UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Um processo para a avaliação da aprendizagem social *online* em ambientes informais

Rogério Ferreira da Silva

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Rogério Ferreira da Silva

Um processo para a avaliação da aprendizagem social *online* em ambientes informais

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. José Carlos Maldonado

USP – São Carlos
Novembro de 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

d586p

da Silva, Rogério Ferreira
Um processo para a avaliação da aprendizagem social online em ambientes informais. / Rogério Ferreira da Silva; orientador José Carlos Maldonado. -- São Carlos, 2021.
153 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Aprendizagem social. 2. Ambientes informais de aprendizagem. 3. Comunidades online. 4. Avaliação. 5. Processo. I. Maldonado, José Carlos, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Rogério Ferreira da Silva

A process for assessing online social learning in informal settings

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. José Carlos Maldonado

USP – São Carlos
November 2021

AGRADECIMENTOS

Agradeço ao professor José Carlos Maldonado, pela orientação e confiança ao longo desta jornada.

À professora Itana Maria de Souza Gimenes, pela coorientação, confiança, paciência, amizade, conversas, ensinamentos e discussões realizadas ao longo dos últimos quatro anos, que agregaram um valor imensurável a este trabalho.

Ao Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, pela formação, sobretudo aos funcionários, professores e colegas de pesquisa, especialmente do Laboratório de Engenharia de Software (Labes).

Ao professor Fabrício Benevenuto de Souza, do Departamento de Computação da Universidade Federal de Minas Gerais, e aos colegas de pesquisa do Laboratório de Computação Social (Locus), que dividiram inúmeros conhecimentos que foram importantes ao desenvolvimento deste trabalho.

À Universidade Federal do Paraná e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio.

Por fim, esta tese é fruto de um trabalho conjunto, portanto, agradeço a todos e todas que contribuíram de alguma forma.

RESUMO

SILVA, R. F. **Um processo para a avaliação da aprendizagem social *online* em ambientes informais**. 2021. 153 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Os ambientes informais de aprendizagem, como as redes sociais, aplicativos de mensagens e grandes comunidades *online*, fazem parte de uma ecologia de contextos tecnológicos em que a aprendizagem ocorre de forma cada vez mais distribuída e em interação dinâmica com diferentes públicos. A área de pesquisa educacional evidencia que tais ambientes são essenciais para despertar e manter o interesse e a participação dos estudantes em práticas colaborativas, um conceito fundamental para a Educação do Século XXI. Assim, existe um interesse crescente em analisar como a aprendizagem social ocorre em ambientes informais. No entanto, investigar a aquisição de conhecimento nesses ambientes é um processo complexo que exige uma abordagem analítica para compreender as várias dimensões do discurso dos participantes e das suas interações sociais. Além disso, existem inúmeras dificuldades metodológicas inerentes a esta investigação, como a natureza não estruturada das atividades de ensino e aprendizagem, a ausência do papel exercido por professores e/ou tutores e a inexistência de processos formais de avaliação. Assim, faltam métodos e não há consenso sobre como analisar as atividades de aprendizagem em ambientes informais. Neste contexto, o presente trabalho de doutorado define um processo intitulado SLIM (*Process for assessing online **S**ocial **L**earning within online communities in **I**nformal environ**M**ents*). SLIM foi concebido a partir das lacunas de pesquisa sobre a investigação de ambiente informais evidenciadas na área de pesquisa educacional. Seu objetivo é estabelecer um conjunto de atividades e diretrizes para guiar a análise de dados produzidos pelos participantes de grandes comunidades *online*. Um estudo de caso para validar a aplicação do SLIM foi conduzido com base na análise dos dados de 131.037 participantes e 594.868 comentários de duas comunidades *online* da rede social Reddit. Os resultados evidenciaram a efetividade do processo na identificação de padrões comportamentais associados à aprendizagem. As principais contribuições do trabalho referem-se à confirmação de hipóteses que ampliam a compreensão sobre como a aprendizagem social *online* se realiza nos ambientes informais. Tais hipóteses evidenciam a troca de informações entre participantes com diferentes níveis de conhecimento, a importância do nível de atividade para que um indivíduo seja reconhecido como especialista pelos seus pares e a influência das emoções positiva e negativa na quantidade de participação detectada entre os membros das comunidades.

Palavras-chave: aprendizagem social, ambientes informais de aprendizagem, comunidades *online*, avaliação, processo.

ABSTRACT

SILVA, R. F. **A process for assessing online social learning in informal settings**. 2021. 153 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Informal learning environments, such as social networks, messaging applications and large *online* communities, are part of an ecology of technological contexts in which distributed learning takes place. These environments involve different audiences in dynamic interactions. Educational research suggests that such environments are essential to sustain students' interest and participation in collaborative practices, a fundamental concept for 21st Century Education. Thus, there is an increasing interest in systematically analyzing social learning that takes place in informal environments. However, investigating the acquisition of knowledge in these environments is a complex process that requires an analytical approach to understand the various dimensions of participants' discourse and their social interactions. In addition, there are several methodological difficulties inherent to this investigation, such as the unstructured nature of teaching and learning activities, the absence of the role played by teachers or tutors, and the inexistence of formal evaluation processes. Consequently, there is a lack of methods and tools to analyze learning activities in informal environments. Thus, this work defines a process named SLIM (*Process for assessing online **S**ocial **L**earning within online communities in **I**nformal environ**M**ents*). SLIM is based on the research gaps described in the educational research. Its objective is to define a set of activities to guide the analysis of data produced by the participants of large online communities. A case study was conducted to validate the application of the SLIM based on data related to 131,037 participants and 594,868 comments from two online communities in social network Reddit. The results demonstrate the effectiveness of the process in identifying behavioral patterns associated with learning. The main contributions of SLIM refer to the confirmation of hypotheses that broaden the understanding of how online social learning takes place in informal environments. Such hypotheses point out the information exchange between participants with different levels of knowledge, the importance of activity level for recognition of a member as an expert by their peers and the influence of positive and negative emotion in the amount of participation detected among community members.

Keywords: social learning, informal learning environments, online communities, assessment, process.

LISTA DE ILUSTRAÇÕES

Figura 1 – Principais características dos ambientes formais, não-formais e informais de aprendizagem . . . 32
Figura 2 – Representação esquemática de um tópico de discussão e suas estruturas de dados . . . 34
Figura 3 – Modelo de referência para LA . . . 36
Figura 4 – Modelo de processo analítico para LA com recursos de análise visual . . . 37
Figura 5 – Efeitos estruturais comumente investigados em métodos estatísticos de rede 46
Figura 6 – Interconexões das três presenças propostas pelo *framework* CoI . . . 51
Figura 7 – Fases do mapeamento sistemático . . . 58
Figura 8 – Bases teóricas referenciadas nos estudos primários . . . 60
Figura 9 – Métodos para análise de dados utilizados nos estudos primários . . . 60
Figura 10 – Descrição dos conceitos-chave do processo SLIM . . . 68
Figura 11 – Representação do processo SLIM por meio de um diagrama de atividades . 76
Figura 12 – Diagrama que apresenta o detalhamento das atividades relacionadas aos modelos . . . 80
Figura 13 – Submissão em que um usuário publica uma dúvida sobre a sintaxe da linguagem C . . . 91
Figura 14 – Evolução do número de submissões e comentários em cada comunidade . . 92
Figura 15 – Distribuição dos valores de Tamanho, Comprimento e Profundidade das submissões . . . 95
Figura 16 – Número de comentários publicados mensalmente por moderadores e não moderadores na comunidade *learnprogramming* . . . 97
Figura 17 – Número de comentários publicados mensalmente por moderadores e não moderadores na comunidade *MachineLearning* . . . 97
Figura 18 – Termos mais frequentes e os três tópicos mais relevantes para a comunidade *learnprogramming* . . . 102
Figura 19 – Termos mais frequentes e os três tópicos mais relevantes para a comunidade *MachineLearning* . . . 104
Figura 20 – Distribuição do *lifespan* em dias. . . . 105
Figura 21 – Distribuição da pontuação de *karma* e a quantidade de participantes em cada intervalo. . . . 106
Figura 22 – Resultado da aplicação do método *Elbow* . . . 117

Figura 23 – Resultado da avaliação do modelo de clusterização para a comunidade *learnprogramming* 117
Figura 24 – Resultado da avaliação do modelo de clusterização para a comunidade *MachineLearning* 118
Figura 25 – Resultado da análise dos centróides para o modelo do estilo de discurso na comunidade *learnprogramming* 119
Figura 26 – Resultado da análise dos centróides para o modelo do estilo de discurso na comunidade *MachineLearning* 120
Figura 27 – Componentes visuais do LAD 121
Figura 28 – Séries temporais das métricas estruturadas 123
Figura 29 – Séries temporais das métricas do discurso 124
Figura 30 – Comportamento das métricas de emoção positiva e emoção negativa no período entre outubro/2019 e dezembro/2020 125

LISTA DE QUADROS

Quadro 1 – Interpretação das métricas egocêntricas em contextos de investigação da aprendizagem 41
Quadro 2 – Interpretação das métricas a nível de rede em contextos de investigação da aprendizagem 43
Quadro 3 – Indicadores para avaliação do aspecto rede da aprendizagem social 50
Quadro 4 – As dez submissões mais bem avaliadas na comunidade *learnprogramming* . 99
Quadro 5 – As dez submissões mais bem avaliadas na comunidade *MachineLearning* . 100

LISTA DE TABELAS

Tabela 1 – Exemplos de atributos extraídos pelo LIWC 48
Tabela 2 – Intersecção entre os indicadores de presença social e *framework* de criação de valor 52
Tabela 3 – Estudos identificados na revisão de literatura 62
Tabela 4 – Lacunas de pesquisa identificadas nos estudos 64
Tabela 5 – Estatística descritiva para a quantidade de participantes envolvidos nos estudos analisados na revisão de literatura descrita neste capítulo e em três revisões sistemáticas 66
Tabela 6 – Métricas estruturadas 70
Tabela 7 – Métricas do discurso 71
Tabela 8 – Detalhes dos dados coletados nas comunidades entre 01/01/2019 e 31/12/2020 93
Tabela 9 – Estatística descritiva para o *score* e quantidades de palavras nas comunidades 94
Tabela 10 – Diversidade na interação entre participantes 95
Tabela 11 – Intensidade na interação entre participantes 96
Tabela 12 – Métricas estruturais de rede e egocêtricas 96
Tabela 13 – Coeficientes de correlação de *Pearson* para o modelo dos participantes especialistas 115
Tabela 14 – Coeficientes de correlação de *Pearson* para o modelo das discussões bem avaliadas 116
Tabela 15 – Valores médios das métricas estruturadas significativas para cada um dos *clusters* 119
Tabela 16 – Coeficientes do teste ADF para avaliação das séries temporais 122
Tabela 17 – Modelo ERG que investiga a interação entre participantes especialistas e regulares (**QP1**) 129
Tabela 18 – Modelo ERG que investiga a influência das métricas de centralidade na quantidade de participação (**QP2**) 131
Tabela 19 – Modelo ERG que investiga o nível de atividade e popularidade entre participantes especialistas e regulares (**QP3**) 132
Tabela 20 – Modelo ERG que investiga a menor frequência de efeitos estruturais de redes associados à aprendizagem social *online* na prevalência de emoção negativa (**QP4**) 133

LISTA DE ABREVIATURAS E SIGLAS

ADF	*Augmented Dickey-Fuller*
AOD	*Asynchronous text-based Online Discussions*
API	*Application Programming Interfaces*
CoI	*Community of Inquiry*
CSCL	*Computer-Supported Collaborative Learning*
DA	*Discourse Analysis*
DEA	*Data Envelopment Analysis*
DGWDSP	*Geometrically Weighted Dyadwise Shared Partner Distribution*
DGWESP	*Geometrically Weighted Edgewise Shared Partner Distribution*
DP	Desvio Padrão
ERG	*Exponential Random Graph*
ERGM	*Exponential Random Graph Model*
GWID	*Geometrically Weighted In-degree Distribution*
GWOD	*Geometrically Weighted Out-degree Distribution*
HITS	*Hyperlink-Induced Topic Search*
JSON	*JavaScript Object Notation*
LA	*Learning Analytics*
LAD	*Learning Analytics Dashboard*
LDA	*Latent Dirichlet Allocation*
LIWC	*Linguistic Inquiry and Word Count*
LMS	*Learning Management Systems*
LPP	Legitimação da Participação Periférica
LSA	*Latent Semantic Analysis*
ML	*Machine Learning*
MOOC	*Massive Open Online Courses*
MS	Mapeamento Sistemático
OLC	*Online Learning Communities*
Q&A	*Question and Answer*
QP	Questões de Pesquisa
RL	Revisão de Literatura
SL	*Social Learning*

SLA	*Social Learning Analytics*
SNA	*Social Network Analysis*
TIC	Tecnologias de Informação e Comunicação
URL	*Uniform Resource Locator*

SUMÁRIO

1 INTRODUÇÃO 23
1.1 Contextualização 23
1.2 Motivação 25
1.3 Objetivos e Questão de Pesquisa 26
1.4 Organização 27

2 FUNDAMENTAÇÃO TEÓRICA 29
2.1 Considerações iniciais 29
2.2 Aprendizagem social 29
2.2.1 Ambientes formais, não-formais e informais de aprendizagem 30
2.2.2 Definições e tipos de comunidades 33
2.3 *Learning Analytics* 35
2.4 *Social learning analytics* 38
2.4.1 Análise de Redes Sociais 39
2.4.1.1 Métricas de SNA no contexto da aprendizagem social 39
2.4.1.2 Análise da estrutura das discussões 42
2.4.1.3 Métodos estatísticos para redes 44
2.4.2 Análise do Discurso 46
2.5 *Frameworks* teóricos e aprendizagem social *online* 48
2.5.1 Criação de Valor 49
2.5.2 Comunidades de inquérito 49
2.6 Considerações finais 53

