Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 31/Mar/2022; 1st round notif.: 27/Jul/2022; New version: 9/Aug/2022; 2nd round notif.: 23/Sep/2022;
Camera ready: 5/Oct/2022; Edition review: 26/Oct/2022; Available online: 27/Oct/2022; Published: 27/Oct/2022;

SLIM: a process for analyzing learners' behavior and discourse within large online communities

Rogério F. da Silva
Universidade Federal do Paraná (UFPR)
ORCID:0000-0003-2151-226X
rogerio.ferreira@ufpr.br

Abstract

*Educational researchers have an increasing interest in systematically assessing social learning that takes place in large online communities, nowadays one of the most important producers of Big Data in education. However, there is no agreement on how to measure the performance of such communities in informal learning settings. Assessing online Social Learning (SL) is a complex process that calls for an analytical approach in order to understand the various dimensions of learner discourse and the structure of the social interactions. This paper presents SLIM (Process for assessing online **S**ocial **L**earning within online communities in **I**nformal environ**M**ents): a process that combines structure and discourse analyses to assess SL indicators within large Online Learning Communities (OLC). Initially, we have used data provided by informal environments to perform Social Network Analysis (SNA) in order to identify conditions and behavioral patterns associated to learning. Next, we have incorporated these data into an unsupervised machine learning method to identify a discourse style related to learning. SLIM has been initially applied to two large online communities from the news sharing site Reddit. We are interested in characterizing and assessing the massively distributed learning, and just-in-time learning associated with the development of sustained online communities in informal environments. The results point out a set of quantitative measures and machine learning models that can be used to outline the evolution of SL indicators over time. They suggest that participation, ongoing collaboration and positive emotion have a fundamental role for knowledge creation and sharing. These findings can be used to take actions in order to regulate social interaction within large OLC.*

Keywords: *Social learning, Informal learning environments, Online communities, Assessment, Process*

Cite as: da Silva, R. F. SLIM: a process for analyzing learners' behavior and discourse within large online communities. Revista Brasileira de Informática na Educação, 30, 573-597. DOI: 10.5753/2022.2614.

1 Introduction

Learning is increasingly seen as most effective when it is collaborative in nature (B. Chen, Chang, Ouyang, & Zhou, 2018). Social Learning (SL) is an interactive and dynamic process that takes place in a multi-actor setting where actors learn and co-create new knowledge in ongoing collaboration (Kent, Rechavi, & Rafaeli, 2019). According to educational research, asynchronous text-based discussions are a resource with high potential for promoting collaboration in online SL, supporting students' interactions in contexts of learning that follow the social constructivist paradigm (De Wever, Schellens, Valcke, & Van Keer, 2006; Ferreira et al., 2020). The advantages of such discussions as compared to synchronous ones have been associated with their spatial and time flexibility, which enable students to refine their commitment, and reflect on learning content at any time (Kim, Yoon, Jo, & Branch, 2018). Thus, investigating the students' interactions and the content of discussions lead to identifying patterns of activity and discourse styles that indicate meaningful knowledge construction (De Laat & Prinsen, 2014; Haythornthwaite et al., 2018).

Online SL based on asynchronous discussions takes place in formal and informal settings (Ferguson & Shum, 2012; De Laat & Prinsen, 2014). The distinction between formal and informal learning can be controversial, and researchers have proposed a variety of definitions, at times, conflicting (Czerkawski, 2016; Hudgins et al., 2020). Despite these differences, in general, formal learning refers to hierarchically structured and chronologically paced learning activities that are facilitated by an instructor (Schreurs & De Laat, 2014; Czerkawski, 2016). It is usually applied to educational institutions and implemented by Learning Management Systems (LMS). It keeps track of student performance and leads to a degree or certification (Czerkawski, 2016; Corbi & Burgos, 2020). On the other hand, informal learning refers to unstructured, not teacher-led, and in most cases, spontaneous learning which occurs outside the conventional educational systems (Ferguson & Shum, 2012; Czerkawski, 2016). Online Learning Communities (OLC) have been referred to, in educational research, as natural environments for informal learning (Nistor, Dascalu, Tarnai, & Trausan-Matu, 2020; Hudgins et al., 2020). Students use such communities to bridge the gap between the traditional curricula and their personal interests, expanding their learning opportunities (Gruzd, Paulin, & Haythornthwaite, 2016). As learners progress through school and towards professional life, informal learning becomes essential to developing knowledge and skills in lifelong learning experiences (Gruzd et al., 2016). OLC offer the possibility of connecting students with their members of diverse expertise and support the involvement in a knowledge-creating culture (Gruzd et al., 2016; Nistor et al., 2020).

In the past few years, the discussion about technologies for learning has moved away from only institutionally managed systems, for example LMS, to informal learning environments (Galanis, Mayol, Alier, & García-Peñalvo, 2016; Haythornthwaite, de Laat, & Schreurs, 2016; Chatti, Muslim, & Schroeder, 2017; Rezaei, Bobarshad, & Badie, 2019). Educational researchers have suggested that informal environments may also be particularly supportive of learning. They have argued that a considerable amount of knowledge exchange occurs through informal interactions with peers, reading, and observation (Greenhow, Gibbins, & Menzer, 2015; Galvin & Greenhow, 2020). Students' digital footprints provide vast amount of implicit knowledge and a new perspective for academics and professionals to understand students' experiences outside the controlled formal settings (X. Chen, Vorvoreanu, & Madhavan, 2014). Learning and collaborating in informal contexts are nowadays an ubiquitous phenomenon and an important producer of Big Data in

education (Nistor, Derntl, & Klamma, 2015). However, the investigation of learning experiences within informal environments have been relatively underrepresented (Hudgins et al., 2020). Thus, little is known about learning processes within such environments (Greenhow et al., 2015).

This paper presents SLIM: a process for assessing online Social Learning indicators within large online communities in informal environments. By large communities we refer to those with more than one million users enrolled. This number refers to the context in which our study is addressed to. The SL indicators are represented by a set of quantitative measures which suggest conditions and behavioral patterns related to learning. They are based on the guidelines proposed in the value creation (Wenger, Trayner, & De Laat, 2011) and social presence (Garrison, Anderson, & Archer, 2010) theoretical frameworks. Whereas our overall aim is to develop a general process that can be applied to different informal learning environments, we have initially defined and refined our method working with two large communities from the online news sharing site Reddit (Weninger, 2014). Thus, the objectives of this paper are the following:

1. Present a process for assessing SL indicators within large online communities, based on asynchronous text-based discussions. The process is not geared towards measuring specific learning outcomes. Instead, by assessing we mean identifying and analyzing a set of conditions and behavioral patterns strongly associated with learning, according to the body of research described along the paper.

2. Combine structure and discourse analyses to analyze a set of quantitative measures that support our process.

3. Employ interactive visual representations and exploratory educational data analysis, in order to investigate the evolution of the most relevant measures over time (see subsection 4.3.1). This investigation can amplify cognition and generate insights to the broader understanding of social learning within large OLC.

The main contribution to the previous literature is that we instantiate our process, characterize and assess two real online communities of different domains from Reddit, emphasizing their similarities and differences in order to provide a diagnostic about large scale social learning in informal settings. The conclusions revealed that the most important measures correlated to behaviors and conditions associated with learning are the ones related to amount of participation. Moreover, the analysis of discussions identified that positive and negative emotion may influence the ongoing collaboration. The findings can be used to extract useful insights about learning in online communities, helping educational researchers to investigate SL process in such environments. In addition, the results aim to help learners to become more aware of the productivity of their social connections and contributions to the community.

The structure of the paper is defined as follows. Section 2 introduces the theoretical background and related works. Section 3 describes our methods and tools. Section 4 presents the application of SLIM process and the main results. Finally, Section 5 describes the conclusions.

2 Background and Related Works

A criticism often voiced about assessing social learning, or analyzing educational data in general, is its atheoretical nature (De Laat & Prinsen, 2014; Toikkanen & Lipponen, 2011). A challenging aspect of educational research lies in the theoretical framework it is embedded in (Nistor et al., 2015). Based on this perspective, we have used the frameworks of value creation (Wenger et al., 2011) and social presence (Garrison et al., 2010) to support SNA and discourse analysis in evaluating online social learning. The next subsections present the background of online learning communities. Moreover, we describe the theoretical frameworks and some observable research gaps that support our process.

2.1 Social Learning and Learning Communities

The underlying premise of SL is that learning is explained as a social process, enacted through interactions, knowledge exchanges, conversations and collaborations (Shum & Ferguson, 2012). The connections between people form links (or ties), which in turn, form networks of actors connected. Thus, learning can emerge as an outcome of forming and maintaining such interaction network (Haythornthwaite, 2018). An important issue refers to when an interaction network becomes a learning community. Haythornthwaite (2018) argues that "a key transition to a learning community entails going from personal information seeking to collective practices associated with a [...] participatory culture of open exchange" (Haythornthwaite, 2018, p. 7). Learning communities serve a personal but shared need organized around a specific subject. They represent a collective intention to steward a domain of knowledge and to sustain learning about it (Haythornthwaite, 2018; Wenger et al., 2011).

