializing, disagreement and debate even though the subreddit's rules and norms stipulate that posts should be reputable, civil, sourced and remain on-topic. We hypothesize that the personal and normative nature of politics inadvertently fuels more argumentative, opinion-based comments between Redditors, where there is no 'right' or 'wrong' answer in an objective sense. This is not to suggest that disagreements are counterproductive to learning. Rather, explanations with disagreements, arguments, debates, negotiation and alternative viewpoints can encourage processes of learning (and unlearning). From 'ask_Politics' we can glean that even with moderated rules/norms, the anonymity of the Reddit platform can sometimes prompt critical and disputational learner conversations, potentially leading to transgressions between Redditors.

In contrast to the above subject-led subreddits, 'AskAcademia' is a professionally-focused subreddit open to anyone interested in academia and academic careers/life. Distribution of results from 'AskAcademia' highlight a new range of self-directed learner practices that do not necessarily have a curricula/subject

counterpart. Rather, comments in this subreddit are found to be more neutral, supportive, reflective and socially positive; appealing to budding academics by focusing on personal needs.

Overall, the research shows that learning processes in open, online social networking sites like Reddit can foster individual learning outcomes which can help self-motivated learners sustain online group dynamics and communities of practice. And while often focused on niche topics and interests, these online participatory practices are part of a much wider trend towards lifelong informal learning that calls for more research attention. University instructors are increasingly looking to incorporate social media into their course curriculum, as a way to connect students and extend classroom learning environments to include discussions occurring in the outside world [7, 15]. In today's social media age, teaching and learning activities are taking place across informal and formal settings, and require new analytical frameworks and coding schemes. Recognizing the need for more precise consideration of these dynamic learning processes, our schema contributes a novel framework to better capture the social, conversational and collaborative elements (all defining features) of informal, online learning environments.

6. Conclusion

This paper has reported on the development and refinement of the proposed 'learning in the wild' coding schema. We have shown the validity and utility of our coding schema when studying unstructured, informal learning processes through analysis of four diverse 'Ask' subreddit communities. We used three independent coders to evaluate the schema, and recorded an alpha of 0.76 (79% intercoder agreement) for the 'AskHistorians' 2016 sample. In doing so, we highlighted different spheres of knowledge, informal learning practices and exploratory dialogue that occur in online settings, outside of traditional educational and professional environments. The research has reasserted the potential of social media sites such as Reddit to support self-motivated learners and sustain communities of practice. We intend to expand this research, first by validating the proposed coding schema with a larger sample of subreddits, and then across other social media platforms (e.g. Twitter, Facebook, LinkedIn). As such, we invite other scholars to apply our schema to their research on informal learning in open, online environments. Upon further validation, we intend to integrate automatic machine learning to our research.

7. Acknowledgments

This work is supported by a Social Sciences and Humanities Research Council of Canada (SSHRC) grant, "Learning Analytics for the Social Media Age", PIs: Anatoliy Gruzd and Caroline Haythornthwaite. The authors would like to thank Nadia Conroy, Michael Pacheco, and Jordan Kilfoy, who helped with the manual coding of reddit posts. We would like to thank anonymous reviewers for providing very helpful comments.

8. References

[1] Alexa: Actionable Analytics for the Web. *Analytics for Reddit.com*. Retrieved on April 11, 2017. http://www.alexa.com/siteinfo/reddit.com

[2] Anderson, K.E., "Ask Me Anything: What is Reddit?", Library Hi Tech News, 32(5), 2015, pp.8-11.

[3] Bransford, J.D., A.L. Brown, and R.R. Cocking, eds, *How People Learn: Brain, Mind, Experience, and School*, National Academy Press, Washington, DC, 1999.

[4] Buckingham Shum, S., and R. Ferguson, "Social Learning Analytics", Educational Technology & Society, 15(3), 2012, pp. 3-26.