3 AVALIAÇÃO DA APRENDIZAGEM EM AMBIENTES INFORMAIS 55
3.1 Considerações iniciais 55
3.2 Ambientes informais e atividades de aprendizagem 55
3.3 Avaliação da aprendizagem em ambientes informais 56
3.4 Mapeamento sistemático sobre a avaliação da aprendizagem em comunidades *online* 57
3.4.1 QP1: uso de bases teóricas 59
3.4.2 QP2: estratégias para análise de dados 59
3.5 Revisão de literatura complementar ao MS 61
3.6 Considerações finais 66

4 SLIM: UM PROCESSO PARA AVALIAÇÃO DA APRENDIZAGEM SOCIAL EM AMBIENTES INFORMAIS . . . 67
4.1 Considerações iniciais . . . 67
4.2 Visão geral . . . 67
4.2.1 Dimensão "O quê?" . . . *68*
4.2.1.1 O espaço de métricas . . . *68*
4.2.1.2 Definição do formato dos dados de aprendizagem . . . *69*
4.2.2 Dimensão "Por quê?" . . . *72*
4.2.3 Dimensão "Quem?" . . . *73*
4.2.4 Dimensão "Como?" . . . *74*
4.2.4.1 Métodos de análise . . . *74*
4.2.4.2 Criação de modelos . . . *74*
4.3 Atividades do processo SLIM . . . 76
4.3.1 Atividades relacionadas aos dados . . . *77*
4.3.1.1 Extração de dados . . . *77*
4.3.1.2 Limpeza e transformação de dados . . . *78*
4.3.1.3 Cálculo das métricas e armazenamento dos dados . . . *78*
4.3.2 Atividades relacionadas aos modelos . . . *79*
4.3.2.1 Definição das métricas de avaliação dos modelos . . . *80*
4.3.2.2 Criação e treinamento dos modelos seccionais . . . *81*
4.3.2.3 Avaliação dos modelos seccionais . . . *83*
4.3.2.4 Disponibilização dos modelos seccionais . . . *83*
4.3.2.5 Criação e treinamento das séries temporais . . . *84*
4.3.2.6 Visualização das séries temporais . . . *85*
4.3.2.7 Avaliação das séries temporais . . . *86*
4.3.2.8 Disponibilização das séries temporais . . . *86*
4.3.3 Atividades relacionadas ao conhecimento . . . *87*
4.3.3.1 Formulação de hipóteses . . . *87*
4.3.3.2 Testes de hipóteses . . . *87*
4.3.3.3 Geração de conhecimento . . . *88*
4.4 Considerações finais . . . 88

5 CARACTERIZAÇÃO DAS COMUNIDADES DE APRENDIZAGEM 89
5.1 Considerações iniciais . . . 89
5.2 A rede social Reddit . . . 89
5.3 Características gerais das comunidades . . . 92
5.4 Análise das submissões e comentários . . . 93
5.4.1 Diversidade e intensidade das interações . . . *94*
5.4.2 O papel dos moderadores . . . *98*
5.4.3 Análise do conteúdo das discussões . . . *98*

5.4.3.1 Submissões de maior repercussão (ou mais bem avaliadas) 98
5.4.3.2 Mineração de textos e temas mais discutidos 100
5.5 Análise das atividades dos participantes 105
5.6 Considerações finais 107

6 UM ESTUDO DE CASO PARA O PROCESSO SLIM 109
6.1 Considerações iniciais 109
6.2 Um estudo de caso para o processo SLIM 109
6.2.1 Planejamento do estudo de caso 110
6.2.2 Preparação para a coleta de dados 111
6.2.3 Coleta de dados 112
6.2.4 Análise dos dados coletados 112
6.2.4.1 Análise dos dados quantitativos 112
6.2.4.2 Análise dos dados qualitativos 113
6.2.5 Redação de relatórios 113
6.3 SLIM: aplicação do estudo de caso 113
6.3.1 Execução da coleta de dados 114
6.3.2 Análise dos dados quantitativos 114
6.3.2.1 Criação e análise dos modelos seccionais 114
6.3.2.2 Criação e análise dos modelos temporais 120
6.3.3 Análise dos dados qualitativos 127
6.3.3.1 Formulação de hipóteses 127
6.3.3.2 Testes de hipóteses e geração de conhecimento 128
6.4 Considerações finais 134

7 CONCLUSÕES 135
7.1 Considerações iniciais 135
7.2 Contribuições da pesquisa 136
7.3 Limitações da pesquisa 138
7.4 Trabalhos futuros 138
7.5 Publicações resultantes 139

REFERÊNCIAS 141

CAPÍTULO

1

INTRODUÇÃO

Esta tese de doutorado define um processo para a avaliação de comunidades *online*, cujo objetivo é ampliar o entendimento e gerar conhecimento sobre como a aprendizagem social se realiza em ambientes informais de aprendizagem. Tal processo especifica um conjunto de atividades e diretrizes para guiar a análise dos dados produzidos pelos participantes destas comunidades, com o intuito de identificar um conjunto de condições e padrões comportamentais associados à aprendizagem social. Este capítulo está estruturado como descrito a seguir. A seção 1.1 faz uma contextualização sobre os ambientes informais de aprendizagem. A seção 1.2 descreve a motivação para o desenvolvimento deste trabalho de doutorado. A seção 1.3 apresenta os objetivos e a questão de pesquisa abordada no trabalho. Por fim, a seção 1.4 descreve a organização da tese de doutorado.

1.1 Contextualização

Educadores e pesquisadores concordam que a educação do século XXI deve incluir habilidades como pensamento crítico, comunicação, capacidade de pesquisar, sintetizar e compartilhar informações por meios tecnológicos, além da compreensão sobre como gerenciar e integrar diferentes fontes de informação no exercício profissional, incluindo a avaliação de tais fontes como confiáveis ou não confiáveis (AHN, 2013; DEDE, 2010; ROTHERHAM; WILLINGHAM, 2010). Essas habilidades estão em conformidade com o grau de importância que as práticas colaborativas na resolução de problemas vêm adquirindo. O trabalho na sociedade moderna baseada em conhecimento está sendo cada vez mais realizado por grupos de pessoas com competências e papéis complementares, que exigem um conjunto de habilidades cognitivas mais sofisticadas, de forma antagônica ao período em que indivíduos trabalhavam de forma mais isolada (AHN, 2013; KAROLY; PANIS, 2004).

As Tecnologias de Informação e Comunicação (TIC) desempenham um papel fundamental ao desenvolvimento de tais habilidades cognitivas. Inicialmente, a pesquisa na área das

tecnologias educacionais abrangia principalmente os sistemas de gestão da aprendizagem, do inglês *Learning Management Systems* (LMS), utilizados formalmente pelas instituições de ensino para, entre outras tarefas, auxiliar no monitoramento do desempenho dos estudantes (CHATTI; MUSLIM; SCHROEDER, 2017). Porém, nas últimas décadas, a discussão sobre tecnologia e aprendizagem passou a incluir ambientes abertos e em rede, como os Cursos Online Abertos e Massivos, do inglês *Massive Open Online Courses* (MOOC), aplicativos de mensagens, sites de redes sociais e comunidades de aprendizagem *online*. Assim, as práticas pedagógicas estão cada vez mais distribuídas no tempo, espaço e em diferentes plataformas tecnológicas (PINKARD, 2019).

Neste contexto, compreender as experiências educacionais que acontecem fora dos ambientes institucionais é fundamental para despertar e manter o interesse e a participação dos estudantes em práticas colaborativas (NACU *et al.*, 2019; WORSLEY; OCHOA, 2020). Uma abordagem promissora é considerar todo o ecossistema de aprendizagem, baseada na visão de que a aquisição de conhecimento ocorre de forma distribuída dentro de uma ecologia de contextos *online* e em interação dinâmica com diversos públicos (PENUEL; LEE; BEVAN, 2014; PINKARD, 2019). À medida que grandes volumes de dados gerados por meio da participação nesses espaços se tornam disponíveis, técnicas como *Learning Analytics* (LA) assumem um papel importante para ajudar a compreender os processos de aprendizagem cada vez mais complexos que acontecem nesse ecossistema de plataformas tecnológicas (CHATTI; MUSLIM; SCHROEDER, 2017; NACU *et al.*, 2019). LA refere-se ao desenvolvimento e a aplicação de métodos de ciência de dados para compreender as características e necessidades de sistemas educacionais e os fluxos de dados que eles geram. Seu propósito é aperfeiçoar os processos de aprendizagem e os ambientes tecnológicos em que eles ocorrem, permitindo a tomada de decisões informada por professores, equipe pedagógica, estudantes, gestores e outras partes interessadas na educação (GARCIA-PENALVO, 2020; WISE, 2019).

As técnicas de LA podem ser aplicadas tanto em ambientes formais quanto informais de aprendizagem, sendo que em cada um deles deve-se levar em consideração as características e especificidades da plataforma tecnológica (AHN, 2013; HUDGINS *et al.*, 2020). No contexto desta tese de doutorado, os ambientes formais são aqueles em que os processos educativos são baseados em normas institucionais e podem acontecer sob a supervisão de professores e/ou tutores; possuem uma ementa pré-definida e conteúdo hierarquicamente estruturado; normalmente, no final do processo são executados procedimentos formais de avaliação que determinam uma nota ou conceito, para fins de certificação (HAYTHORNTHWAITE; LAAT; SCHREURS, 2016; SCHREURS; De Laat, 2014). São exemplos de ambientes formais de aprendizagem os LMS e os MOOC.

Por outro lado, os ambientes informais visam a complementar a aprendizagem ocorrida em ambientes formais, conectando estudantes a outros segmentos da sociedade por meio de plataformas como as redes sociais, comunidades de aprendizagem *online* e sites do tipo

pergunta e resposta, do inglês *Question and Answer* (Q&A). Tais ambientes são usados para preencher a lacuna entre os currículos tradicionais e as necessidades ou interesses pessoais dos estudantes, expandindo as oportunidades de aprendizagem e análise crítica de informações por meio do envolvimento com um público maior e contexto mais abrangente (GRUZD; PAULIN; HAYTHORNTHWAITE, 2016; NISTOR *et al.*, 2020). À medida que os indivíduos progridem na vida escolar, acadêmica e profissional, a aprendizagem informal se torna cada vez mais importante para o desenvolvimento das habilidades e competências necessárias para a educação do século XXI. Assim, os ambientes informais apoiam o envolvimento em uma cultura de participação, colaboração e de compartilhamento de informações que potencializa a criação de uma sociedade baseada no conhecimento (AHN, 2013; KLAMMA, 2013; GRUZD; PAULIN; HAYTHORNTHWAITE, 2016; HUDGINS *et al.*, 2020; WORSLEY; OCHOA, 2020).

1.2 Motivação

Os ambientes informais não foram inicialmente projetados para fins de ensino e aprendizagem. Eles surgiram com o advento da Internet, se popularizaram rapidamente com a Web 2.0 e atualmente são utilizados por estudantes e profissionais para a execução de atividades colaborativas, o que tem levado pesquisadores a investigar sua utilização como contextos em que a aprendizagem acontece de forma espontânea (GREENHOW; GIBBINS; MENZER, 2015; GRUZD *et al.*, 2020; NISTOR; DERNTL; KLAMMA, 2015). Contudo, grande parte dos estudos em educação *online* e LA possuem como objetivo investigar os ambientes formais de aprendizagem, como os LMS ou os MOOC, em que o processo educativo acontece predominantemente de forma roteirizada. Este cenário faz com que os ambientes informais permaneçam sub-representados na pesquisa em aprendizagem *online*, mesmo considerando a sua onipresença em contextos pedagógicos (NISTOR; DERNTL; KLAMMA, 2015; GRUZD *et al.*, 2020; HUDGINS *et al.*, 2020).

A sub-representação dos ambientes informais de aprendizagem em pesquisas da área de educação *online* se deve, principalmente, às dificuldades inerentes à investigação de tais ambientes, como a complexidade em determinar as características dos seus participantes e a inexistência de mecanismos formais de verificação da aprendizagem ou taxas de inscrição e abandono (HUDGINS *et al.*, 2020). A análise de ambientes informais é um processo abrangente e complexo que exige uma abordagem analítica para compreender as várias dimensões do discurso dos participantes e a estrutura das suas interações sociais (JOKSIMOVIC *et al.*, 2019; SILVA; GIMENES; MALDONADO, 2020c). Apesar das dificuldades, pesquisadores e profissionais têm um interesse crescente em investigar o processo de aprendizagem em ambientes informais (HAFEEZ *et al.*, 2019; SILVA; GIMENES; MALDONADO, 2020a). Contudo, não existe consenso sobre como esta investigação pode ser realizada sistematicamente. Entre as principais lacunas de pesquisa nesta área estão a falta de ferramentas e métodos para investigar comportamentos relacionados à aprendizagem social *online* (LAAT; PRINSEN, 2014; NISTOR;

DERNTL; KLAMMA, 2015; PESARE; ROSELLI; ROSSANO, 2016; SILVA; GIMENES; MALDONADO, 2020c).

Com base nas lacunas de pesquisa em relação à investigação da aprendizagem em ambientes informais, esta tese de doutorado visa a avaliar condições e padrões comportamentais associados à aprendizagem social que ocorrem nos ambientes informais.

1.3 Objetivos e Questão de Pesquisa

A partir do contexto e motivação apresentados, o objetivo principal deste trabalho de doutorado é ampliar o entendimento e gerar conhecimento sobre como a aprendizagem social ocorre nos ambientes informais, por meio da avaliação de comunidades de aprendizagem *online*. Neste trabalho entende-se avaliação como a identificação, medição e análise de um conjunto de condições e padrões comportamentais fortemente associados à aprendizagem social, de acordo com as teorias educacionais. Portanto, o trabalho não visa a relacionar aspectos de aprendizagem a resultados específicos, como notas ou conceitos para fins de certificação, pois tais resultados não estão presentes de forma explícita em ambientes informais.