Learning communities have emerged from informal learning environments, such as social networking sites, open discussion forums or question and answer sites, where the traditional roles of teacher and learner become defined based on their behavior in the environment, rather than on pre-defined roles. For example, moderators arise in many online communities, recognized from their peers, who manage knowledge exchanges, monitor adherence to the discussion topic and appropriate user's behavior (Haythornthwaite, 2018). We use the term online communities or OLC referring to large online learning communities in informal settings based on asynchronous text-based discussions. Our specific interest have been the communities which focus in the co–construction of knowledge, meaning and understanding based on a set of SL activities, such as asking questions, sharing information or tips, learning from each other's experience, helping each other with their challenges and creating knowledge in ongoing collaboration. Educational researchers have addressed the lack of tools and methods to measure the learning effectiveness in such communities (Pesare, Roselli, & Rossano, 2016). Our process aims to cover this research lack by using the theoretical frameworks as described in next subsections.

2.2 Value Creation Framework and Social Network Analysis

Value creation refers to the knowledge created by the involvement of learning communities when they are used to promote SL activities (Wenger et al., 2011). Such communities can generate several sorts of qualitative and quantitative data about their activities. The value creation framework

provides a set of indicators for an evaluation process that can integrate heterogeneous sources and types of data to create a picture of how online communities create value for their members, hosting organizations and sponsors (Wenger et al., 2011). Wenger et al. (2011) suggest that SNA provide a good basis for talking about the value that networking has for community members.

SNA is a method and set of principles for studying relational connections between actors in a network (Haythornthwaite, 2018). In the context of SL, SNA helps to understand how learners are connected and how they interact with each other. The measures defined in this method can provide detailed information about the nature of student participation. At an individual level, SNA takes an egocentric perspective where the network of a particular node (user) is being studied. Measures as degree, betweenness centrality, authority and pagerank are generally associated to how influential a node is within the network. On the other hand, a social network perspective can be used to study the whole network. Measures as density, diameter and reciprocity enable accounting for the importance of group dynamics and provides comprehensive insights into the quantity and quality of social interactions within the network (Shum & Ferguson, 2012; Haythornthwaite et al., 2016; Joksimovic et al., 2019). The Appendix B describes how these SNA measures have been interpreted in social learning investigation contexts. In general, SNA has played a prominent role in the learning sciences for evaluation of educational settings (Gašević, Joksimović, Eagan, & Shaffer, 2019; Joksimovic et al., 2019). However, the analysis of the network of connections is not enough for deeply understanding patterns of interactions in learning environments. Therefore, we argue that discourse analysis should be applied together with SNA and with the purpose of creating a holistic approach.

2.3 Discourse Analysis and Social Presence

Discourse analysis is the collective term for a wide variety of approaches for analyzing series of communicative events. Some of these approaches provide ways of understanding the large amounts of text generated within online environments, supporting the comprehension of the discourse dimension of online interactions (Shum & Ferguson, 2012). The social constructivist theoretical framework named Community of Inquiry (CoI) (Garrison et al., 2010) is one of the most used models to outline how asynchronous online communication shapes student learning (Ferreira et al., 2020; Kovanović et al., 2018). CoI proposes three key dimensions, known as presences (Ferreira et al., 2020): ***(i)*** **Social presence** measures the ability to humanize the relationships among participants in a discussion. It focuses on social interactions and tries to model the social climate within a group of learners. ***(ii)*** **Cognitive presence** is highly related to the development of learning outcomes. It aims to capture the progress of interactions throughout students' cognitive processes that support the development of critical thinking and knowledge construction. ***(iii)*** **Teaching presence** concerns teaching role during online courses.

The most important purpose of social presence is to provide a comfortable environment for participants to exchange ideas freely, explore different perspectives and solve problems collectively (Joksimovic, Gasevic, Kovanovic, Riecke, & Hatala, 2015). It motivates participants to post their tentative ideas and also to criticize others' hypotheses (Rourke, Anderson, Garrison, & Archer, 2001). Social presence, as defined in the CoI model, includes three categories, described as follow (Ferreira et al., 2020). ***(i)*** **Affective**: this category analyses the translation of real emotions into text. It encompasses expression of emotions, feelings, and self-disclosure.

(ii) **Interactive**: this category focuses on the interactivity of the messages exchanged amongst participants. It includes continuing a discussion thread, asking questions, referring explicitly to others' message, expressing appreciation and agreement. ***(iii)*** **Group Cohesion**: this category investigates the sense of union and group commitment among students. It encompasses vocatives, references to the group and salutations.

The participants' ability to project themselves within an online community and the level of their communication with peers is initially identified in the social presence. It supports cognitive objectives through its ability to instigate, sustain, and encourage critical thinking in a community of learners (Rourke et al., 2001). Thus, we have used the indicators of social presence in discourse analysis, because it is essential to support the expression of the participants' thoughts in text messages (Ferreira et al., 2020). SLIM process has used the well-known linguistic framework LIWC (Linguistic Inquiry Word Count) (Pennebaker, Boyd, Jordan, & Blackburn, 2015) to extract indications of social presence from textual contributions. LIWC is one of the most suitable tools for analysis of online discussion messages, and assessment of various psychological constructs (Joksimovic, Gasevic, Kovanovic, Adesope, & Hatala, 2014; Lin, Yu, & Dowell, 2020). It extracts 93 measures divided into the categories as summary of language variables, linguistic dimensions, grammar, social processes, affective processes, and others. LIWC is based on the word count strategies, which is geared toward revealing the psychological meaning of words, independently from their literal and semantic contexts (Pennebaker et al., 2015).

2.4 Related Work

Educational researchers have done a significant effort in the last decade to understand how social interactions can leverage learning. This subsection describes some research gaps of available studies, both in formal and informal settings, that led to the development of our approach.

2.4.1 ***Lack of combination of analytical methods***

The connection between social interactions and discourse analysis within a network is well established in numerous anthropological, sociological and sociolinguistic studies (Scott & Carrington, 2011). However, the combination of these methods in an analytic approach is notably an educational research gap (Gašević et al., 2019; Joksimovic et al., 2015). Joksimovic et al. (2019) propose an approach that extracts speech acts as representations of knowledge construction processes. The authors integrate the use of discovered speech acts to explain the formation of social ties and predicting course outcomes. Statistical network analysis and regression models showed that the combined use of measures derived from discourse analysis and social ties predicted learning outcomes. However, the authors have focused on investigating only SNA centrality measures.

2.4.2 ***Small sample sizes and the investigation of specific contexts***

The sample size of majority of the studies that investigates online learning settings is small. Some authors explicitly state this as a limitation (Dascalu, McNamara, Trausan-Matu, & Allen, 2018; Ferreira et al., 2020; Jan, 2019; Swiecki & Shaffer, 2020). Swiecki and Shaffer (2020) proposed the integrated social-epistemic network signature (iSENS), an approach that affords the simultaneous investigation of cognitive and social patterns. They modeled such patterns using an integrated

Epistemic Network Analysis (ENA). The approach was tested on data collected from collaborative problem solving (CPS) of military teams in training. Their findings suggested that these teams are defined by specific patterns of cognitive and social connections. However, according to the authors, the results use data from one context only. Thus, it could not be concluded that iSENS is a better approach for CPS data in general.

*2.4.3 **The need to investigate dynamics over time***

A pressing need for investigating informal learning settings is a move away from static analyses that observe an OLC at one point in time to pursue instead systematic accounts of how such communities change over time (Schreurs & De Laat, 2014). Becheru, Calota, and Popescu (2018) proposed a SNA-based platform for visualizing students' collaboration patterns that integrates several social media tools, such as blogs and wikis. The authors implemented a list of visualization needs outlined by teachers, such as exhibit the general status of collaboration and the status of collaboration for each learner. However, they did not provide more details about the behavior trends to explore the temporal dynamics of interactions.

SLIM aims to bridge the research gaps above mentioned based on a multistage methodology that performs educational data analysis in large scale, as described in next Section.

3 Methods and tools

Generally, informal learning environments have no compulsory assessment procedures. For that reason, OLC commonly provide a peer assessment process performed by participants when interacting with each other. It is based on a reward system and displays for all community members a point scheme that recompenses the frequency and quality of individual participation (Hudgins et al., 2020). SLIM has been applied to data obtained from OLC within online news sharing site Reddit[1]. Nowadays, Reddit comprises approximately 52 million daily active users, 303.4 million posts and 2 billion comments per year[2]. Participants, known as *redditors*, can evaluate (positively or negatively) the discussion topics, creating their score. In general, the positive votes associated with a particular discussion indicate the community's opinion about it. Thus, the topics with the best answers will likely be rated with higher scores (Hudgins et al., 2020). *Redditors* are also able to assign points to each other responses. These points, named *karma*, indicate the members' expertise and reflect their popularity (Silva, Gimenes, & Maldonado, 2020). Discussion score and *karma* points comprise the Reddit peer assessment data. These data are used to support the SLIM process.

The Fig. 1 depicts an activity diagram that represents the SLIM process. The activities are divided in three partitions. They are described in next subsections.