[5] DeCuir-Gunby, J.T., P.L. Marshall, and A.W. McCulloch, "Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project", Field Methods, 23(2), 2011, pp.136–155.

[6] Dolezal, M., L. Ennser-Jedenastik, W.C Muller, and A.K, Winkler, "How Parties Compete for Votes: A test of Saliency Theory", European Journal of Political Research, 53(1), 2014, pp.57-76.

[7] Esteve Del Valle, M., A. Gruzd., C. Haythornthwaite, D. Paulin, and S. Gilbert, "Social Media in Educational Practice: Faculty Present and Future Use of Social Media in Teaching", in Proceedings of the 50th Hawaii International Conference on System Sciences, 2017.

[8] Ezen-Can, A., and K.E. Boyer, "Understanding Student Language: An Unsupervised Dialogue Act Classification Approach", JEDM-Journal of Educational Data Mining, 7(1), 2015, pp.51-78.

[9] Feng, G. C, "Intercoder Reliability Indices: Disuse, Misuse, and Abuse" Quality and Quantity, (48), 2014, pp.1803–1815.

[10] Ferguson, R., and S.B. Shum, "Learning Analytics to Identify Exploratory Dialogue within Synchronous Text Chat", in P. Long, G. Siemens, G. Conole, & D. Gasevic, eds., Proceedings of the 1st International Conference on Learning Analytics and Knowledge, ACM, New York, USA, 2011, pp.99-103.

[11] Ferguson, R., Z. Wei, Y. He, and S. Buckingham Shum, "An Evaluation of Learning Analytics to Identify Exploratory Dialogue in Online Discussions", in D. Suthers, K. Verbert, E. Duval, X. Ochoa, eds., Proceedings of LAK'13, Leuven, Belgium, 2013, pp.85-93.