Para atingir o objetivo principal deste trabalho de doutorado, destacam-se os seguintes objetivos específicos:

- identificar as principais teorias educacionais, abordagens existentes e lacunas de pesquisa relacionadas ao domínio da avaliação da aprendizagem social *online* em ambientes informais;
- definir um processo intitulado **SLIM** (*Process for assessing online **S**ocial **L**earning within online communities in **I**nformal environ**M**ents*), cujo objetivo é guiar estudantes, pesquisadores e profissionais na avaliação de comunidades de aprendizagem *online* em ambientes informais;
- implementar os elementos necessários para automatizar a extração, análise e visualização dos dados educacionais utilizados ao longo da execução do processo;
- planejar e executar um estudo de caso envolvendo a avaliação de diferentes comunidades de aprendizagem *online*, com o intuito de investigar a efetividade do processo na geração de conhecimento sobre a aprendizagem social em ambientes informais.

Desta maneira, a principal questão de pesquisa que guia este trabalho de doutorado é:

- "como a análise dos dados gerados pelos participantes de comunidades *online* pode contribuir para ampliar a compreensão e gerar conhecimento sobre a aprendizagem social em ambientes informais de aprendizagem?"

1.4 Organização

Esta tese está organizada como descrito a seguir. Neste capítulo foram discutidos o contexto, a motivação, os objetivos e a principal questão de pesquisa do trabalho de doutorado. O capítulo 2 apresenta a fundamentação teórica com os conceitos necessários ao entendimento do trabalho. O capítulo 3 descreve o conceito de avaliação da aprendizagem em ambientes informais e evidencia as lacunas de pesquisa que motivaram o desenvolvimento do trabalho de doutorado. O capítulo 4 descreve o processo SLIM e a especificação das suas atividades. O capítulo 5 apresenta a caracterização das comunidades *online* avaliadas pelo processo SLIM, com o objetivo de efetivamente reconhecê-las como comunidades de aprendizagem. O capítulo 6 apresenta o estudo de caso cujo intuito é investigar a efetividade da aplicação do processo SLIM na geração de conhecimento sobre a aprendizagem social *online* em ambientes informais. Por fim o capítulo 7 descreve as conclusões deste trabalho de doutorado.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações iniciais

Os principais temas deste trabalho de doutorado estão relacionados à avaliação da aprendizagem social em ambientes informais; aos métodos e técnicas que apoiam a análise de dados obtidos de contextos educacionais; e aos *frameworks* teóricos que estabelecem os fundamentos pedagógicos que amparam a avaliação de comunidades *online*.

Neste sentido, este capítulo apresenta uma síntese dos fundamentos necessários para a compreensão deste trabalho de doutorado. Primeiramente, na seção 2.2 são apresentados os conceitos de aprendizagem social, ambientes formais e informais de aprendizagem. Em seguida, na seção 2.3 é apresentado o conceito de *learning analytics*, uma atividade que apoia a análise de dados educacionais. Na sequência, na seção 2.4 discute-se o conceito de *social learning analytics*, com destaque para as técnicas de análise de redes sociais e análise do discurso. Em seguida, na seção 2.5 são apresentados os *frameworks* teóricos que sustentam a abordagem apresentada neste trabalho de doutorado. Por fim, na seção 2.6 são apresentadas as considerações finais.

2.2 Aprendizagem social

A interatividade é frequentemente citada como um dos conceitos mais importantes para uma aprendizagem efetiva (VYGOTSKY, 1978; KENT; RECHAVI, 2020; LAAT; PRINSEN, 2014; SILVA; GIMENES; MALDONADO, 2020b). Aprender em um contexto social é um processo de construção de significados que pode ser baseado tanto em experiências anteriores, quanto no contexto mais imediato em que algo é aprendido (LAAT; PRINSEN, 2014; HAYTHORNTHWAITE, 2018). A aprendizagem social, do inglês *Social Learning* (SL), pode ser definida como um processo interativo e dinâmico que conta com a participação de diversos

atores e é caracterizado pela troca de conhecimento entre eles. Os atores aprendem e criam novos conhecimentos por meio da negociação de significados, interação e colaboração contínuas (HAYTHORNTHWAITE, 2018; KENT; RECHAVI, 2020). O foco da aprendizagem está em processos nos quais os atores não agem de forma isolada. Eles estão comprometidos em atividades sociais, como fazer ou responder perguntas, compartilhar informações ou dicas, relatar experiências, auxiliar na resolução de desafios e estimular mudanças de comportamento (WENGER; TRAYNER; De Laat, 2011; SHUM; FERGUSON, 2012).

Em contraste à abordagem instrutivista, que se baseia na transferência direta de conhecimento do instrutor para o estudante, na aprendizagem social construtivista o professor age como um facilitador e as participações ocorrem em um ambiente com foco nas interações do grupo, com o objetivo de construir as compreensões coletiva e individual por meio do diálogo (GARRISON, 2006; De Wever *et al.*, 2006). A comunicação entre o instrutor e os estudantes cria uma rede complexa de troca de conhecimentos. A quantidade de participação e o volume de interações mútuas são indicadores comuns usados para analisar a efetividade do processo de aprendizagem (KENT; LASLO; RAFAELI, 2016).

As discussões assíncronas *online* baseadas em texto, do inglês *Asynchronous text-based Online Discussions* (AOD) são reconhecidas como um recurso de grande potencial para promover a colaboração e intensificar o engajamento em ambientes baseados nas teorias da aprendizagem social (De Wever *et al.*, 2006; CHEN; HUANG, 2019; FERREIRA *et al.*, 2020). A análise de tais discussões pode revelar padrões de atividade importantes que evidenciem a aprendizagem e a construção de conhecimento (HAYTHORNTHWAITE *et al.*, 2018). As principais vantagens das AOD quando comparadas às discussões síncronas estão associadas à sua flexibilidade de tempo e espaço, que permite aos participantes refletir sobre o conteúdo da aprendizagem a qualquer momento e pesquisar ou sintetizar diferentes fontes de informação antes de realizar a sua contribuição para a discussão (WISE; ZHAO; HAUSKNECHT, 2014; KENT; LASLO; RAFAELI, 2016). O fato dos elementos de comunicação estarem explícitos nas contribuições escritas torna o processo de colaboração mais transparente, pois a transcrição dessas mensagens pode ser usada para investigar tanto o processo colaborativo do grupo quanto a participação do indivíduo neste processo (De Wever *et al.*, 2006).

2.2.1 Ambientes formais, não-formais e informais de aprendizagem

As atividades de aprendizagem social *online* podem ser realizadas em ambientes formais, não-formais ou informais de aprendizagem (LAAT; PRINSEN, 2014; SHUM; FERGUSON, 2012). A distinção entre esses termos pode ser controversa, e pesquisadores da área educacional propõem uma variedade de definições, muitas vezes conflitantes (CZERKAWSKI, 2016; HUDGINS *et al.*, 2020). A popularização das TIC e a grande quantidade de cursos e materiais disponíveis na Internet criaram um ecossistema de ambientes de aprendizagem que, muitas vezes, torna essa distinção mais difícil (GALANIS *et al.*, 2016; CZERKAWSKI, 2016). Neste sentido,

existem duas linhas principais de pensamento; na primeira, argumenta-se que os aspectos formal e informal são modos distintos de aprendizagem (ESHACH, 2007; MARSICK; WATKINS, 2001); na segunda, declara-se que eles apenas descrevem atributos diferentes da aprendizagem, e que quase todas as plataformas de ensino apresentam alguma combinação de elementos formais e informais (MALCOLM; HODKINSON; COLLEY, 2003; GALANIS *et al.*, 2016).

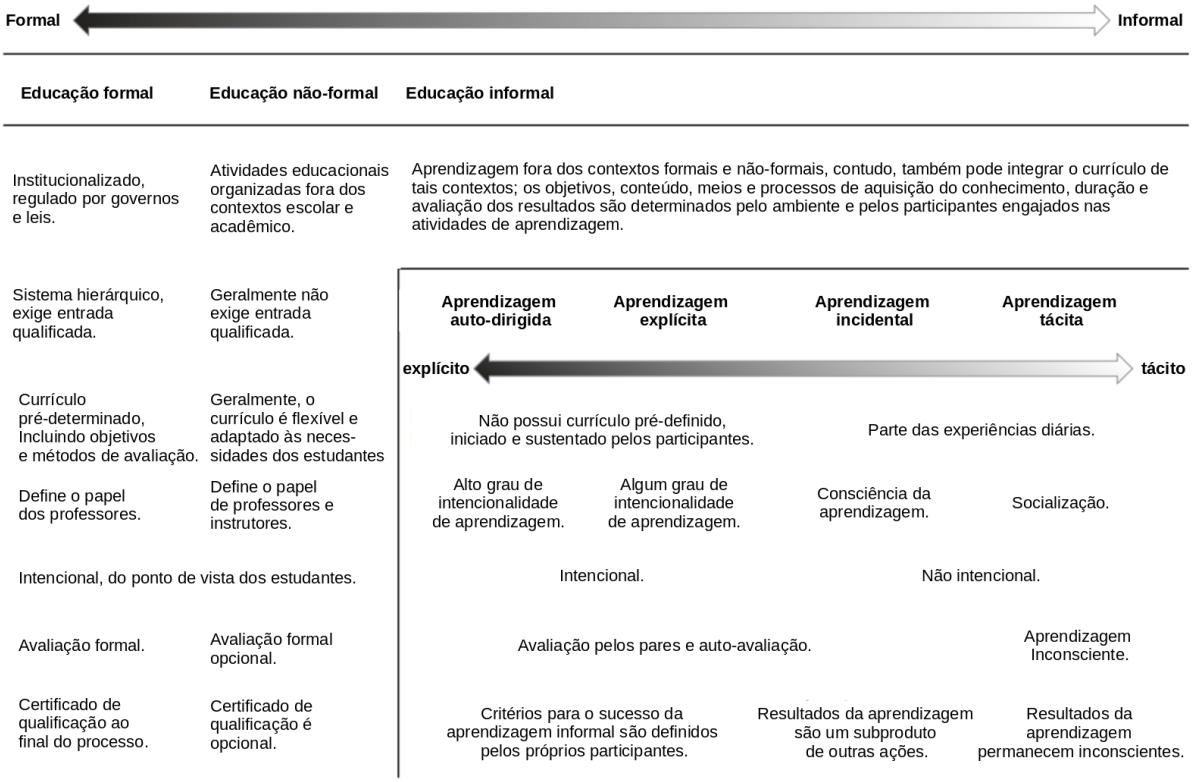
Apesar dessas diferenças, é possível estabelecer alguns consensos, como o que divide os contextos educacionais em ambientes formais, não-formais e informais de aprendizagem (ESHACH, 2007; CZERKAWSKI; HERNÁNDEZ, 2013; GALANIS *et al.*, 2016; SCHUMACHER, 2018). A Figura 1 apresenta as principais características dos três conceitos, que podem ser resumidos como descrito a seguir.

- Os **ambientes formais** apoiam atividades de aprendizagem realizadas nas instituições formais de ensino, como escolas e universidades. As atividades desenvolvidas são hierarquicamente estruturadas, possuem currículo e cronograma pré-definidos, são facilitadas por um professor ou instrutor e oferecem meios para obtenção de certificação (CZERKAWSKI, 2016; GALANIS *et al.*, 2016; SCHREURS; De Laat, 2014). Um exemplo de ambiente formal são os LMS que, entre outras tarefas, podem auxiliar no monitoramento do desempenho dos estudantes (CZERKAWSKI, 2016; CORBI; BURGOS, 2020).

- Os **ambientes não-formais** apoiam uma atividade educacional organizada e sistemática que ocorre fora dos ambientes formais de aprendizagem, portanto, apresenta um grau maior de flexibilidade em relação às tarefas desenvolvidas pelos estudantes (SCHUMACHER, 2018; CZERKAWSKI; HERNÁNDEZ, 2013). Tais ambientes oferecem meios para obtenção de certificação, contudo, ela é opcional. Um exemplo de ambiente não-formal são os MOOC.

- Os **ambientes informais** oferecem apoio à aprendizagem não estruturada, sem currículo ou cronograma pré-definidos e que, na maioria dos casos, acontece de forma espontânea fora dos contextos formais e não-formais (SHUM; FERGUSON, 2012; KLAMMA, 2013; GALANIS *et al.*, 2016; CZERKAWSKI, 2016). Como exemplo de ambientes informais de aprendizagem pode-se citar os sites de redes sociais, aplicativos de mensagens, sites de perguntas e respostas, *blogs* e comunidades *online* (NISTOR *et al.*, 2020; SPEILY *et al.*, 2020; GRUZD *et al.*, 2020). Os estudantes usam tais ambientes para enriquecer os currículos tradicionais, expandindo suas oportunidades de aprendizagem para além dos contextos formais e não-formais (AHN, 2013). À medida que os indivíduos progridem na vida escolar e acadêmica em direção à vida profissional, a aprendizagem informal desempenha um papel cada vez mais importante, e se torna essencial para o desenvolvimento de habilidades e aquisição de conhecimentos (GRUZD; PAULIN; HAYTHORNTHWAITE, 2016; NISTOR *et al.*, 2020).

Figura 1 – Principais características dos ambientes formais, não-formais e informais de aprendizagem



Fonte: Adaptada de Schumacher (2018).

Em reconhecimento às possíveis diferenças que o contexto pode trazer à interpretação destes conceitos, o presente trabalho de doutorado se refere aos contextos formais e não-formais como ambientes formais de aprendizagem. Além disso, com base nas definições apresentadas na Figura 1, considera-se que os ambientes informais possuem as características descritas a seguir.

- Apresentam, no mínimo, algum grau de intencionalidade da aprendizagem. Portanto, as atividades de aprendizagem social são iniciadas e sustentadas pelos próprios participantes.
- A aprendizagem é democrática. Embora, em alguns casos, as atividades de aprendizagem social possam ser facilitadas por um pequeno grupo, os ambientes informais se concentram no conhecimento criado e compartilhado por um grande número de participantes.
- O conhecimento dos participantes não é avaliado de forma padronizada ou integrada ao ambiente de aprendizagem. Portanto, não são aplicados processos formais de avaliação. Contudo, tais processos podem estar presentes de outras formas, como por meio de um mecanismo de avaliação pelos pares.
- A participação dos usuários não é autenticada com qualquer forma de certificação. Porém, eles podem receber o reconhecimento por suas contribuições no ambiente por meio de um sistema de pontuação exibido aos participantes.