[1] http://www.reddit.com

[2] http://redditblog.com/2020/12/08/reddits-2020-year-in-review/

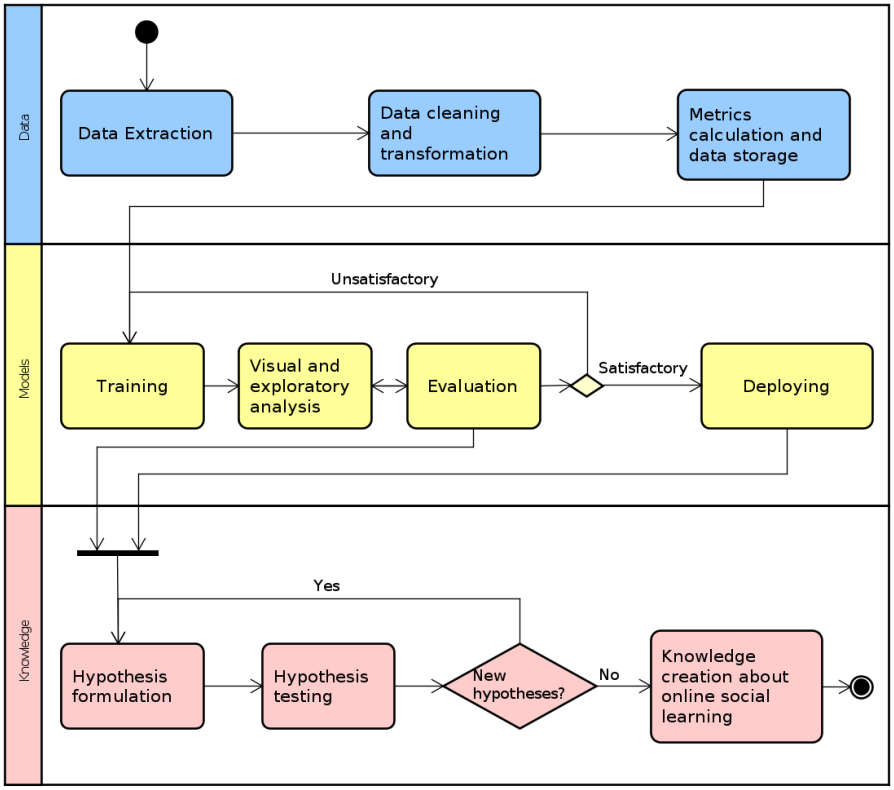

Figure 1: SLIM process and its activities.

3.1 Activities related to Data

The partition named Data comprises activities related to extracting, cleaning and transforming, calculating metrics, and storing data. We have extracted data from OLC by using two Python packages for Reddit official Application Programming Interface (API): PSAW[3] and PRAW[4]. The data extracted are related to community participants, their interactions, discussions and peer assessment information. These data have been used to compute the measures that assess user behavior and discourse, described as follow:

- structured measures: they refer to SNA measures that helps to understand how participants are connected and how they interact with each other. In order to analyze the hierarchical structure of messages posted by participants, we have computed in this stage additional measures, such as number of participants in the discussion, discussion size and time of first reply.
- Discourse measures: we have used the results pointed out in studies which aim to investigate how participants' social presence is manifested in online discussions (Silva et al., 2020; Ferreira et al., 2020; Lin et al., 2020; Zhu, Herring, & Bonk, 2019; Zou et al., 2021). These studies recognize that the LIWC measures presented in Table 1 are important for identifying social presence in online discussions.

[3]http://github.com/dmarx/psaw
[4]http://praw.readthedocs.io

- Peer assessment data: they refer to discussion score and participant *karma* points. They are used to fit the models described in next subsection.

Table 1: LIWC measures of Discourse Analysis (adapted from Silva et al. (2020)).

Category	Measures	Description
Affective	liwc.pronoun	Number of pronouns
	liwc.ppron	Number of personal pronouns
	liwc.i	First-person singular pronouns
	liwc.ipron	Number of impersonal pronouns
	liwc.affect	Affective processes
	liwc.posemo	Positive emotion
	liwc.negemo	Negative emotion
	liwc.work	Personal concerns: work
	liwc.power	Words related to power
	liwc.drives	Words related to drives
	liwc.percept	Perceptual processes
	liwc.negate	Negations
Interactive	liwc.interrog	Interrogatives
	liwc.focuspresent	Focus on the present
	liwc.auxverb	Auxiliary verbs
	liwc.you	Second-person pronouns
	liwc.assent	Words related to assent
	liwc.focuspast	Focus on the past
Cohesive	liwc.we	First-person plural pronouns
	liwc.affiliation	Affiliation
	liwc.social	Social processes

3.2 Activities related to Models

The partition named Models represents activities related to training, evaluation and deploying three machine learning models created by Silva et al. (2020). These models identified the significant structured and discourse measures associated with the best rated discussions. They are described as follow.

- Relevant structured measures: a linear regression model has identified the most significant structured measures, analyzing which of them are more strongly associated with the best rated discussions. The results have pointed out the measures related to the amount of participation: *(i)* number of participants in the discussion; *(ii)* discussion size; *(iii)* discussion width; *(iv)* number of bottlenecks; and *(v)* number of triangles (or triads).
- Relevant discourse measures: the clustering algorithm *Kmeans* have grouped the discussion topics in order to identify which measures are more strongly associated with the best rated discussions. The results have pointed out eight LIWC measures: *(i)* positive emotion; *(ii)* affective processes; *(iii)* drive words; *(iv)* perceptual processes; *(v)* assent words; *(vi)* affiliation words; *(vii)* focus present; and *(viii)* negative emotion.

- Evolution of measures: it refers to multiple time series models to explore temporal dynamics of structured and discourse measures, in order to reveal the evolution of participants' behavior and discourse in the period under investigation.

The activity Visual and exploratory analysis aims to help data analysts in the evaluation of the models described previously. The analysts can perform this activity to realize exploratory data analysis in order to create insights about social learning. In addition, informing participants of their level of interaction and increasing awareness of the status of collaboration with their peers, may lead to enhanced self-regulation of social interaction and knowledge sharing in online communities (Joksimovic et al., 2015). In order to operationalize the activity Visual and exploratory analysis, we have created a Social Learning Analytics Dashboard (SLAD) that aims to visually trace participants' behavior detected in the machine learning models. Figure 2 shows the aspect of our SLAD. The main characteristics are described as follow:

- Define parameters (see Fig. 2-**A**) - it refers to parameters that configure the data visualization: ***(i)*** Select OLC - it allows to choose one or more OLC data, with the purpose of comparing their similarities and differences; ***(ii)*** Select trend scale - it applies a method to standardize the time series, in order to present the measures at the same scale; and ***(iii)*** Select type of analysis - it allows to exhibit data of structured or discourse measures.
- Outline temporal trends (see Fig. 2-**B**) - it shows the behavior over time of the most relevant structured and discourse measures according to the models fitted previously. The data viewing period can be shortened in order to investigate specific time intervals.
- Select measures (see Fig. 2-**C**) - it allows to disable some measures, with the purpose of emphasizing that ones which are important for the activity of visual and exploratory data analysis.

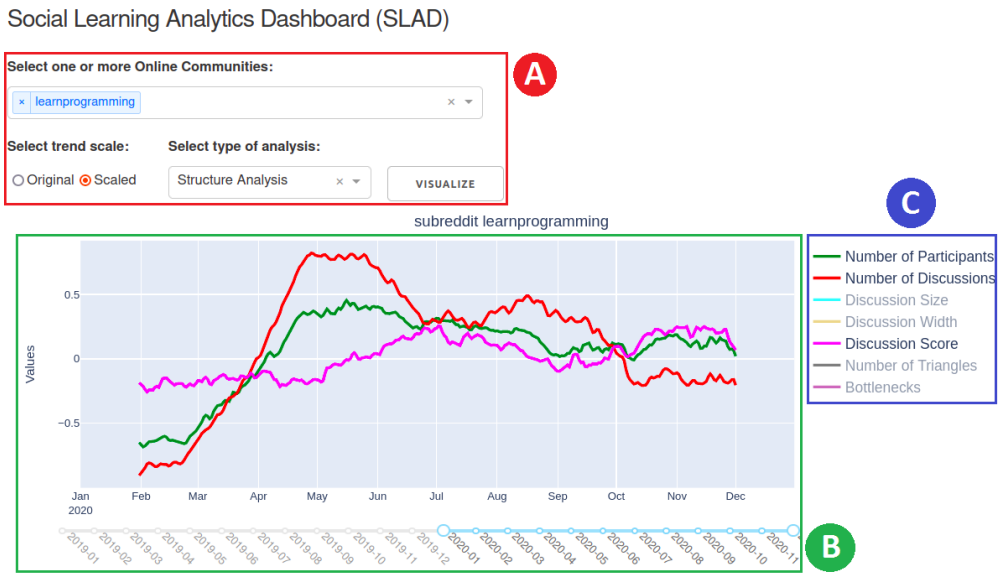

Figure 2: Design of our SLAD.