[12] Finlay, S.C., "Age and Gender in Reddit Commenting and Success", Journal of Information Science Theory and Practice 2(3), 2014, pp.18-28. 16
[13] Gilbert. S., "Learning in a Twitter-based Community of Practice: an Exploration of Knowledge Exchange as a Motivation for Participation in #hcsmca", Information, Communication & Society, 19(9), 2016, pp. 1214-1232.
[14] Gruzd, A., and C. Haythornthwaite, *Networking online: Cybercommunities*, in J. Scott & P. Carrington, eds., *Handbook of Social Network Analysis*, Sage, London, 2011.
[15] Gruzd, A., C. Haythornthwaite, D. Paulin, S. Gilbert and Esteve del Valle, M. "Uses and gratifications factors for social media use in teaching: Instructors' perspectives", *New Media and Society*, doi: 10.1177/1461444816662933.
[16] Gruzd, A., D. Paulin, and C. Haythornthwaite, "Analyzing Social Media and Learning through Content and Social Network Analysis: A Faceted Methodological Approach", Journal of Learning Analytics, 3(3), 2016, pp.46-71.
[17] Hase, S., and C. Kenyon, "From Andragogy to Heutagogy", Ultibase Articles, 5(3), 2000, pp. 1–10.
[18] Haythornthwaite, C., "Facilitating Collaboration in Online Learning", Journal of Asynchronous Learning Networks, 10(1), 2006.
[19] Haythornthwaite, C., *Social Networks and Online Community*, in Joinson, A., K. McKenna, U. Reips, and T. Postmes, eds., *Oxford Handbook of Internet Psychology*, Oxford University Press, Oxford, UK, 2007.
[20] Haythornthwaite, C., *Learning Networks*, in R. Alhajj and J. Rokne, eds., *Encyclopedia of Social Network Analysis and Mining*, Springer Science+Business Media, New York, 2014.
[21] Haythornthwaite, C., and R. Andrews, *E-learning Theory and Practice*, Sage, London, 2011.
[22] Haythornthwaite, C., and E. Meyers, eds., "New Media, New Literacies and New Forms of Learning", International Journal of Learning and Media, 4(3-4), 2012, pp.1-8.
[23] Haythornthwaite, C., M. de Laat, and B. Schreurs, *A Social Network Analytic Perspective on E-Learning*, in C. Haythornthwaite, R. Andrews, J. Fransman and E. Meyers, eds., *Handbook of E-Learning Research*, Sage, London, 2016.
[24] Hernández-García, Á., I. González-González, A.I. Jiménez-Zarco, and J. Chaparro-Peláez, "Applying Social Learning Analytics to Message Boards in Online Distance Learning: A Case Study", Computers in Human Behavior, 47, 2015, pp. 68-80.
[25] Jackson, N., *Learning for a Complex World: A Lifewide Concept of Learning, Education and Personal Development*, Author House Publishing, UK, 2011.
[26] Knight, S., and K. Littleton, "Discourse, computation and context – sociocultural DCLA revisited', Paper presented at 1st International Workshop on Discourse-Centric Learning Analytics 2013, Leuven, Belgium, April 2013.
[27] Krippendorff, K., "Reliability in Content Analysis", Human Communication Research, 30(3), 2004, pp. 411–433.
[28] Krippendorff, K., *Content Analysis: An Introduction to Its Methodology*, Sage, Newbury Park, CA, 1980.
[29] Lave, J., and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press, Cambridge, UK, 1991.
[30] Loudon, M., "Research in the wild in online communities: Reddit's resistance to SOPA", First Monday, 19(2), 2014.
[31] Luckin, R., *Re-Designing Learning Contexts: Technology-Rich, Learner-Centred Ecologies*, Routledge, Abingdon, UK, 2010.
[32] Mercer, N., "Sociocultural Discourse Analysis: Analysing Classroom Talk as a Social Mode of Thinking", Journal of Applied Linguistics, 1(2), 2004, pp.137-168.
[33] Miyanke, M., *Computer Supported Collaborative Learning*, in R. Andrews, and C. Haythornthwaite, eds., *Handbook of E-Learning Research*, Sage, London, 2007.
[34] Moore, C., and L. Chuang, "Redditors Revealed: Motivational Factors of the Reddit Community", in Proceedings of the 50th Hawaii International Conference on System Sciences, 2017, pp. 2313-2322.
[35] Nistor, N., B. Baltes, M. Dascălu, D. Mihăilă, G. Smeaton, and Ş. Trăuşan-Matu, "Participation in Virtual Academic Communities of Practice under the Influence of Technology Acceptance and Community Factors. A Learning Analytics Application", Computers in Human Behavior, 34, 2014, pp. 339-344.
[36] Paulin, D., and C. Haythornthwaite, "Crowdsourcing the Curriculum: Redefining e-learning Practices through Peer-Generated Approaches", The Information Society, 32(2), 2016, pp. 130–142.
[37] Reddit, About Us. Retrieved on April 11, 2017. https://about.reddit.com
[38] Sankin, A., "7 sites to try during Reddit's meltdown", *The Daily Dot*. Retrieved on May 27, 2017. https://www.dailydot.com/layer8/reddit-alternatives-goodbye-cruel-world/
[39] Shulman, S., "DiscoverText: Software Training to Unlock the Power of Text", in Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times, ACM, NY USA, 2011, pp. 373-373.
[40] Siemens, G., "Learning Analytics: The Emergence of a Discipline", American Behavioral Scientist, 57(10), 2013, pp.1380-1400.
[41] Wenger, E., *Communities of Practice: Learning, Meaning, and Identity*, Cambridge University Press, Cambridge, 1998. 54
[42] Wilson, L., and A. Gruzd, "MOOCs – International Information and Education Phenomenon?", Bulletin of the American Society for Information Science and Technology, 40(5), pp. 35-40.
[43] Zhou, L., B. Li, W. Gao, Z. Wei, and K.F. Wong, "Unsupervised Discovery of Discourse Relations for Eliminating Intra-Sentence Polarity Ambiguities" in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp.162-171