Na área da pesquisa educacional, as comunidades de aprendizagem *online*, do inglês *Online Learning Communities* (OLC), são consideradas ambientes naturais para o desenvolvimento de atividades de aprendizagem social em contextos informais (NISTOR *et al.*, 2020; SPEILY *et al.*, 2020; KENT; RECHAVI, 2020; HUDGINS *et al.*, 2020). Elas oferecem a possibilidade de conectar estudantes com seus membros de diversos níveis de conhecimento e apoiar o envolvimento em uma cultura de compartilhamento de informações (NISTOR *et al.*, 2020; GRUZD; PAULIN; HAYTHORNTHWAITE, 2016; KLAMMA, 2013). Os conceitos de comunidade e comunidade de aprendizagem no contexto dos ambientes informais são descritos com mais detalhes na próxima subseção.

2.2.2 Definições e tipos de comunidades

Uma comunidade é um grupo de pessoas mutuamente engajadas que interage por longos períodos de tempo, compartilha um repertório de conhecimentos e relaciona-se de acordo com um certo tipo de orientação, como interesses comuns (em comunidades de interesse), objetivos comuns (em comunidades orientadas a objetivos), aprendizagem ou construção de conhecimento (em comunidades de aprendizagem) ou práticas sociais (em comunidades de prática) (HENRI; PUDELKO, 2003; HOD; BIELACZYC; BEN-ZVI, 2018; WENGER, 1999; NISTOR *et al.*, 2020).

De uma perspectiva ligeiramente diferente, Wenger, Trayner e De Laat (2011, p. 9) definem comunidade como "o desenvolvimento de uma identidade compartilhada em torno de um assunto ou conjunto de desafios". Essa identidade compartilhada implica o desenvolvimento de outras características sociais, cognitivas e emocionais, como história e intimidade compartilhadas (BRINT, 2001; MALINEN, 2015), memória coletiva (NISTOR *et al.*, 2020; HOSKINS, 2017), redes sociais e capital social (BRINT, 2001), coesão de grupo (GARRISON, 2007; HENRI; PUDELKO, 2003) e senso de comunidade (ROVAI, 2002). Em comunidades *online*, os participantes utilizam as TIC para estabelecer e manter contato, possibilitando a interação de uns com os outros por longos períodos de tempo (BRINT, 2001; HENRI; PUDELKO, 2003; MACIÀ; GARCÍA, 2016).

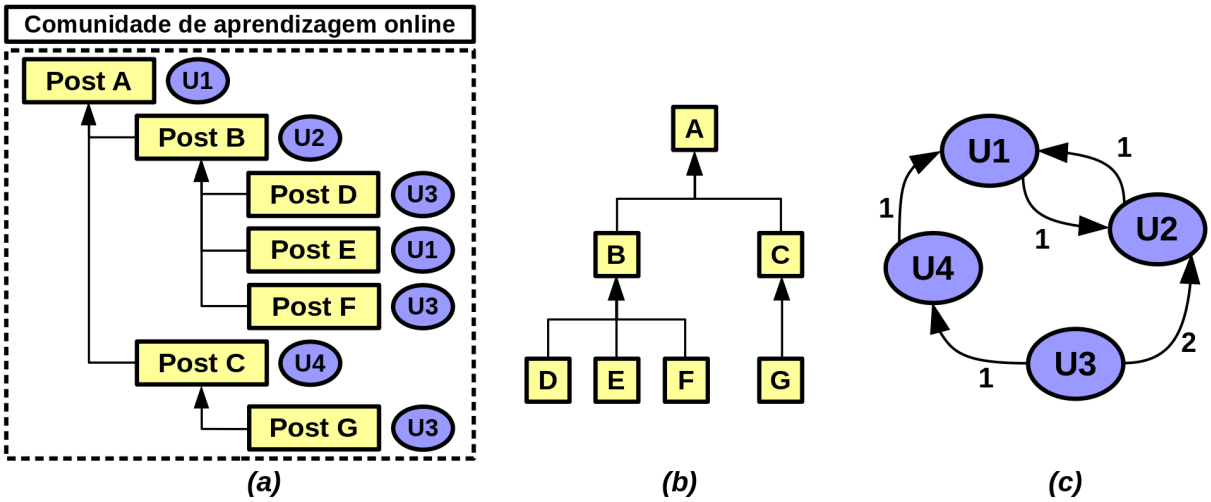
Diante destes conceitos, uma questão importante é distinguir quando uma comunidade *online* torna-se, de fato, uma comunidade de aprendizagem *online*. Um ponto fundamental envolve a transição da busca de informações pessoais para práticas coletivas que atendam a uma necessidade individual, mas compartilhada, organizada em torno de um domínio de conhecimento específico (JENKINS, 2009; HAYTHORNTHWAITE, 2018). Nesse contexto, as OLC podem ser definidas como grupos públicos ou privados de membros mutuamente engajados que interagem por meios digitais de forma predominantemente assíncrona, com o objetivo de atender às necessidades de aprendizagem dos seus participantes (SPEILY *et al.*, 2020; HAYTHORNTHWAITE, 2018; GALANIS *et al.*, 2016). Elas podem se estabelecer em diferentes plataformas tecnológicas baseadas na Web, como sites de redes sociais, sites de perguntas e

respostas e fóruns de discussão (NISTOR *et al.*, 2020; HAYTHORNTHWAITE, 2018).

Os papéis tradicionais de professor e estudante não são pré-estabelecidos nas OLC. Pelo contrário, as atividades que envolvem ensinar e aprender são definidas com base no conhecimento dos membros da comunidade sobre determinado assunto e no que eles fazem no ambiente. Ambos estão envolvidos no apoio à aprendizagem, na construção e sustentação da comunidade (HAYTHORNTHWAITE, 2018). Outros papéis surgem de acordo com as interações realizadas pelos participantes. Por exemplo, é comum a designação de moderadores que são responsáveis por gerenciar as trocas de conhecimentos, o comportamento relacionado às regras da comunidade e a aderência das contribuições à discussão (GILBERT, 2017).

Nas OLC, as discussões assíncronas normalmente ocorrem por meio de tópicos de discussão criados pelos participantes, organizados por tema ou ordem cronológica (SILVA; GIMENES; MALDONADO, 2020a; HAYTHORNTHWAITE *et al.*, 2018). A Figura 2 mostra a representação esquemática de um tópico de discussão e duas estruturas de dados comumente criadas com o objetivo representá-lo para fins de investigação. O participante que busca conhecimento faz a primeira postagem descrevendo o problema; outros participantes se engajam na discussão, contribuindo no mesmo tópico. Assim, a contribuição (ou participação) de um membro da comunidade é detectada quando ele publica ou responde um tópico de discussão. Na Figura 2**(a)**, os círculos retratam os participantes que publicaram ou responderam um tópico de discussão, representado pelos retângulos. A Figura 2**(b)** mostra a estrutura em árvore que representa o tópico de discussão. Por fim, a Figura 2**(c)** mostra a rede que representa as interações entre os participantes da discussão. Essas duas estruturas são descritas com mais detalhes na subseção 2.4.1.

Figura 2 – Representação esquemática de um tópico de discussão e suas estruturas de dados



Fonte: Adaptada de Silva, Gimenes e Maldonado (2020b).

No contexto educacional, as grandes comunidades de aprendizagem *online* disponíveis

na Internet são atualmente uma das fontes de produção de dados mais importantes do fenômeno conhecido como *Big Data* (NISTOR; DERNTL; KLAMMA, 2015; SILVA; GIMENES; MALDONADO, 2020a). À medida que grandes volumes de dados gerados por meio da participação em OLC tornam-se disponíveis, técnicas como LA assumem um papel importante para ajudar a compreender os processos de aprendizagem cada vez mais complexos que ocorrem nestes ambientes (CHATTI; MUSLIM; SCHROEDER, 2017; NACU *et al.*, 2019).

2.3 *Learning Analytics*

LA é uma técnica que se refere à medição, coleta, análise e documentação de dados sobre estudantes e seus contextos de aprendizagem, com o objetivo de compreender e otimizar os processos educativos e os ambientes em que eles ocorrem, sejam em contextos formais ou informais de aprendizagem (CHATTI; MUSLIM; SCHROEDER, 2017; SIEMENS; LONG, 2011). Chatti *et al.* (2012) fornecem uma visão geral de LA e seus conceitos-chave por meio de um modelo de referência baseado em quatro dimensões, conforme apresentado na Figura 3 e descrito a seguir.

- Dimensão "**o quê?**": refere-se aos dados dos participantes, do processo de aprendizagem e do ambiente em que ele ocorre, seja em um contexto formal ou informal. Os dados podem ser coletados em diferentes formatos e podem estar distribuídos ao longo do tempo e espaço.

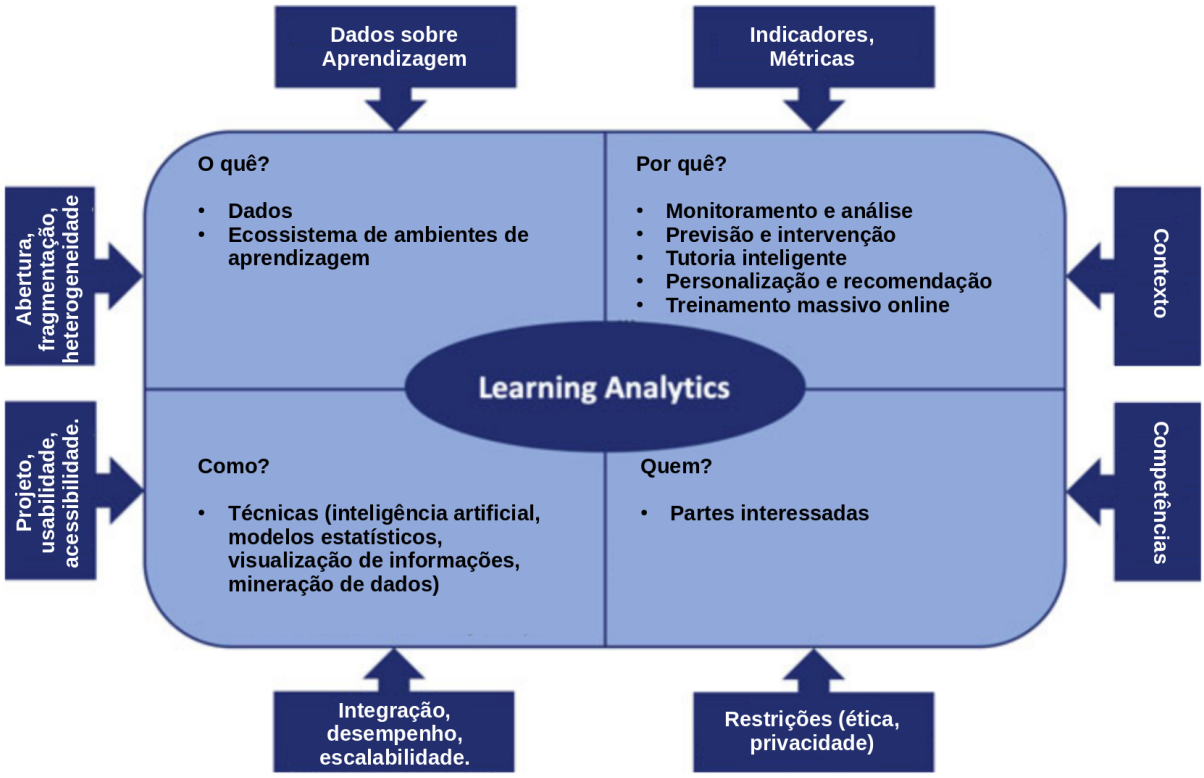
- Dimensão "**por quê?**": refere-se aos motivos pelos quais se analisam os dados coletados. Os objetivos da LA são diversos e precisam ser articulados de acordo com o ponto de vista das diferentes partes interessadas. Os possíveis objetivos que motivam à aplicação de LA em um determinado contexto incluem monitoramento, análise, previsão, intervenção, tutoria, avaliação, adaptação, personalização, recomendação, conscientização e reflexão.

- Dimensão "**quem?**": refere-se aos atores envolvidos no processo de aprendizagem. A aplicação de LA pode ser orientada para diferentes partes interessadas, incluindo estudantes, professores, tutores, moderadores, instituições educacionais, gestores e profissionais da educação, pesquisadores e projetistas de sistemas educacionais com diferentes perspectivas, objetivos e expectativas em relação à atividade de LA.

- Dimensão "**como?**": refere-se a como o sistema realiza a análise dos dados coletados. LA aplica diferentes métodos para detectar padrões ocultos em conjuntos de dados educacionais. Os métodos possíveis incluem modelos estatísticos, algoritmos de aprendizagem de máquina, visualização de informações e mineração de dados.