3.3 Activities related to Knowledge

The partition named Knowledge comprises the activities responsible for formulating and testing hypotheses that aim to generate knowledge and provide a broader understanding of social learning in OLC. Such hypotheses are formulated based on insights about social learning supported by the SLAD in the activity Visual and exploratory analysis. With the aim of testing the hypotheses we have used the Exponential Random Graph (ERG) modelling. ERG is a true generative statistical model of network structure and characteristics. This model can reveal the association between network effects and behavioral patterns traditionally related to learning. Such network effects are described as follow:

- **Reciprocity** reflects participants' tendency to form reciprocal ties and cluster together (Fincham, Gasevic, & Pardo, 2018). This effect is an indicator of mutual exchange depth, continuity, collaboration and negotiation of meaning (Jan, 2019). It reveals whether participants tend to continue interaction with peers who replied to their posts (Gašević et al., 2019).
- **Simple connectivity** reveals the propensity to participate. It is represented by a relationship between sending and receiving messages, that is, users who receive messages are more likely to send them and vice versa (Mamas, Bjorklund Jr, Daly, & Moukarzel, 2020).
- **Popularity**: we have uncovered In-degree as a relevant measure for identifying popular or expert users (Silva et al., 2020). Thus, popularity was modelled by the geometrically weighted in-degree distribution (gwidegree), an ERG statistic that captures the popularity effect (Fincham et al., 2018).
- **Transitivity** indicates the creation of alternative paths that facilitate the information flow in the interaction network (Jan & Vlachopoulos, 2018).

4 Applying the SLIM process

4.1 Data Collection

The SLIM process was applied to data obtained from two Reddit OLC, named *subreddits*, *learnprogramming*[5] and *MachineLearning*[6]. They had, respectively, 3,440,477 and 1,935,702 members enrolled at the evaluation snapshot in June 2021. In both *subreddits* we have extracted all discussion topics, replies and peer assessment data posted between 2019-Jan-01 and 2020-Dec-31. These *subreddits* were chosen because they are domain oriented OLC, where the members are focused on learning a specific domain. Table 2 shows the details of data extracted and analyzed.

4.2 Characterization of online communities

This section presents the characterization of the OLC evaluated in this paper, their participants, moderators and activity levels. Such characterization is important to effectively recognize the communities investigated as learning communities.

[5]http://www.reddit.com/r/learnprogramming
[6]http://www.reddit.com/r/MachineLearning

Table 2: Data extracted and analyzed (from 2019-Jan-01 to 2020-Dec-31).

	learningprogramming	***MachineLearning***
Discussion topics	69,447	22,124
Unique active users	95,335	35,702
Replies or interactions	442,243	152,625

4.2.1 The role of moderators and community guidelines

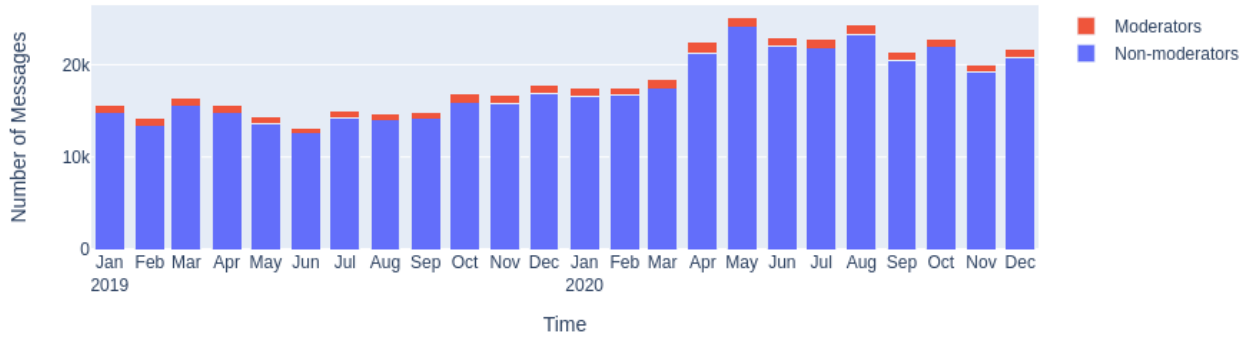

Figure 3: Number of messages posted monthly by moderators and non-moderators of ***subreddit learnprogramming***. Descriptive statistics: **Monthly messages** - Mean: 18410.16; Median: 17521; Std. Dev.: 3632.24. **Monthly messages from moderators** - Mean: 869.12; Median: 855; Std. Dev.: 144.95..

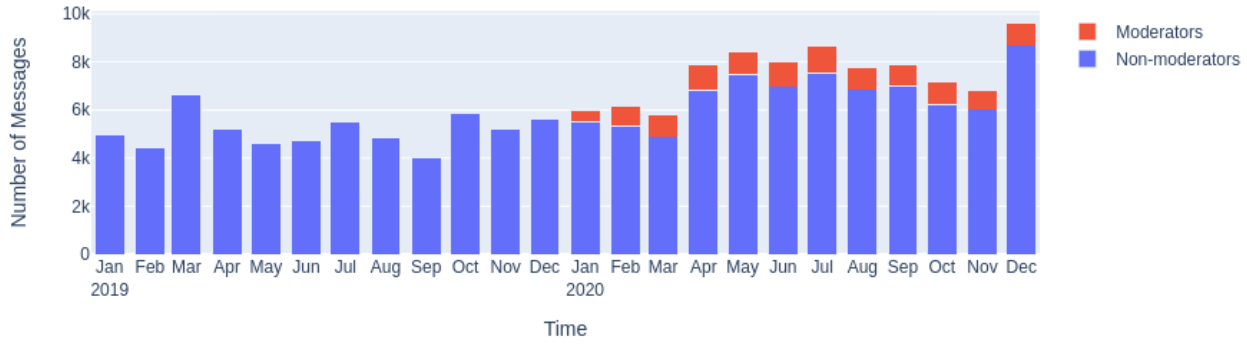

Figure 4: Number of messages posted monthly by moderators and non-moderators of ***subreddit MachineLearning***. Descriptive statistics: **Monthly messages** - Mean: 6331.37; Median: 5932; Std. Dev.: 1530.09. **Monthly messages from moderators** - Mean: 460.25; Median: 254; Std. Dev.: 465.84..

The *subreddits learnprogramming* and *MachineLearning* present a handful of guidelines about posting and answering questions in the correct way. The messages must provide all contextual information in the form of a good description and include a descriptive short title. Community members are requested to check out older messages before posting a new question. They can also learn and help to improve the knowledge creation and sharing through moderating. Moderating in Reddit requires a reputation, measured above a certain number of *karma* points. Participants find beneficial the presence of community moderators that follows their activity and regularly intervenes either to guide them or to monitor the adherence to discussion topic (Haythornthwaite, 2018). At our evaluation snapshot, the *subreddits learnprogramming* and *MachineLearning* had eight and nine moderators, respectively. Fig. 3 and Fig 4 show the number of messages posted

monthly by moderators and non-moderators over time. The *subreddit learnprogramming* presented a consistent number of messages from moderators, ranging from 3.76% (December 2020) to 5.93% (February 2019) of total messages. The *subreddit MachineLearning* had few messages posted by moderators in 2019, around 0.37% of total messages. However, this rate increased to 12% percent in 2020.

4.2.2 Interaction network structure analysis

To quantifying the properties evident in the overall community member's interaction network, we have computed the SNA measures shown in Table 3 and Table 4. They reveal the structural aspects of the online community as it re-configures itself according to the messages exchanged by users. At a network level, we computed the diameter (largest distance between two nodes) and number of triangles (sets of three nodes, each of which is connected to each other). At an egocentric level, we have computed the measures degree, eccentricity (the maximum distance from a node to all others) and number of triangles. The network was built considering all unique active community users in the period under analysis. The interaction patterns are very similar across both *subreddits*. The large number of triangles combined with high diameter, node degree and node eccentricity corroborate the fact that user interactions are based mostly on the content of the discussion topics and comments, regardless of the users who generate them (Fraga, da Silva, & Murai, 2018). This behavior is in line with the arguments described by Haythornthwaite (2018), which argues that a key transition to a learning community entails going from personal information seeking to collective practices associated with culture of exchanging information. We also observe a large standard deviation (SD) in the distribution of number of triangles, representing a consequence of the large variation in the number of activities performed by different users.

Table 3: Measures from interaction network at a network level.

subreddit	Nodes (users)	Links (interactions)	Diameter	Triangles
lp	95,335	442,243	67	118,836
ML	35,702	152,625	82	43,565

Note: lp means *learnprogramming*; ML means *MachineLearning*

Table 4: Measures from interaction network at an egocentric level.

subreddit	**Degree**			**Eccentricity**			**Triangles**		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
lp	9.32	3	78.36	10.11	10	1.26	3.76	0	109.68
ML	8.55	2	67.03	8.98	9	1.14	3.65	0	49.52

Note: lp means *learnprogramming*; ML means *MachineLearning*

4.3 Results

After extracting the data, the models described in section 3.2 were created. Such models are reported in more details in Appendix A. Thus, the next subsections describe the activity Visual and exploratory analysis and the activities related to Knowledge.

4.3.1 The SLAD and Visual and exploratory analysis

The SLAD depicts the general trend model of time series to reveal the evolution of the most significant structure and discourse measures, according section 3.2. Time series aim to capture the long run trend that can be fitted as linear regression of the time index. The SLAD has applied a method to standardize the trend models by removing the mean and scaling to unit variance. Thus, different trend measures are presented at the same scale. In addition, we have used moving averages of 60 days to have an effect of smoothing the original time series by eliminating random noise. The results are described as follow.

The Fig. 5 shows the temporal trend models that represent the behavior of structured measures over time in both *subreddits*. The *subreddit learnprogramming* has presented an increasing trend of measures number of participants and number of discussions. This scenario has produced a growth trend of measures related to amount of participation (size, width and score of discussions), although they were less intense. High levels of activity and participation in OLC are the key to the success of such environments. A learner as a member of these communities is both a producer and consumer of information. Thus, they have an important role in creating knowledge artifacts and sharing them to the their peers (Speily, Rezvanian, Ghasemzadeh, Saghiri, & Vahidipour, 2020). The *subreddit MachineLearning* has presented a similar growth trend of measures number of participants and number of discussions. However, these increasing trends have not produced a greater amount of participation, because the measures size, width and score of discussions have presented consistent decreasing trends over time. The Figure 6, described as follow, could help to clarify this scenario.

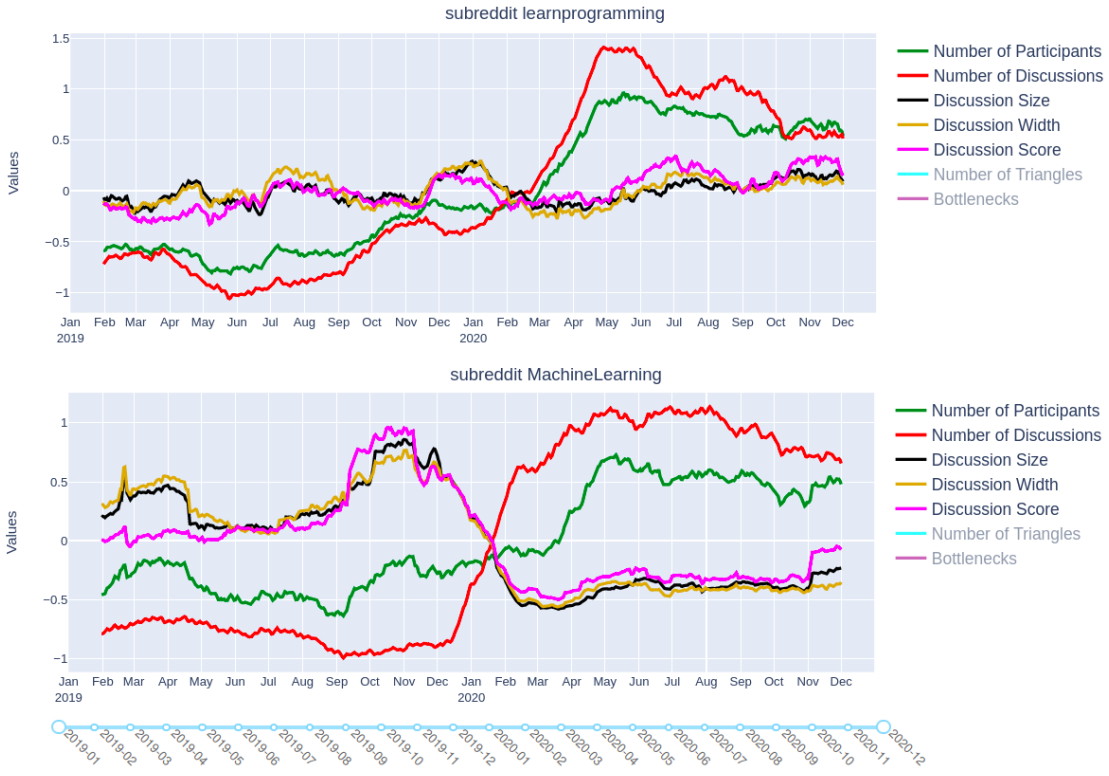

Figure 5: Temporal trend models of structured measures.

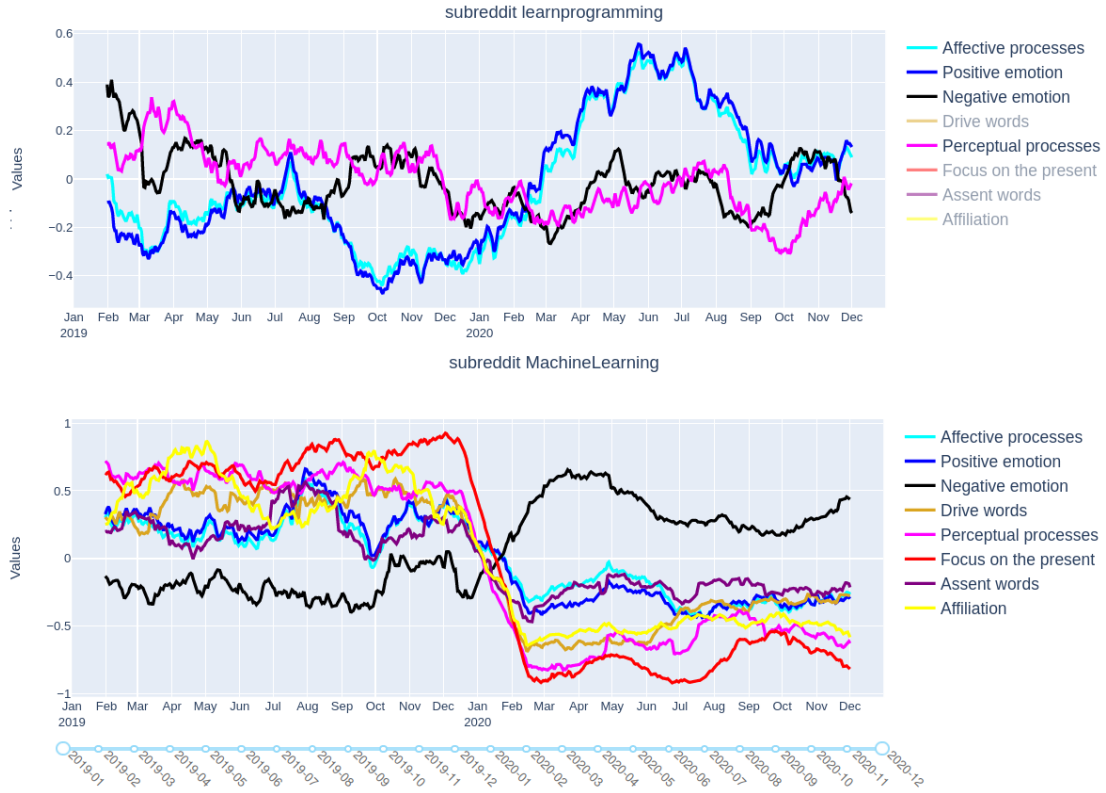

Figure 6: Temporal trend models of discourse measures.

Fig. 6 shows the temporal trend models that represent the behavior of discourse measures. In *subreddit learnprogramming* we have emphasized the measures related to emotions, affective and perceptual processes, because the other measures have not exhibited significant increasing or decreasing trends. The measures positive emotion (words like love, good and nice) and affective processes (words like admire, interesting and laugh) have presented increasing trends, whilst the measures perceptual processes (words like look, hear, feeling) and negative emotion (words like angry, bad and nasty) have exhibited smooth decreasing trends over time. On the other hand, in *subreddit MachineLearning* all discourse measures have presented decreasing trends, except measure negative emotion which has exhibited increasing trend over time. Emotions play a critical role during the learning process and problem solving with educational technologies (Azevedo et al., 2017). Experimental research has assumed that positive emotions facilitate the use of flexible and creative learning strategies; whilst activating negative emotions leads to more rigid strategies like simple rehearsal and superficial ways of processing information (Pekrun, 2006). In addition, studies have shown that comments with the prevalence of negative emotion in online discussions are less likely to receive responses from other participants, and resulted in lower prestige for their authors (Zou et al., 2021).

In order to investigate the behavior of positive and negative emotion with more details, we have shortened the analysis interval from Oct-2019 to Dec-2020. The result is shown in Fig. 7. The *subreddit learnprogramming* has presented a positive emotion increasing trend most of the time. However, *subreddit MachineLearning* has exhibited higher negative emotion increasing trend near from Feb-2020 to the end of period under analysis. Thus, our SLAD has helped to

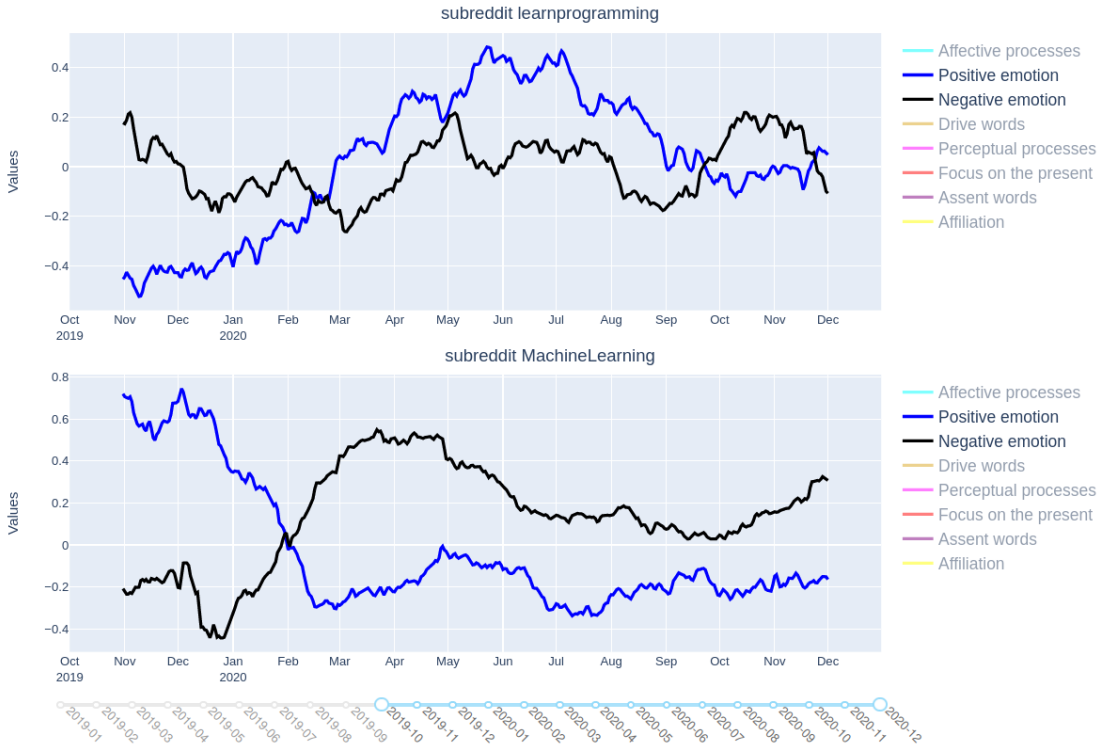

Figure 7: Behavior of positive and negative emotion in a shortened time interval.