Na área da pesquisa educacional são definidos processos analíticos que visam a guiar a aplicação de LA. Tais processos são ciclos iterativos que geralmente compreendem quatro etapas

Figura 3 – Modelo de referência para LA



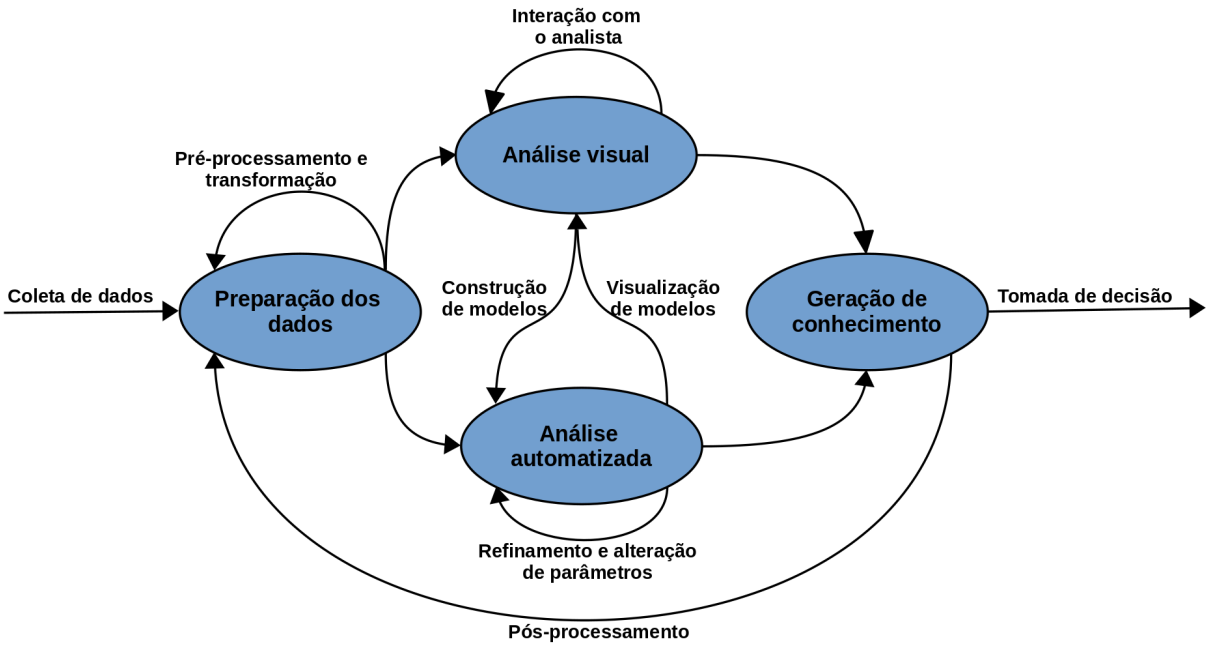
Fonte: Adaptada de Chatti *et al.* (2012).

principais (GARCIA-PENALVO, 2020): ***(i)*** pré-processamento dos dados; ***(ii)*** análise dos dados; ***(iii)*** tomada de decisão; e ***(iv)*** pós-processamento dos dados. Garcia-Penalvo (2020) sugere que essas etapas devem ser automatizadas, no entanto, a tomada de decisão requer a participação humana com base na visualização dos dados processados. Nesse contexto, surge um processo analítico visual, cujo intuito é obter conhecimento a partir da análise dos dados da aprendizagem (KEIM *et al.*, 2008; GÓMEZ-AGUILAR *et al.*, 2015). A Figura 4 mostra um modelo de processo analítico para LA (CHATTI *et al.*, 2012) com recursos de análise visual (KEIM *et al.*, 2010). As formas ovais representam as etapas e as setas representam as possibilidades de transição entre elas.

Após a coleta dos dados do ambiente de aprendizagem, o processo se inicia com a **preparação dos dados**. Em muitos cenários de aplicação de LA, diversas fontes de dados precisam ser integradas antes que as etapas de análise visual ou automatizada possam ser executadas. Portanto, geralmente é necessário pré-processar e transformar os dados a fim de criar diferentes representações para análise subsequente. Algumas tarefas típicas da etapa de preparação incluem limpeza de dados, normalização, agrupamento ou integração de fontes de dados heterogêneas.

Depois da preparação e transformação dos dados, a geração de conhecimento ocorre por meio da interação entre as etapas de análise visual e análise automatizada (criação de

Figura 4 – Modelo de processo analítico para LA com recursos de análise visual



Fonte: Adaptada de Chatti *et al.* (2012), Keim *et al.* (2008).

modelos). Se primeiramente for realizada a **análise automatizada**, algoritmos de aprendizagem de máquina podem ser aplicados para a criação de modelos que capturem padrões presentes nos dados (CHATTI *et al.*, 2012). Na sequência, as técnicas de visualização permitem que os analistas avaliem tais modelos e, caso necessário, os modifiquem por meio da alteração de parâmetros ou a escolha de outros algoritmos. A alternância entre os métodos visual e automatizado é característica dos processos de análise visual e leva a um refinamento contínuo dos resultados.

Caso os responsáveis pela análise optem por realizar primeiramente a **análise visual**, as possíveis hipóteses sobre os dados devem ser confirmadas em seguida por meio da análise automatizada. A interação com os métodos de visualização é necessária para revelar *insights* sobre os dados quando, por exemplo, o responsável pela atividade de LA concentra a sua análise em diferentes intervalos de tempo ou utiliza diferentes técnicas para visualizar os dados. A iteração entre as análises visual e automatizada leva a **geração de conhecimento** e tomada de decisão a partir dos dados de aprendizagem. Por fim, após a última etapa é possível reiniciar o processo por meio do pós-processamento dos dados.

Os processos analíticos que investigam dados educacionais normalmente são centrados em resultados baseados no desempenho dos estudantes em atividades de aprendizagem (SHUM; FERGUSON, 2012). Na educação formal, o desempenho geralmente é definido por meio de tarefas de avaliação, como exames ou questionários, destinados a avaliar o domínio do conhecimento em um determinado assunto. Este modelo é comum em ambientes como escolas, universidades ou MOOC, porém, é menos relevante nos contextos informais que envolvem

atividades de aprendizagem social realizadas ao longo da vida, baseadas em recursos disponíveis na Internet com o objetivo de solucionar problemas do cotidiano, muitas vezes sem o papel de um professor ou instituição de ensino formal (SHUM; FERGUSON, 2012; AHN, 2013; DAWSON, 2011; LAAT; PRINSEN, 2014). Nesse contexto, emerge o conceito de *Social Learning Analytics* (SLA).

2.4 *Social learning analytics*

Social learning analytics é uma sub-área de LA fortemente fundamentada em teorias educacionais. Ela tem por objetivo demonstrar que novas habilidades e ideias não são apenas conquistas individuais, mas são desenvolvidas e transmitidas por meio de interação e colaboração (SHUM; FERGUSON, 2012; LAAT; PRINSEN, 2014; JOKSIMOVIC; KOVANOVIC; DAWSON, 2019a; KENT; RECHAVI; RAFAELI, 2019; AHN, 2013). SLA concentra seus esforços em elementos relevantes da aprendizagem social *online*, evidenciando processos educativos nos quais os participantes estão envolvidos em atividades sociais, seja por interação direta com outras pessoas ou por meio de plataformas digitais que registram tais atividades para que elas sejam experimentadas por outros participantes no futuro (SHUM; FERGUSON, 2012; CHEN *et al.*, 2018; HERNÁNDEZ-GARCÍA *et al.*, 2015; KENT; RECHAVI, 2020).

SLA utiliza dados produzidos por estudantes, professores, tutores e outros atores envolvidos no processo educativo com o objetivo de identificar padrões comportamentais que evidenciem a efetividade da aprendizagem social *online* (LAAT; PRINSEN, 2014). Nesse contexto, Shum e Ferguson (2012) propuseram uma taxonomia para SLA formada por cinco analíticas, do inglês *analytics*, divididas em duas categorias: analíticas inerentemente sociais, do inglês *inherently social analytics*, e analíticas socializadas, do inglês *socialised analytics*.

As analíticas inerentemente sociais podem ser associadas a dois métodos de análise que investigam processos de aprendizagem em contextos coletivos (SHUM; FERGUSON, 2012):

- **análise de redes sociais, do inglês *Social Network Analysis* (SNA)** - investiga os processos intrínsecos a uma rede para entender como as pessoas desenvolvem e mantêm relações com o intuito de apoiar a aprendizagem;
- **análise do discurso, do inglês *Discourse Analysis* (DA)** - estabelece a linguagem como recurso primário para a construção do conhecimento e negociação dos significados.

As analíticas socializadas podem ser associadas a métodos de análise que, embora sejam relevantes para investigação da aprendizagem do ponto de vista individual, possuem atributos importantes para um contexto coletivo:

- **análise de conteúdo, do inglês *content analysis*** - analisa o conteúdo gerado pelos participantes, umas das principais características da Web 2.0;

- **análise de disposição, do inglês *disposition analysis*** - investiga as motivações que levam os participantes a se engajar nas atividades de aprendizagem;
- **análise de contexto, do inglês *context analysis*** - analisa os meios de acesso aos ambientes de aprendizagem *online*, tanto no contexto formal quanto informal.

Este trabalho de doutorado concentra-se em investigar o processo de aprendizagem social em OLC. Portanto, os métodos de análise relacionados às analíticas socializadas estão fora do escopo desta pesquisa. As próximas subseções descrevem os dois métodos de análise associados às analíticas inerentemente sociais.

2.4.1 Análise de Redes Sociais

No contexto da aprendizagem social, SNA é um método baseado na teoria dos grafos que visa a investigar os processos de formação de uma rede social e as propriedades das suas conexões relacionais, com o objetivo de compreender como a criação e manutenção de tal rede pode apoiar a aprendizagem (SHUM; FERGUSON, 2012; HAYTHORNTHWAITE, 2018). A Figura 2**(c)**, apresentada na seção 2.2.2, mostra a estrutura típica de uma rede social em contextos informais de aprendizagem e seus dois conceitos básicos: atores (ou nós) e conexões (ou ligações). Os atores representam os participantes da discussão e as conexões representam as interações entre eles. As conexões podem conter atributos. Na Figura 2**(c)** o atributo numérico mostrado nas ligações representa frequência de interação. Diante destes conceitos, este trabalho de doutorado referencia a rede social formada pelos atores de um ambiente educacional e suas respectivas conexões como uma rede de aprendizagem (HAYTHORNTHWAITE, 2018).

SNA permite investigar inúmeras atividades dos processos educacionais, como por exemplo, quais conexões descrevem mais precisamente o fluxo de informações dentro da rede, ou evidenciar quais atores prioritariamente fornecem ou recebem mais informações (LAAT; PRINSEN, 2014; HAYTHORNTHWAITE, 2018). As métricas definidas por este método podem fornecer detalhes sobre a natureza da participação dos atores em redes de aprendizagem. Em um nível individual, SNA assume uma perspectiva egocêntrica em que a rede de um ator em particular é investigada. Por outro lado, sob a perspectiva de analisar a rede como um todo, métricas a nível de rede consideram a importância da dinâmica do grupo e fornecem percepções abrangentes sobre a quantidade e a qualidade das interações sociais ocorridas na rede (SHUM; FERGUSON, 2012; HAYTHORNTHWAITE, 2018; JOKSIMOVIC *et al.*, 2019).

2.4.1.1 Métricas de SNA no contexto da aprendizagem social

SNA tem desempenhado um papel importante nas ciências educacionais para a caracterização e avaliação de ambientes de aprendizagem (JOKSIMOVIC *et al.*, 2019; GAŠEVIĆ *et al.*, 2019; PALAZUELOS; GARCÍA-SAIZ; ZORRILLA, 2013). A seguir é apresentada a definição de algumas das principais métricas egocêntricas de SNA.

- **Centralidade de grau, do inglês *degree centrality***: o grau refere-se à quantidade total de conexões em que um nó participa. Em redes direcionadas, o **grau de entrada** representa o número de conexões dirigidas ao nó, e o **grau de saída** o número de conexões com origem no nó. Intuitivamente, nós com um número maior de conexões são mais importantes do que aqueles com uma única conexão.

- **Centralidade de intermediação, do inglês *betweenness centrality***: esta métrica tem um papel importante para determinar o controle que determinados nós exercem sobre a rede. Ela mede a fração de todos os caminhos possíveis que passam por um determinado nó. Se o seu valor for alto, o nó potencialmente cumpre um papel importante no sentido de intermediar a comunicação na rede, e a sua remoção possivelmente dividiria a rede em dois ou mais componentes desconectados.

- **Centralidade de proximidade, do inglês *closeness centrality***: esta métrica representa a integração ou isolamento de um nó na rede. Ela representa o quão o nó está acessível para interagir com os demais sem depender de outros para transmitir as suas mensagens.

- **Centralidade de autovetor, do inglês *eigenvector centrality***: baseia-se na premissa de que a importância de um nó é determinada pela relevância ou influência de seus vizinhos. O cálculo é realizado com base em um processo recursivo em que o valor de cada nó é proporcional à soma das centralidades de autovetor dos nós aos quais ele está conectado.

- **Excentricidade, do inglês *eccentricity***: representa a maior distância mínima de um nó para todos os outros da rede.

- ***Pagerank*, *HITS Hub* e *Authority***: estas três métricas visam a classificar a importância de um determinado nó. A classificação é calculada como a probabilidade de que o nó seja alcançado ao percorrer um caminho aleatoriamente na rede (ZINOVIEV, 2018). No contexto da SNA, o algoritmo *Pagerank* (PAGE *et al.*, 2019) classifica os nós de acordo com a frequência com que um usuário hipotético alcançará o nó de forma não aleatória, quantificando assim a importância relativa de um indivíduo dentro da rede. O algoritmo *Hyperlink-Induced Topic Search* (HITS) (KLEINBERG, 1999) calcula duas métricas para cada nó: ***(i)*** *Hub* mede a qualidade das conexões de um indivíduo, considerando nós que geralmente estão vinculados a muitos outros (alto grau de entrada); ***(ii)*** *Authority* mede o quão valiosa é a informação de um determinado nó (alto grau de saída); normalmente, um nó com alto valor de *authority* está conectado a muitos *hubs* diferentes (ZINOVIEV, 2018; DE-MARCOS *et al.*, 2016).

Diversos trabalhos na área educacional descrevem como as métricas de SNA têm sido interpretadas em contextos de investigação da aprendizagem (CELA; SICILIA; SÁNCHEZ, 2015; DADO; BODEMER, 2017; JAN; VLACHOPOULOS; PARSELL, 2019). O Quadro 1 apresenta interpretações comumente atribuídas às métricas egocêntricas em tais contextos.