identify a specific context that could lead to lower levels of activity and participation in *subreddit MachineLearning*. Consequently, we have investigated whether the negative emotion prevalence could influence the occurrence of network effects and behavioral patterns traditionally related to learning, as described in next subsection.

4.3.2 Knowledge creation about online social learning

Based on the insight that negative emotion prevalence could influence the occurrence of network effects related to learning, we have formulated the following hypothesis:

- **H1**: Do periods with prevalence of negative emotion produce, less frequently, structural network effects associated with online social learning?

With the aim of testing such hypothesis, we have created an ERG model to investigate whether negative emotion could influence the occurrence of the following network effects: reciprocity, simple connectivity, popularity and transitivity. The result is shown in Table 5. The *subreddit learnprogramming* has presented higher significant estimates for all network effects. This means that such effects occurred less frequently in *subreddit MachineLearning*. Thus, we accept H1 because the prevalence of negative emotion has produced less network effects associated with learning than expected by chance. Many studies in online contexts found that the emotional representation of information can result in more attention and participation (Xiong, Feng, & Tang, 2020). Our findings are in line with other studies that describe the role of positive and negative emotions in learning settings (Azevedo et al., 2017; Zou et al., 2021).

Table 5: Results of ERG model for network effects.

subreddit learnprogramming			***subreddit MachineLearning***		
Estimates	**Coefficient**	**SE**	**Estimates**	**Coefficient**	**SE**
Baseline (edges)	-12.1109***	0.0092	Baseline (edges)	-11.1476***	0.0087
Reciprocity	8.0950***	0.3342	Reciprocity	6.2750***	0.0805
Simple connectivity	0.2756***	0.0077	Simple connectivity	0.0367*	0.0001
Popularity	1.1508***	0.0751	Popularity	0.8606***	0.3337
Transitivity	14.8768***	0.8773	Transitivity	7.3114***	0.2025

Notes: *** means *p-value*<0.001; * means *p-value*<0.05; SE means Standard Error.

5 Conclusions

This paper presented SLIM: a process that combines SNA and discourse analysis to provide valuable information about student interaction and discourse style in large online learning communities, an underrepresented learning environment in the educational research. SLIM can be applied to asynchronous text-based discussions where users can rate each other's responses and discussion topics. We have applied it to two large *subreddits* from online news sharing site Reddit: *learnprogramming* and *MachineLearning*.

The contributions to the broader understanding of social learning within large OLC refer to the identification of a set of quantitative measures and machine learning models that outline the evolution of SL indicators over time. The combination of trend models visualization and exploratory educational data analysis was able to point out that the prevalence of negative emotion could explain the decreasing participation in online communities. We confirmed this claim by fitting an ERG model that evaluated network effects associated with the amount of participation. The results showed that the period with negative emotion increasing trend produced such effects less frequently than expected by chance.

The SLIM process offers means to analyze large scale educational data in informal environments; however, there are still some limitations. Considering discourse analysis, we have used a limited set of LIWC measures in the clustering method. Though we have found empirical evidences that pointed out such measures as important to recognize social presence in OLC, additional measures could be investigated. In future works, we intend to improve the discourse analysis and apply SLIM in other online communities, with the purpose of investigating how the expression of positive or negative sentiment and intense emotional states in the messages exchanged by the users may affect the level of participation or the discourse style in informal learning settings.

6 Extended Awarded Article

This publication is an extended version of the best paper award winner in Brazilian Simposio of Informatics on Education (SBIE - 2021), entitled "Analyzing learners' behavior and discourse within large online communities: a Social Learning Analytics Dashboard", DOI: 10.5753/sbie.2021.217468

References

Azevedo, R., Taub, M., Mudrick, N. V., Millar, G. C., Bradbury, A. E., & Price, M. J. (2017). Using data visualizations to foster emotion regulation during self-regulated learning with advanced learning technologies. In *Informational environments* (pp. 225–247). Springer. doi: 10.1007/978-3-319-64274-1_10 [GS Search]

Becheru, A., Calota, A., & Popescu, E. (2018). Analyzing students' collaboration patterns in a social learning environment using studentviz platform. *Smart Learning Environments*, *5*(1), 1–18. doi: 10.1186/s40561-018-0063-0 [GS Search]

Chatti, M. A., Muslim, A., & Schroeder, U. (2017). Toward an open learning analytics ecosystem. In B. K. Daniel (Ed.), *Big data and learning analytics in higher education* (pp. 195–219). Springer International Publishing. doi: 10.1007/978-3-319-06520-5_12 [GS Search]

Chen, B., Chang, Y.-h., Ouyang, F., & Zhou, W. (2018). Fostering student engagement in online discussion through social learning analytics. *The Internet and Higher Education*, *37*(January), 21–30. doi: 10.1016/j.iheduc.2017.12.002 [GS Search]

Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on learning technologies*, *7*(3), 246–259. doi: 10.1109/TLT.2013.2296520 [GS Search]

Corbi, A., & Burgos, D. (2020). How to integrate formal and informal settings in massive open online courses through a transgenic learning approach. In *Radical solutions and learning analytics* (pp. 173–191). Springer. doi: 10.1007/978-981-15-4526-9_11 [GS Search]

Czerkawski, B. C. (2016). Blending formal and informal learning networks for online learning. *Int. Review of Research in Open and Distributed Learning*, *17*(3), 138–156. doi: 10.19173/irrodl.v17i3.2344 [GS Search]

Dascalu, M., McNamara, D. S., Trausan-Matu, S., & Allen, L. K. (2018). Cohesion network analysis of cscl participation. *Behavior Research Methods*, *50*(2), 604–619. doi: 10.3758/s13428-017-0888-4 [GS Search]

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers and Education*, *46*(1), 6–28. doi: 10.1016/j.compedu.2005.04.005 [GS Search]

De Laat, M., & Prinsen, F. (2014). Social Learning Analytics: Navigating the Changing Settings of Higher Education. *Research & Practice in Assessment*, *9*(1), 51–60. [GS Search]

Ferguson, R., & Shum, S. B. (2012). Social learning analytics: five approaches. In *Proc. of 2nd int. conf. on learning analytics & knowledge* (pp. 23–33). doi: 10.1145/2330601.2330616 [GS Search]

Ferreira, M., Rolim, V., Mello, R. F., Lins, R. D., Chen, G., & Gasevic, D. (2020). Towards automatic content analysis of social presence in transcripts of online discussions. In *Proc. of the 10th int. conf. on learning analytics and knowledge.* doi: 10.1145/3375462.3375495 [GS Search]

Fincham, E., Gasevic, D., & Pardo, A. (2018). From social ties to network processes: Do tie definitions matter? *Journal of Learning Analytics*, *5*(2), 9–28. doi: 10.18608/jla.2018.52.2 [GS Search]

Fraga, B. S., da Silva, A. P. C., & Murai, F. (2018). Online social networks in health care: a study of mental disorders on reddit. In *2018 ieee/wic/acm int. conf. on web intelligence (wi)* (pp. 568–573). doi: 10.1109/WI.2018.00-36 [GS Search]

Galanis, N., Mayol, E., Alier, M., & García-Peñalvo, F. J. (2016). Supporting, evaluating and validating informal learning. a social approach. *Computers in Human Behavior*, *55*, 596–603. doi: 10.1016/j.chb.2015.08.005 [GS Search]

Galvin, S., & Greenhow, C. (2020). Educational networking: A novel discipline for improved k-12 learning based on social networks. In *Educational networking* (pp. 3–41). Springer. doi: 10.1007/978-3-030-29973-6_1 [GS Search]

Garrison, D. R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. *Internet and Higher Education*, *13*(1-2), 5–9. doi: 10.1016/j.iheduc.2009.10.003 [GS Search]

Gašević, D., Joksimović, S., Eagan, B. R., & Shaffer, D. W. (2019). Sens: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior*, *92*, 562–577. doi: 10.1016/j.chb.2018.07.003 [GS Search]

Greenhow, C., Gibbins, T., & Menzer, M. M. (2015). Re-thinking scientific literacy out-of-school: Arguing science issues in a niche Facebook application. *Computers in Human Behavior*, *53*, 593–604. doi: doi.org/10.1016/j.chb.2015.06.031 [GS Search]

Gruzd, A., Paulin, D., & Haythornthwaite, C. (2016). Analyzing social media and learning through content and social network analysis: A faceted methodological approach. *Journal of Learning Analytics*, *3*(3), 46–71. doi: 10.18608/jla.2016.33.4 [GS Search]

Haythornthwaite, C. (2018). Learning , connectivity and networks. *Information and Learning Science*. doi: 10.1108/ILS-06-2018-0052 [GS Search]

Haythornthwaite, C., de Laat, M., & Schreurs, B. (2016). A social network analytic perspective on e-learning. In *The sage handbook of e-learning research* (pp. 251–269). London: SAGE Publications. [GS Search]

Haythornthwaite, C., Kumar, P., Gruzd, A., Gilbert, S., Esteve del Valle, M., & Paulin, D. (2018). Learning in the wild: coding for learning and practice on reddit. *Learning, media and technology*, *43*(3), 219–235. doi: 10.1080/17439884.2018.1498356 [GS Search]

Hudgins, W., Lynch, M., Schmal, A., Sikka, H., Swenson, M., & Joyner, D. A. (2020). Informal Learning Communities: The Other Massive Open Online 'C'. *L@S 2020 - Proceedings of the 7th ACM Conference on Learning @ Scale*, 91–101. doi: 10.1145/3386527.3405926 [GS Search]

Jan, S. K. (2019). Investigating virtual communities of practice with social network analysis: guidelines from a systematic review of research. *Int. Journal of Web Based Communities*, *15*(1), 25. doi: 10.1504/ijwbc.2019.098697 [GS Search]

Jan, S. K., & Vlachopoulos, P. (2018). Influence of learning design of the formation of online communities of learning. *International Review of Research in Open and Distributed Learning*, *19*(4). doi: 10.19173/irrodl.v19i4.3620 [GS Search]

Jan, S. K., & Vlachopoulos, P. (2019). Social network analysis: A framework for identifying communities in higher education online learning. *Technology, Knowledge and Learning*, *24*(4), 621–639. doi: 10.1007/s10758-018-9375-y [GS Search]

Joksimovic, S., Gasevic, D., Kovanovic, V., Adesope, O., & Hatala, M. (2014). Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions. *Internet and Higher Education*, *22*, 1–10. doi: 10.1016/j.iheduc.2014.03.001 [GS Search]

Joksimovic, S., Gasevic, D., Kovanovic, V., Riecke, B. E., & Hatala, M. (2015). Social presence in online discussions as a process predictor of academic performance. *Journal of Computer*

Assisted Learning, *31*(6), 638–654. doi: 10.1111/jcal.12107 [GS Search]

Joksimovic, S., Jovanovic, J., Kovanovic, V., Gasevic, D., Milikic, N., Zouaq, A., & van Staalduinen, J. P. (2019). Comprehensive analysis of discussion forum participation: From speech acts to discussion dynamics and course outcomes. *IEEE Transactions on Learning Technologies*, *13*(1), 38–51. doi: 10.1109/TLT.2019.2916808 [GS Search]

Kent, C., Rechavi, A., & Rafaeli, S. (2019). Networked learning analytics: A theoretically informed methodology for analytics of collaborative learning. In *Learning in a networked society* (pp. 145–175). Springer. doi: 10.1007/978-3-030-14610-8_9 [GS Search]

Kim, D., Yoon, M., Jo, I.-H., & Branch, R. M. (2018). Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in south korea. *Computers & Education*, *127*, 233–251. doi: 10.1016/j.compedu.2018.08.023 [GS Search]

Kovanović, V., Joksimović, S., Poquet, O., Hennis, T., Čukić, I., De Vries, P., ... Gašević, D. (2018). Exploring communities of inquiry in massive open online courses. *Computers & Education*, *119*, 44–58. doi: 10.1016/j.compedu.2017.11.010 [GS Search]

Lin, Y., Yu, R., & Dowell, N. (2020). Liwcs the same, not the same: Gendered linguistic signals of performance and experience in online stem courses. In *Int. conf. on artificial intelligence in education* (pp. 333–345). doi: 10.1007/978-3-030-52237-7_27 [GS Search]

Mamas, C., Bjorklund Jr, P., Daly, A. J., & Moukarzel, S. (2020). Friendship and support networks among students with disabilities in middle school. *International Journal of Educational Research*, *103*, 101608. doi: 10.1016/j.ijer.2020.101608 [GS Search]

Nistor, N., Dascalu, M., Tarnai, C., & Trausan-Matu, S. (2020). Predicting newcomer integration in online learning communities: Automated dialog assessment in blogger communities. *Computers in Human Behavior*, *105*(September 2019), 106202. doi: 10.1016/j.chb.2019.106202 [GS Search]

Nistor, N., Derntl, M., & Klamma, R. (2015). Learning Analytics : Trends and Issues of the Empirical Research of the Years 2011 – 2014. *Design for Teaching and Learning in a Networked World*, *4*, 453–459. doi: 10.1007/978-3-319-24258-3 [GS Search]

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, *18*(4), 315–341. doi: 10.1007/s10648-006-9029-9 [GS Search]

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of liwc2015.
[GS Search]

Pesare, E., Roselli, T., & Rossano, V. (2016). Visualizing student engagement in e-learning environment. In *Proc. 22th int. conference on distributed multimedia systems* (pp. 26–33). doi: 10.18293/DMS2016-028 [GS Search]

Rezaei, M. S., Bobarshad, H., & Badie, K. (2019). Toward enhancing collaborative learning groups formation in q&a website using tag-based next questions predictor. *Int. Journal of Technology Enhanced Learning*, *11*(4), 441–462. doi: 10.1504/IJTEL.2019.102570 [GS Search]

Rourke, L., Anderson, T., Garrison, R., & Archer, W. (2001). Assessing social presence in asynchronous text-based computer. *Journal of Distance Education*, *14*, 1–18. [GS Search]

Schreurs, B., & De Laat, M. (2014). The Network Awareness Tool: A web 2.0 tool to visualize informal networked learning in organizations. *Computers in Human Behavior*, *37*(1), 385–

394. doi: 10.1016/j.chb.2014.04.034 [GS Search]

Scott, J., & Carrington, P. J. (2011). *The sage handbook of social network analysis*. SAGE publications. [GS Search]

Shum, S. B., & Ferguson, R. (2012). Social learning analytics. *Journal of educational technology & society*, *15*(3), 3–26. [GS Search]

Silva, R. F., Gimenes, I. M. S., & Maldonado, J. C. (2020). An Approach for Assessing Large Online Communities in Informal Learning Environments. In *Anais do xxxi simpósio brasileiro de informática na educação* (pp. 642–651). doi: 10.5753/cbie.sbie.2020.642 [GS Search]

Speily, O. R. B., Rezvanian, A., Ghasemzadeh, A., Saghiri, A. M., & Vahidipour, S. M. (2020). Lurkers versus posters: Investigation of the participation behaviors in online learning communities. In *Educational networking* (pp. 269–298). Springer. doi: 10.1007/978-3-030-29973-6_8 [GS Search]

Swiecki, Z., & Shaffer, D. W. (2020). ISENS: An integrated approach to combining epistemic and social network analyses. In *Proc. of the 10th int. conf. on learning analytics & knowledge* (pp. 305–313). doi: 10.1145/3375462.3375505 [GS Search]

Toikkanen, T., & Lipponen, L. (2011). The applicability of social network analysis to the study of networked learning. *Interactive Learning Environments*, *19*(4), 365–379. doi: 10.1080/10494820903281999 [GS Search]

Wenger, E., Trayner, B., & De Laat, M. (2011). Promoting and assessing value creation in communities and networks: a conceptual framework. , *18*(August), 1–60. [GS Search]

Weninger, T. (2014). An exploration of submissions and discussions in social news: Mining collective intelligence of reddit. *Social Network Analysis and Mining*, *4*(1), 173–192. doi: 10.1007/s13278-014-0173-9 [GS Search]

Xiong, J., Feng, X., & Tang, Z. (2020). Understanding user-to-user interaction on government microblogs: An exponential random graph model with the homophily and emotional effect. *Information Processing & Management*, *57*(4), 102229. doi: 10.1016/j.ipm.2020.102229 [GS Search]

Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise Networks in Online Communities: Structure and Algorithms. In *International world wide web conference* (pp. 221–230). doi: 10.1145/1242572.1242603 [GS Search]

Zhu, M., Herring, S. C., & Bonk, C. J. (2019). Exploring presence in online learning through three forms of computer-mediated discourse analysis. *Distance Education*, *40*(2), 205–225. doi: 10.1080/01587919.2019.1600365 [GS Search]

Zou, W., Hu, X., Pan, Z., Li, C., Cai, Y., & Liu, M. (2021). Exploring the relationship between social presence and learners' prestige in mooc discussion forums using automated content analysis and social network analysis. *Computers in Human Behavior*, *115*, 106582. doi: 10.1016/j.chb.2020.106582 [GS Search]

Appendix A

The Appendix A reports the machine learning models fitted with the purpose of identifying the most relevant structure and discourse measures, described as follow.