Quadro 1 – Interpretação das métricas egocêntricas em contextos de investigação da aprendizagem

Métrica	**Interpretação**
Centralidade de grau	Fornece informações sobre a influência e a popularidade de um ator na rede de aprendizagem. Um alto valor do grau de saída é associado à influência, enquanto um alto grau de entrada significa popularidade (JAN; VLACHOPOULOS, 2019; ANAYA *et al.*, 2015). Um ator influente é capaz de transmitir informações, alcançando atores mais distantes na rede, enquanto um ator popular atrai a interação de outros atores. Investigar o nível de influência e popularidade evidencia a função ou papel que o ator desempenha no contexto de aprendizagem (RISSER; BOTTOMS, 2014).
Centralidade de intermediação	É comumente usada para identificar atores considerados especialistas, atores que realizam a mediação do fluxo de informação ou conectam diferentes grupos presentes na rede de aprendizagem (JAN, 2019; GRUZD; PAULIN; HAYTHORNTHWAITE, 2016; JOKSIMOVIC *et al.*, 2015).
Centralidade de proximidade	Significa o quão um ator está acessível, assim, determina a facilidade de propagação das informações de um ator para o restante da rede de aprendizagem (ANAYA *et al.*, 2015; HERNÁNDEZ-GARCÍA *et al.*, 2015; GAŠEVIĆ *et al.*, 2019).
Centralidade de autovetor	Mensura o grau de importância de um ator com base nas suas conexões, assim, evidencia atores candidatos à disseminação de informação na rede de aprendizagem (ANAYA *et al.*, 2015; KENT; RECHAVI; RAFAELI, 2019; SOUSA-VIEIRA *et al.*, 2017).
Excentricidade	Utilizada para identificar atores centrais e periféricos. Aqueles com alta excentricidade ocupam as posições periféricas da rede de aprendizagem (NISTOR *et al.*, 2020; DE-MARCOS *et al.*, 2016; SAQR; FORS; NOURI, 2018).
Pagerank, *HITS Hub* e *Authority*	São comumente usadas para detectar especialistas na rede de aprendizagem, assim, identificam atores que auxiliam os demais membros, ou são procurados por fornecerem as melhores respostas (MENSHIKOVA, 2018; DE-MARCOS *et al.*, 2016; HERNÁNDEZ-GARCÍA *et al.*, 2015; ZHANG; ACKERMAN; ADAMIC, 2007).

A seguir é apresentada a definição de algumas das principais métricas de SNA em nível de rede.

- **Densidade, do inglês *density***: mede a fração de conexões existentes entre todas as potencialmente possíveis na rede. A densidade é um número entre 0 e 1, inclusive. Uma rede com densidade 0 não possui conexões, e uma rede com densidade 1 é chamada de completa ou totalmente conectada.

- **Reciprocidade, do inglês *reciprocity***: calculada em redes direcionadas, esta métrica mede a fração de conexões recíprocas, ou simétricas, em relação ao número total conexões.

- **Número de componentes, do inglês *connected components***: determina o número de componentes conectados. Um componente conectado é um subconjunto de nós da rede em que existe, no mínimo, um caminho que conecta cada nó a todos os outros nós do subconjunto.

- **Caminho médio, do inglês *average path length***: determina a média do comprimento dos caminhos existentes na rede. Um caminho é capaz de conectar indiretamente os nós que não são adjacentes, permitindo que haja interação entre eles.

- **Coeficiente de agrupamento, do inglês *clustering coefficient***: é uma medida da prevalência da quantidade de triângulos (três nós conectados) em uma rede. O coeficiente de agrupamento representa a fração de triângulos existentes entre os potencialmente possíveis. Pode ser usada para mensurar a extensão em que uma rede possui áreas de alta e baixa densidade.

- **Diâmetro, do inglês *diameter***: representa o valor da excentricidade máxima, ou o caminho mais longo entre os caminhos mais curtos que conectam dois nós quaisquer da rede.

- **Transitividade, do inglês *transitivity***: representa o número de conexões transitivas entre três nós A, B e C, em que A está conectado a B, B está conectado a C, e C fecha a estrutura conectando-se em A.

O Quadro 2 apresenta a interpretação comumente atribuída às métricas a nível de rede em contextos de investigação da aprendizagem.

2.4.1.2 *Análise da estrutura das discussões*

Para complementar a investigação dos padrões comportamentais e de interação em ambientes de aprendizagem, além das métricas de SNA descritas na subseção anterior, podem ser utilizadas métricas baseadas na estrutura das discussões para adicionar mais informações contextuais à análise da aprendizagem (WANG *et al.*, 2015; KOVANOVIĆ *et al.*, 2016; MENSHIKOVA,

Quadro 2 – Interpretação das métricas a nível de rede em contextos de investigação da aprendizagem

Métrica	**Interpretação**
Densidade	O valor da densidade tende a diminuir com o aumento do tamanho da rede, pois torna-se mais difícil conectar todos os atores (HERNÁNDEZ-GARCÍA *et al.*, 2015). Seu valor é uma evidência do índice de conectividade, que reflete a facilidade com que as informações podem alcançar os atores da rede (JAN, 2019; HERNÁNDEZ-GARCÍA *et al.*, 2015; HAYTHORNTHWAITE; LAAT; SCHREURS, 2016).
Reciprocidade	Um alto número de conexões recíprocas tem o potencial de apoiar o processo colaborativo de forma iterativa. Sendo assim, é interpretado como um indicador da troca mútua de conhecimentos, construção e negociação dos significados (KENT; RECHAVI, 2020; JAN, 2019; GAŠEVIĆ *et al.*, 2019).
Número de componentes	As redes de aprendizagem deveriam ser preferencialmente coesas para que o conhecimento possa fluir livremente. Assim, muitos componentes representam uma barreira à transferência de conhecimento (TOIKKANEN; LIPPONEN, 2011; SILVA; GIMENES; MALDONADO, 2020b; HELMS *et al.*, 2010).
Caminho médio	Indica o alcance que os atores têm um em relação aos outros. Um processo de aprendizagem eficaz tende a diminuir o valor do caminho médio (KENT; RECHAVI; RAFAELI, 2019; HAYTHORNTHWAITE; LAAT; SCHREURS, 2016; MISLOVE *et al.*, 2007).
Coeficiente de agrupamento	Indica a coesão ou proximidade entre os atores da rede. Espera-se um alto valor de coeficiente de agrupamento em contextos de aprendizagem que tendem a criar redes com densidade relativamente alta (JAN, 2019; KENT; RECHAVI; RAFAELI, 2019; DE-MARCOS *et al.*, 2016).
Diâmetro	Valores maiores de diâmetro sugerem que as informações compartilhadas podem atingir atores mais distantes na rede (GRUZD; PAULIN; HAYTHORNTHWAITE, 2016). A alteração do diâmetro reflete nas possibilidades de interação entre os atores, e afeta a memória coletiva do grupo (KENT; RECHAVI; RAFAELI, 2019; HERNÁNDEZ-GARCÍA *et al.*, 2015).
Transitividade	Indica a ocorrência de estruturas circulares fechadas (JAN; VLACHOPOULOS, 2018). Do ponto de vista da aprendizagem social, a diminuição da transitividade sugere que a rede reduziu tais estruturas, facilitando o fluxo de informações (KENT; RECHAVI, 2020; KENT; RECHAVI; RAFAELI, 2019).

2018; HAFEEZ *et al.*, 2019; FARROW; MOORE; GAŠEVIĆ, 2020; SILVA; GIMENES; MALDONADO, 2020b; SILVA; GIMENES; MALDONADO, 2020a). Tais métricas são chamadas de métricas de estrutura e podem ser calculadas com base na árvore da discussão, como a mostrada na Figura 2**(b)**, apresentada na seção 2.2.2. Elas analisam a estrutura hierárquica das mensagens publicadas pelos participantes da discussão para investigar os padrões de comunicação dos membros da comunidade, a fim de melhorar a efetividade do compartilhamento de conhecimento (WANG *et al.*, 2015).

As métricas de estrutura comumente utilizadas para a investigação da aprendizagem em contextos educacionais são as seguintes:

- **Número de participantes**: representa o número total de participantes que interagiram ao longo de um tópico de discussão (WENGER; TRAYNER; De Laat, 2011; SILVA; GIMENES; MALDONADO, 2020a; SILVA; GIMENES; MALDONADO, 2020b).
- **Tamanho da discussão**: representa o número total de respostas ou interações do tópico de discussão (FARROW; MOORE; GAŠEVIĆ, 2020; KOVANOVIĆ *et al.*, 2016).
- **Comprimento da discussão**: representa o número de respostas diretas da primeira publicação, representada pela raiz da árvore de discussão (FARROW; MOORE; GAŠEVIĆ, 2020; WANG *et al.*, 2015). Na Figura 2**(b)** o comprimento é igual a 2.
- **Profundidade da discussão**: representa a profundidade da raiz até a publicação mais profunda na árvore de discussão (KOVANOVIĆ *et al.*, 2016; WANG *et al.*, 2015). Na 2**(b)** a profundidade é igual a 3.
- **Intensidade da discussão**: representa a fração entre o número de respostas e o número de participantes da discussão (HAFEEZ *et al.*, 2019; SILVA; GIMENES; MALDONADO, 2020a).
- **Tempo da primeira resposta**: representa o tempo (em horas) da primeira resposta da discussão (MENSHIKOVA, 2018; SILVA; GIMENES; MALDONADO, 2020b).
- **Duração da discussão**: representa a diferença (em horas) entre o tempo de publicação do tópico de discussão e a sua última resposta (HAFEEZ *et al.*, 2019; SILVA; GIMENES; MALDONADO, 2020a).
- **Qualidade da discussão**: representa o *feedback* dos participantes à contribuição de determinado usuário (WENGER; TRAYNER; De Laat, 2011; SILVA; GIMENES; MALDONADO, 2020b; SILVA; GIMENES; MALDONADO, 2020a).

2.4.1.3 *Métodos estatísticos para redes*

Estudos que aplicam as métricas de SNA baseiam-se em modelos matemáticos para descrever relações entre variáveis, revelar características importantes e identificar condições

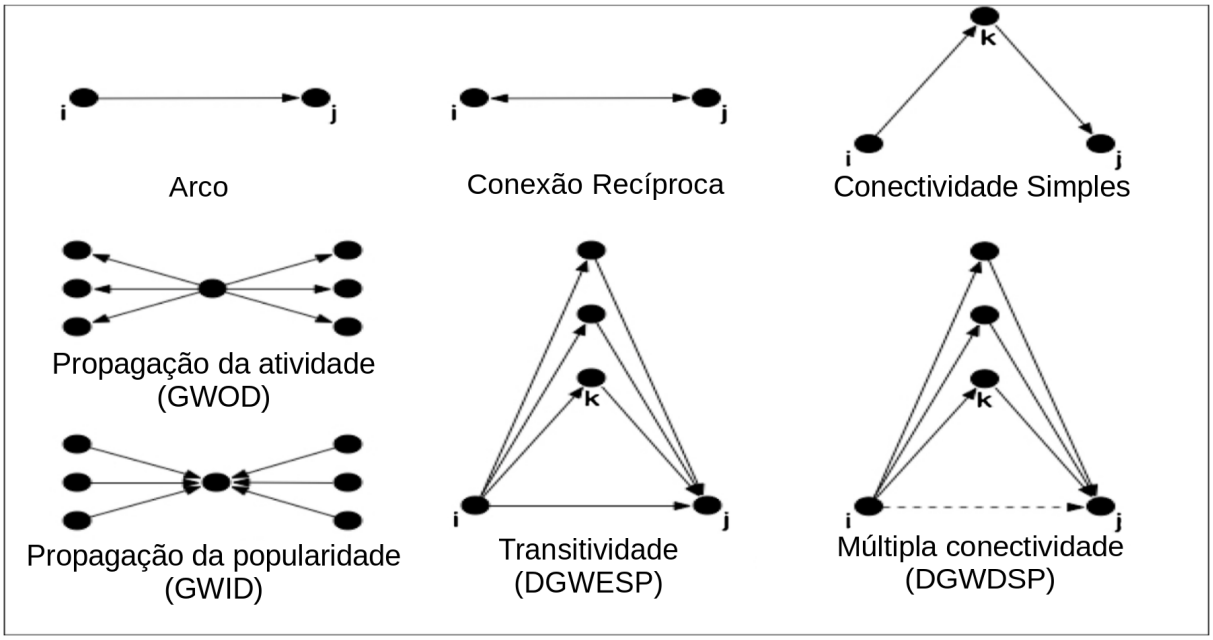
dentro da rede de aprendizagem investigada (CARRINGTON; SCOTT; WASSERMAN, 2005; FINCHAM; GAŠEVIĆ; PARDO, 2018). Por exemplo, modelos descritivos permitem identificar se condições associados à reciprocidade ou transitividade ocorrem em uma determinada rede. No entanto, para entender se tais condições ocorrem ou não com mais frequência que o esperado em redes geradas aleatoriamente, é necessário aplicar métodos estatísticos específicos para redes (GOODREAU; KITTS; MORRIS, 2009; ZHU *et al.*, 2016).

Um método comumente utilizado é o modelo de grafos randômicos exponenciais, do inglês *Exponential Random Graph Model* (ERGM) (LUKE, 2015; JOKSIMOVIĆ *et al.*, 2016; FINCHAM; GAŠEVIĆ; PARDO, 2018; DUBOIS; BUTTS; SMYTH, 2013). Ele pertence a uma família de modelos de probabilidade que permite criar inferências generalizáveis sobre as bases dos padrões de comportamento social em redes (FINCHAM; GAŠEVIĆ; PARDO, 2018). ERGM trata as conexões entre os nós como variáveis aleatórias e modela a estrutura geral da rede por meio de um conjunto de processos estatísticos. Cada conexão (aresta) dentro desses processos é condicionalmente dependente, indicando que "as arestas em redes empíricas não se formam ao acaso, mas se auto-organizam em vários padrões decorrentes de processos sociais subjacentes" (WANG *et al.*, 2013, p. 3). Portanto, ERGM pode confirmar hipóteses e contribuir metodologicamente para uma investigação mais profunda dos padrões de comportamento dos participantes em ambientes baseados em aprendizagem social (MAMAS *et al.*, 2020).