- Table 6 presents the linear regression model for identifying the most significant structured measures.
- We have used the results of algorithm *Kmeans* to support discourse analysis so that discussion topics with similar discourse styles could be grouped in clusters. Thus, we have investigated which cluster is associated with the best structured measures, according to results presented in Table 6. Analyzing Table 7, we can investigate which cluster has the highest average for all measures (in bold). Thus, we argue that such cluster has the most valuable discussions for learners. Cluster 2 presented the highest average for all measures in *subreddit learnprogramming*, except number of bottlenecks. The reason why this measure has not the highest average in Cluster 2 needs more investigation. Cluster 1 presented the highest average for all measures in *subreddit MachineLearning*.
- In order to identify the most relevant discourse measures, we have analyzed the normalized coordinates of the cluster *centroids*: a value in the interval [0, 1] which indicates the cluster average value for each measure. The most significant ones are the closest to one. The results depicted in Fig. 8 and Fig 9 were quite similar for both *subreddits*. They presented seven discourse measures which can be considered important for identifying a discourse style related to most valuable discussions for learners: *liwc.affect*, *liwc.negemo*, *liwc.percept*, *liwc.drives*, *liwc.focuspresent*, *liwc.affiliation* and *liwc.assent*.

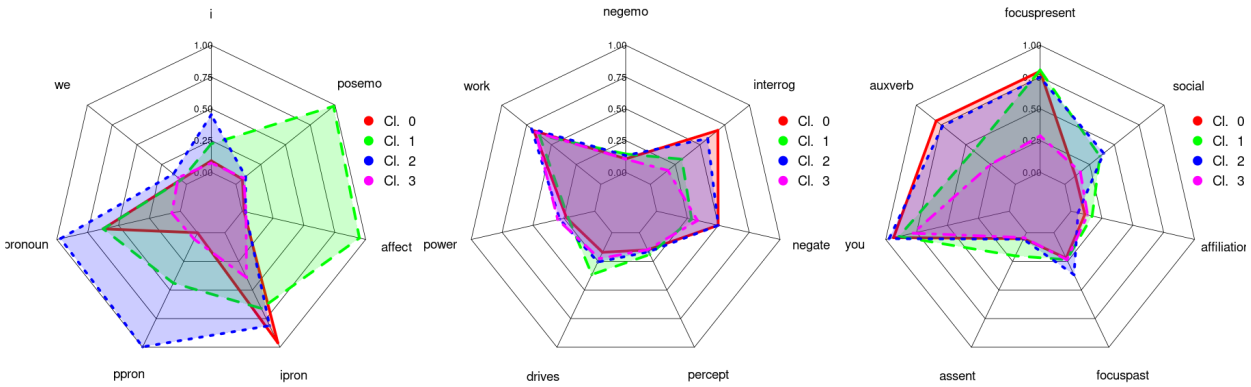

Figure 8: Results of *Kmeans* centroids from model for identifying the most relevant discourse measures - *subreddit learnprogramming*. **Cluster 1** presented the most valuable discussions for users. Their most significant measures were: *posemo*, *affect*, *negemo*, *percept*, *drives*, *focuspresent*, *affiliation* and *assent*..

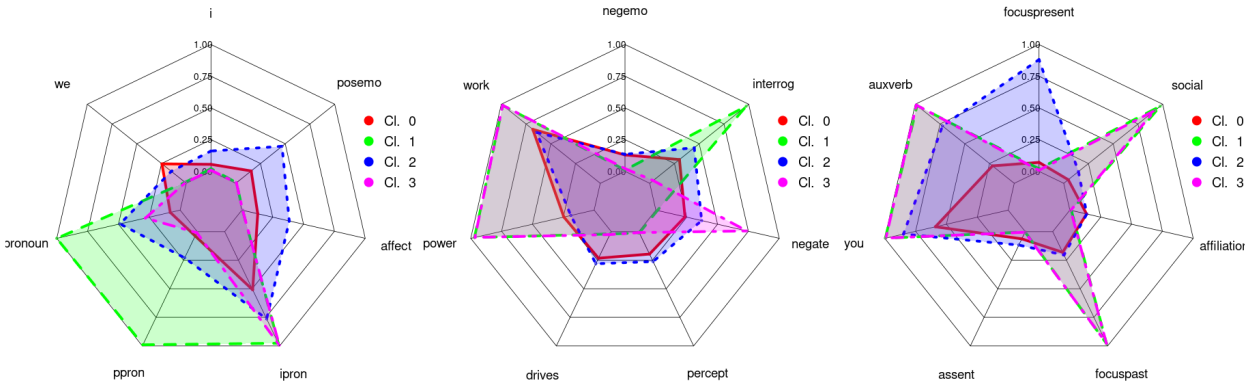

Figure 9: Results of *Kmeans* centroids from model for identifying the most relevant discourse measures - *subreddit MachineLearning*. **Cluster 2** presented the most valuable discussions for users. Their most significant measures were: *i*, *posemo*, *affect*, *negemo*, *percept*, *drives*, *focuspresent*, *affiliation* and *assent*..

Table 6: Linear regression model for identifying the most significant structured measures.

subreddit learnprogramming				
Structural measures			**Network measures**	
1. Discussion Score	*1.00*		9. Density	-0.18
2. Participants	**0.87*****		10. Reciprocity	-0.02
3. Size	**0.84*****		11. Components	0.00
4. Time 1st reply	0.00		12. Avg. shortest path	-0.03
5. Width	**0.79*****		13. Clustering coeff.	0.01
6. Depth	0.25**		14. Diameter	0.37***
7. Disc. intensity	0.03		**15. Triangles**	**0.64*****
8. Disc. duration	0.23		**16. Bottlenecks**	**0.56*****
subreddit MachineLearning				
Structural measures			**Network measures**	
1. Discussion Score	*1.00*		9. Density	-0.26
2. Participants	**0.85*****		10. Reciprocity	0.05
3. Size	**0.81****		11. Components	0.00
4. Time 1st reply	0.00		12. Avg. shortest path	0.04
5. Width	**0.70*****		13. Clustering coeff.	0.05
6. Depth	0.45**		**14. Diameter**	**0.51*****
7. Disc. intensity	0.20		**15. Triangles**	**0.52*****
8. Disc. duration	0.21		**16. Bottlenecks**	**0.78*****

Notes: *** means *p-value*<0.001; ** means *p-value*<0.01
Adjusted R-squared for *learnprogramming*: 0.8200
Adjusted R-squared for *MachineLearning*: 0.7980

Table 7: Results of clustering algorithm applied to discussion topics.

subreddit learnprogramming				
Cluster	Cluster 0	**Cluster 1**	Cluster 2	Cluster 3
Size of Cluster	18,524	12,814	15,123	16,495
% of Total	29.43%	20.35%	24.02%	26.20%
Score	42.58	**85.18**	36.08	14.75
Participants	5.23	**5.85**	4.74	3.69
Size	7.22	**8.63**	6.75	3.88
Width	3.51	**4.17**	3.28	2.50
Diameter	2.35	**2.40**	2.29	1.60
Triangles	0.14	**0.15**	0.10	0.05
Bottlenecks	1.89	2.23	2.05	**2.30**
subreddit MachineLearning				
Cluster	Cluster 0	Cluster 1	**Cluster 2**	Cluster 3
Size of Cluster	5411	3118	5636	3265
% of Total	31.04%	17.89%	32.34%	18.73%
Score	71.58	10.25	**116.77**	48.78
Participants	6.34	2.36	**8.54**	3.82
Size	8.32	1.66	**13.04**	4.46
Width	3.91	1.22	**5.32**	1.92
Diameter	2.17	1.10	**2.94**	1.33
Triangles	0.30	0.02	**0.44**	0.39
Bottlenecks	2.45	2.02	**2.50**	2.27

Appendix B

This appendix describes the common interpretation of SNA egocentric measures in social learning investigation contexts.

Table 8: Common interpretation of SNA egocentric measures in social learning investigation contexts.

Measure	Interpretation
Degree centrality	It refers to the number of connections of a node. In a directed network, the in-degree centrality measures the incoming edges and the out-degree centrality represents outgoing edges. The centrality of a node has also been linked to power, influence, prestige, and performance. Out-degree centrality has been used as indicator of influence and prestige. In-degree centrality has been associated with popularity. (Jan & Vlachopoulos, 2019).
Betweenness centrality	It is commonly used to identify actors considered experts, actors that mediate the flow of information or connect different groups present in the learning network. (Gruzd et al., 2016).
Pagerank and Authority	They are commonly used to recognize experts in the learning network, thus identifying actors who help other members, or are sought after for providing the best answers. (Zhang et al., 2007).
Density	The density value tends to decrease as the network size increases, as it becomes more difficult to connect all actors. Its value is evidence of the connectivity index, which reflects the ease that information can reach network actors. (Haythornthwaite et al., 2016).
Reciprocity	A high number of reciprocal connections has the potential to support interactively the collaborative process. Therefore, this measure is interpreted as an indicator of knowledge mutual exchange, construction and negotiation of meanings. (Gašević et al., 2019).
Diameter	High diameter value suggests that shared information can reach more distant actors in the learning network. The change in diameter reflects on the possibilities of interaction between the actors, and affects the collective memory of the group (Kent et al., 2019).