A Figura 5 apresenta sete efeitos estruturais proeminentes na teoria das redes sociais utilizados em contextos de aprendizagem social (ROBINS; LUSHER, 2013; FINCHAM; GAŠEVIĆ; PARDO, 2018; MAMAS *et al.*, 2020). Eles podem ser definidos como descrito a seguir.

- **Arcos**: é interpretado como a propensão para criação de conexões na rede.
- **Reciprocidade**: representa o efeito da propensão à reciprocidade na rede, ou seja, quando o ator i recebe uma resposta do ator j, aumenta a probabilidade do ator i enviar uma resposta ao ator j.
- **Conectividade simples**: representa o efeito que identifica a relação entre receber e enviar respostas, ou seja, os atores que recebem respostas são mais propensos a enviá-las a outros atores, e vice-versa.
- **Propagação da atividade, usando *Geometrically Weighted Out-degree Distribution* (GWOD)**: representa o efeito da distribuição do grau de saída na rede.
- **Propagação da popularidade, usando *Geometrically Weighted In-degree Distribution* (GWID)**: representa o efeito da distribuição do grau de entrada na rede.
- **Transitividade ou triangulação, usando *Geometrically Weighted Edgewise Shared Partner Distribution* (DGWESP)**: representa o efeito de que dois atores associados a um ator em comum são mais propensos a formar uma conexão.

Figura 5 – Efeitos estruturais comumente investigados em métodos estatísticos de rede



Fonte: Adaptada de Mamas *et al.* (2020).

- **Múltipla conectividade, usando *Geometrically Weighted Dyadwise Shared Partner Distribution* (DGWDSP)**: representa o efeito de profundidade da conectividade local entre pares de atores.

2.4.2 Análise do Discurso

Pesquisadores geralmente utilizam as técnicas de SNA para explorar as interações entre os participantes de discussões *online* (JOKSIMOVIC *et al.*, 2019; GAŠEVIĆ *et al.*, 2019). No entanto, a análise da rede de conexões não é suficiente para compreender de forma abrangente os padrões de interação de um determinado ambiente de aprendizagem mediado por computador (De Wever *et al.*, 2006; JOKSIMOVIC *et al.*, 2015; DASCALU *et al.*, 2018; JOKSIMOVIC *et al.*, 2019). A linguagem desempenha um papel único em tais ambientes, pois é o canal predominante usado pelos participantes de discussões *online* para trocar conhecimento (JOKSIMOVIC *et al.*, 2015). O discurso representa uma das fontes mais ricas de informação sobre os processos educativos, pois sua análise permite compreender as dimensões cognitivas, metacognitivas, afetivas e motivacionais dos atores envolvidos na aprendizagem (AZEVEDO, 2015; GAŠEVIĆ *et al.*, 2016).

Análise do discurso é o termo coletivo para uma variedade de abordagens para a investigação de eventos relacionados à comunicação. Algumas dessas abordagens oferecem meios de compreender a grande quantidade de texto gerado em ambientes de aprendizagem, apoiando a compreensão das dimensões sociais e cognitivas das interações *online* (SHUM; FERGUSON,

2012; KALIISA; MØRCH; KLUGE, 2019). Os métodos de análise automatizada do discurso são particularmente adequados para lidar com a escala crescente de dados produzidos em contextos educacionais mediados por tecnologia (JOKSIMOVIC *et al.*, 2015; FERREIRA *et al.*, 2020). Nesse contexto, a investigação do discurso produzido pelos participantes ao longo do tempo pode fornecer inúmeros *insights* sobre o nível de compreensão dos diferentes atores em discussões *online* (LEE; TAN, 2017). Atualmente, entre as ferramentas que apoiam a análise de grandes quantidades de texto, destaca-se o *framework* de investigação linguística e contagem de palavras, do inglês *Linguistic Inquiry and Word Count* (LIWC) (PENNEBAKER *et al.*, 2015).

Na atualidade, LIWC é uma das ferramentas mais adequadas para análise e avaliação de vários atributos presentes em mensagens de texto, sendo amplamente validada na literatura como um recurso adequado para revelar as principais características das mensagens criadas em discussões *online* (JOKSIMOVIC *et al.*, 2014; LIN; YU; DOWELL, 2020; FERREIRA *et al.*, 2020; GRUZD *et al.*, 2020). Ao todo, LIWC extrai do texto analisado 94 atributos divididos em quatro categorias:

- **resumo das variáveis da linguagem** - apresenta alguns atributos que descrevem as características gerais do texto analisado, como a quantidade de palavras, palavras por sentença e palavras com mais de seis letras;
- **dimensões linguísticas** - contém atributos que descrevem as principais dimensões linguísticas do texto, como pronomes pessoais e impessoais, artigos, preposições, verbos auxiliares, conjunções e negações;
- **gramática** - descreve outras dimensões linguísticas, como verbos comuns, adjetivos comuns, palavras relacionadas a comparação, interrogação e numerais;
- **processos psicológicos** - é a categoria com a maior quantidade de atributos, sendo formada por subcategorias como processos afetivos, processos sociais, processos cognitivos, processos perceptivos, processos biológicos, palavras relacionadas a orientações no tempo e preocupações pessoais, entre outras.

A partir de um determinado texto, LIWC é capaz de extrair o percentual de cada um dos atributos em relação ao total de palavras. Assim, é possível evidenciar os atributos ou categorias que são predominantes no texto sob investigação. Para fins de comparação, a Tabela 1 mostra as categorias e valores extraídos pelo LIWC da seguinte mensagem publicada em uma comunidade *online* sobre aprendizagem de programação:

> "*If you're good enough (or can get good enough) and you're willing to do the legwork then you can most likely find a remote gig without a degree. The demand for good developers is extremely high. You just have to sell yourself. Also since the cost of living in your area is almost certainly lower than in tech hubs you can live comfortably with a lower*

salary. If necessary that can be used as a bargaining chip. No matter where you live if you're a competent developer and can get people to believe it opportunity will follow."

Tabela 1 – Exemplos de atributos extraídos pelo LIWC

Categoria	Atributo	Valor
Linguistic Dimensions	*Auxiliary verbs*	14,4330
	Conjunctions	9,2784
	Negations	2,0619
	Prepositions	12,3711
Other Grammar	*Common verbs*	19,5876
	Interrogatives	1,0309
Psychological Processes	*Affective processes*	9,2784
	Cognitive processes	16,4948
	Drives	12,3711
	Negative emotion	2,0619
	Past focus	1,0309
	Personal concerns	4,1237
	Positive emotion	6,1856
	Present focus	21,6495
	Relativity	11,3402
	Social processes	10,3093

Fonte: Elaborada pelo autor.

2.5 *Frameworks* teóricos e aprendizagem social *online*

Uma crítica frequentemente apresentada na discussão de abordagens que propõem avaliar a aprendizagem social *online*, ou analisar dados educacionais de forma geral, é a falta de embasamento teórico (LAAT; PRINSEN, 2014; TOIKKANEN; LIPPONEN, 2011; NISTOR; DERNTL; KLAMMA, 2015; POQUET; DAWSON; DOWELL, 2017; JOKSIMOVIC; KOVANOVIC; DAWSON, 2019b). Aceita-se amplamente que a aplicação explícita e coerente de fundamentos pedagógicos ampara estudos empíricos na área de tecnologias educacionais (NISTOR; DERNTL; KLAMMA, 2015; SILVA; GIMENES; MALDONADO, 2020a). A importância da teoria na análise da aprendizagem também decorre da noção de validade, conceito que é definido como o grau em que a teoria e as evidências encontradas apoiam a interpretação dos resultados (JOKSIMOVIC; KOVANOVIC; DAWSON, 2019b). Portanto, esforços são necessários para alcançar uma maior convergência em relação aos modelos teóricos e a metodologia na avaliação de contextos educativos (MEIER; SPADA; RUMMEL, 2007). Com base nesta perspectiva, as próximas subseções descrevem os *frameworks* teóricos de criação de valor (WENGER; TRAYNER; De Laat, 2011) e presença social (GARRISON; ANDERSON; ARCHER, 1999) que visam apoiar as atividades de SNA e análise do discurso no contexto deste trabalho de doutorado.

2.5.1 Criação de Valor

Criação de valor refere-se ao conhecimento que redes ou comunidades criam quando elas promovem atividades de aprendizagem social (WENGER; TRAYNER; De Laat, 2011). O *framework* teórico proposto por Wenger, Trayner e De Laat (2011) permite avaliar a criação e o compartilhamento de conhecimento de forma proativa por meio de um conjunto de indicadores para a coleta e análise de dados. Os autores argumentam que grupos de pessoas que promovem a aprendizagem social podem ser avaliados considerando-se dois aspectos diferentes, porém complementares:

- o aspecto **comunidade** refere-se ao desenvolvimento de uma identidade compartilhada em torno de um assunto ou conjunto de desafios; representa uma intenção coletiva de administrar um domínio de conhecimento e sustentar o aprendizado sobre ele;

- o aspecto **rede social** refere-se ao conjunto de relacionamentos, interações pessoais e conexões entre participantes; a rede é representada como um conjunto de nós e ligações que possibilita o fluxo de informações para a resolução conjunta de problemas, criação e compartilhamento de conhecimento.

Comunidades e redes podem gerar diversos tipos de dados quantitativos e qualitativos sobre suas atividades. Alguns desses dados podem ser coletados facilmente, como diálogos mediados por recursos tecnológicos e registros de *download* de documentos. Outros indicadores mais sutis requerem uma avaliação substancial para serem úteis, como o nível de confiança ou a qualidade dos relacionamentos. O objetivo do *framework* de criação de valor é fornecer a base para um processo de avaliação que pode integrar fontes e tipos de dados heterogêneos para criar um panorama convincente sobre como as comunidades e redes criam valor para seus membros.

Este trabalho de doutorado se baseia nos indicadores de avaliação do aspecto rede social sugeridos por Wenger, Trayner e De Laat (2011) como fundamentação teórica para as atividades de SNA e análise da estrutura das discussões. Tais indicadores fornecem um panorama que evidencia diferentes aspectos das atividades de aprendizagem social em comunidades *online*. O Quadro 3 apresenta um resumo dos indicadores e seus principais elementos propostos no *framework* de criação de valor.

2.5.2 Comunidades de inquérito

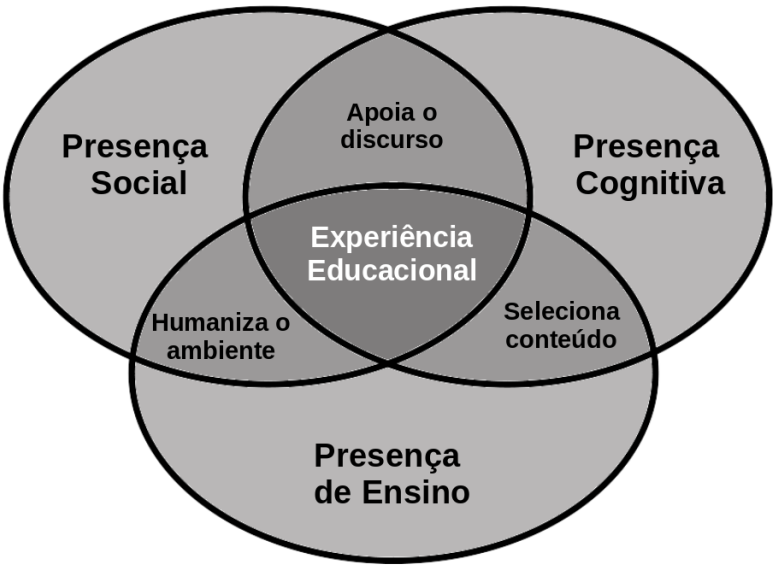
Várias abordagens têm sido propostas para investigar o discurso dos participantes em ambientes que promovem a aprendizagem social. Dentre eles, o *framework* teórico de comunidades de inquérito, do inglês *Community of Inquiry* (CoI), é um dos mais utilizados quando o objetivo é analisar as interações sociais e construção de conhecimento em discussões assíncronas *online* (FERREIRA *et al.*, 2020; KOVANOVIĆ *et al.*, 2018; JOKSIMOVIĆ *et al.*, 2015; GARRISON;

Quadro 3 – Indicadores para avaliação do aspecto rede da aprendizagem social

Indicadores	**Elementos**
Nível de participação	• Número de participantes ativos. • Características dos participantes ativos.
Nível de atividade	• Quantidade de tópicos de discussão. • Tempos de resposta dos tópicos de discussão.
Nível de engajamento	• Intensidade das discussões. • Comprimento dos tópicos de discussão. • Responder perguntas de outros participantes.
Qualidade das interações	• *Feedback* da qualidade das respostas dos tópicos de discussão.
Valor da participação	• Identificação de participantes que voltam a se engajar na rede.
Inspiração	• Taxas de retenção de membros.
Formato estrutural da rede	• Número de componentes da rede. • Coesão estrutural da rede. • Identificação de participantes que agem como "gargalos", do inglês *bottlenecks*, que promovem a conexão entre diferentes componentes da rede.

ANDERSON; ARCHER, 2010; SWAN; ICE, 2010). Ele é fundamentado nos modelos de aprendizagem construtivistas, portanto, evidencia o desenvolvimento do pensamento crítico baseado na interação que acontece em uma comunidade de aprendizagem (KOVANOVIĆ *et al.*, 2018). CoI propõe três dimensões que explicam os processos de construção do conhecimento social com o objetivo de descrever uma experiência educacional efetiva. As três dimensões propostas por Garrison, Anderson e Archer (1999), também chamadas de presenças, são representadas na Figura 6, e podem ser conceituadas como descrito a seguir.

Figura 6 – Interconexões das três presenças propostas pelo *framework* CoI



Fonte: Adaptada de Garrison, Anderson e Archer (2010).

- A **presença cognitiva** está diretamente relacionada aos resultados das atividades de aprendizagem. Seu objetivo é evidenciar o progresso das interações nos processos cognitivos dos estudantes, de forma a investigar o desenvolvimento do pensamento crítico, construção de conhecimento e resolução de problemas.
- A **presença de ensino** visa a facilitar e direcionar os processos cognitivos e sociais dos estudantes, com o propósito de alcançar resultados de aprendizagem mais significativos. Ela se baseia no papel do professor ou instrutor que inicialmente estabelece o formato dos procedimentos didáticos, e continua durante a instrução direta e/ou facilitação de tais procedimentos.
- A **presença social** visa a humanizar e apoiar a interação entre os participantes das discussões. Seu objetivo mais importante é proporcionar um ambiente confortável para que os estudantes possam trocar conhecimentos livremente, explorar diferentes perspectivas de um determinado assunto e resolver problemas de forma coletiva. Mais precisamente, o foco da presença social está no desenvolvimento progressivo das relações entre os membros da comunidade por meio de colaboração.

A capacidade que permite aos participantes interagir de forma efetiva em uma comunidade de aprendizagem e o nível de sua comunicação com os pares é inicialmente identificada na dimensão da presença social (ROURKE *et al.*, 1999; JOKSIMOVIĆ *et al.*, 2015). Ela apoia os processos cognitivos por meio de sua capacidade de instigar, sustentar e encorajar o pensamento crítico (ROURKE *et al.*, 1999). Garrison, Anderson e Archer (1999) definem doze indicadores que evidenciam a presença social em comunidades de aprendizagem. Eles são divididos em três categorias, descritas a seguir.

- **Afetiva**: analisa a tradução de emoções reais em texto. Abrange a expressão de emoções, sentimentos e humor.
- **Interativa**: evidencia a interatividade das mensagens trocadas entre os participantes. O principal objetivo desta categoria é melhorar a comunicação entre os membros da comunidade.
- **Coesiva**: investiga o senso de união e compromisso de grupo entre os participantes.

Alguns indicadores da presença social também são sugeridos por Wenger, Trayner e De Laat (2011) como evidências da criação de valor em comunidades. Esta evidência reitera a importância da dimensão presença social como recurso para avaliação da aprendizagem social *online*. A Tabela 2 mostra as três categorias da presença social, seu indicadores e a intersecção destes com os indicadores sugeridos no *framework* de criação de valor.

Tabela 2 – Intersecção entre os indicadores de presença social e *framework* de criação de valor

Categoria	Indicador da presença social	Indicador da criação de valor
Afetiva	Expressar emoções	Expressão de emoções
	Uso de humor	Evidências de diversão
	Auto-divulgação	-
Interativa	Continuar uma discussão	Comprimento das discussões
	Responder perguntas	Quantidade e tempo das respostas
	Referenciar mensagens de outros	Referências à própria comunidade
	Fazer perguntas a outros membros	Quantidade de perguntas
	Elogiar ou expressar apreciação	-
	Expressar concordância	-
Coesiva	Uso de vocativos	-
	Uso de pronomes para referenciar o grupo	Referências à própria comunidade
	Cumprimentos e saudações	-

A identificação desses doze indicadores revela o nível de presença social em uma comunidade *online*. A baixa frequência de tais indicadores evidencia que o ambiente de aprendizagem é impessoal, ou seja, os participantes estão usando a comunidade de maneira pragmática para trocas concisas de informações. Por outro lado, a alta frequência de presença social indica que os membros possuem uma sensação de afiliação e se identificam com o grupo. Este contexto

de proximidade evidencia que os participantes consideram a comunidade como um ambiente que promove a aprendizagem. Isto, por sua vez, os apoia no ato de publicar e desenvolver suas ideias provisórias, inclusive oferendo críticas construtivas às hipóteses dos outros participantes (ROURKE *et al.*, 1999).

2.6 Considerações finais

Neste capítulo, foi sintetizada a fundamentação teórica sobre aprendizagem social e comunidades *online* em ambientes informais. Inicialmente, apresentou-se a definição de ambientes formais, não-formais e informais de aprendizagem, com o objetivo de descrever o contexto educacional ao qual se aplica este trabalho de doutorado. Foram apresentados ainda os conceitos de SLA e métodos relacionados que apoiam a análise de dados educacionais. Por fim, foram descritos os *frameworks* teóricos que fundamentam o desenvolvimento do trabalho. No próximo capítulo, é apresentado o mapeamento sistemático e as lacunas de pesquisa nas quais este trabalho de doutorado se baseia.

CAPÍTULO

3

AVALIAÇÃO DA APRENDIZAGEM EM AMBIENTES INFORMAIS

3.1 Considerações iniciais

Este capítulo discute a avaliação de OLC em ambientes informais de aprendizagem e descreve lacunas de pesquisa evidenciadas neste tema. Inicialmente, a seção 3.2 contextualiza o interesse de pesquisadores da área educacional na investigação de ambientes informais; e a seção 3.3 apresenta conceitos básicos sobre a avaliação de tais ambientes. Na sequência, a seção 3.4 apresenta um Mapeamento Sistemático (MS) que analisa abordagens de avaliação de comunidades *online* e evidencia duas lacunas de pesquisa. Em seguida, são descritas as conclusões de uma Revisão de Literatura (RL), cujo objetivo é complementar as conclusões do MS e evidenciar três lacunas de pesquisa adicionais às identificadas anteriormente. Por fim, a seção 3.6 faz as considerações finais do capítulo.

3.2 Ambientes informais e atividades de aprendizagem

Educadores e pesquisadores estão cada vez mais interessados no uso e investigação dos ambientes informais, como as OLC, em atividades de aprendizagem (GARCÍA-PEÑALVO; COLOMO-PALACIOS; LYTRAS, 2012; GREENHOW; GIBBINS; MENZER, 2015; GALANIS *et al.*, 2016; SPEILY *et al.*, 2020; SILVA; GIMENES; MALDONADO, 2020a; SILVA; GIMENES; MALDONADO, 2020b). Tal interesse tem se intensificado pois enquanto os ambientes formais conectam estudantes a uma base de usuários relativamente homogênea, as OLC comumente são mais extensas e possibilitam a conexão dos participantes em uma rede de aprendizagem heterogênea, formada por pessoas com diferentes níveis de conhecimento (GALANIS *et al.*, 2016; SPEILY *et al.*, 2020). Assim, elas oferecem um ambiente que potencializa o desenvolvimento de habilidades que são importantes para a Educação do Século XXI, como

pensamento crítico, comunicação, capacidade de pesquisar, sintetizar e compartilhar informações por meios tecnológicos, além da compreensão sobre como gerenciar e integrar diferentes fontes de informação em seu próprio trabalho, incluindo a avaliação de tais fontes como confiáveis ou não confiáveis (AHN, 2013; ROTHERHAM; WILLINGHAM, 2010). O registro em larga escala das ações dos participantes fornece uma vasta quantidade de conhecimento implícito, e uma nova perspectiva para que pesquisadores compreendam as experiências de aprendizagem dos estudantes fora do ambiente controlado dos contextos formais (CHEN; VORVOREANU; MADHAVAN, 2014).

Os dados produzidos pelos participantes e registrados pelas OLC podem ser utilizados para avaliação da aprendizagem em ambientes informais, cujo objetivo é evidenciar condições e padrões comportamentais associados à aprendizagem social que acontece em tais ambientes. A próxima subseção descreve com mais detalhes este contexto.

3.3 Avaliação da aprendizagem em ambientes informais

Geralmente, ambientes informais de aprendizagem, como as OLC, não possuem procedimentos formais de avaliação. Pelo contrário, ela é realizada de modo informal e coletivo, por meio da interação e troca de conhecimento entre os participantes (GALANIS *et al.*, 2016; HAYTHORNTHWAITE, 2018). Além disso, destaca-se o processo de autoavaliação, que acontece baseado no *feedback* que as contribuições feitas à comunidade recebem dos outros participantes (BOEKAERTS; MINNAERT, 1999). O papel exercido pelos moderadores da comunidade é considerado importante, pois devido à natureza não estruturada dos ambientes informais de aprendizagem, eles são responsáveis por acompanhar as atividades desenvolvidas pelos participantes, realizando intervenções regularmente para motivá-los, avaliá-los ou monitorar a aderência das contribuições ao tema da discussão (LEE, 2013; GALANIS *et al.*, 2016; HAYTHORNTHWAITE, 2018).

Neste contexto, apesar de algumas possíveis diferenças, normalmente as OLC disponibilizam algum procedimento de avaliação pelos pares, com base em um sistema de pontos exibido a toda comunidade que recompensa a frequência e a qualidade das contribuições individuais (HUDGINS *et al.*, 2020; GALANIS *et al.*, 2016). Assim, os participantes podem avaliar uns aos outros atribuindo pontos positivos ou negativos às suas respostas. Esta avaliação indica o quanto os membros da comunidade são ativos e respeitados, o que reflete o tempo e o esforço investidos na preparação de respostas de alta qualidade antes da publicação (WENINGER, 2014). Participantes relatam confiança no valor das informações compartilhadas por seus membros mais ativos, pois a frequência na participação é interpretada como comprometimento com o grupo (POQUET; DAWSON; DOWELL, 2017). Além disso, os participantes ativos e com bom desempenho na pontuação do sistema de avaliação por pares são comumente reconhecidos como especialistas, ou *experts*, pelos demais membros da comunidade (KLAMMA, 2013).

Em relação aos padrões de qualidade das discussões, eles são criados implicitamente quando os participantes avaliam, de forma positiva ou negativa, os tópicos de discussão e suas respostas (HUDGINS *et al.*, 2020). Geralmente, os votos positivos associados a uma determinada discussão indicam a opinião da comunidade sobre ela. Os temas mais populares, ou com as melhores respostas, provavelmente serão bem avaliados, enquanto aqueles impopulares, ou com as piores respostas, não receberão uma pontuação elevada. Assim, de forma geral, o saldo de votos positivos em um tópico indica a opinião da comunidade sobre a discussão (MIEGHEM, 2011; HUDGINS *et al.*, 2020).

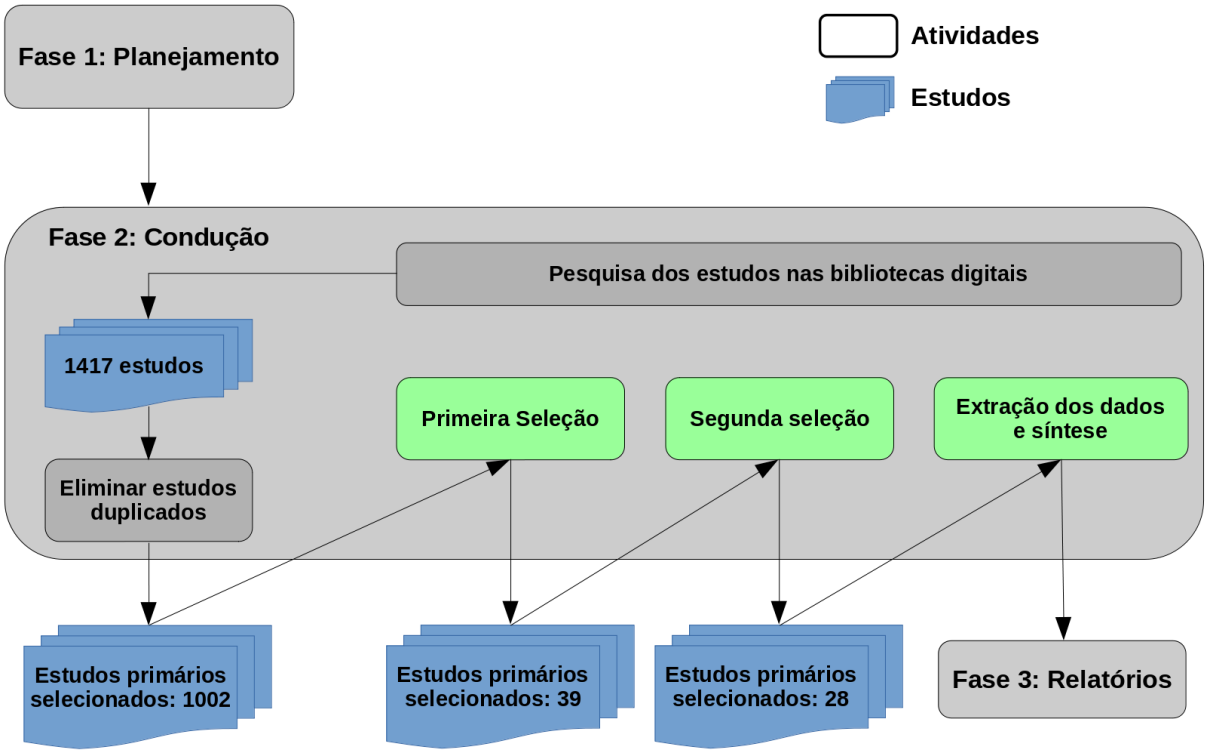
3.4 Mapeamento sistemático sobre a avaliação da aprendizagem em comunidades *online*

A investigação de ambientes informais de aprendizagem oferece dificuldades metodológicas referentes à análise dos dados gerados pelos participantes para fins educacionais (CHEN; VORVOREANU; MADHAVAN, 2014). A natureza heterogênea e não estruturada das atividades, a ausência de processos formais de avaliação, mecanismos de inscrição e taxas de abandono dificultam a determinação da efetividade da aprendizagem (HUDGINS *et al.*, 2020). Tais dificuldades tornam a investigação de contextos informais uma área sub-representada na pesquisa educacional, quando comparada aos contextos formais (HUDGINS *et al.*, 2020). Assim, não existe consenso sobre como mensurar e pouco se sabe de que modo os processos de aprendizagem ocorrem em ambientes como as OLC (CHEN; VORVOREANU; MADHAVAN, 2014; WANG *et al.*, 2015; GREENHOW; GIBBINS; MENZER, 2015; SPEILY *et al.*, 2020). Pesquisadores e profissionais relatam a necessidade de métodos e ferramentas que mensurem a efetividade e revelem os benefícios das atividades de aprendizagem social em contextos informais (SCHREURS; De Laat, 2014; LAAT; PRINSEN, 2014; NISTOR; DERNTL; KLAMMA, 2015; SILVA; GIMENES; MALDONADO, 2020b).

Nesta perspectiva, Silva, Gimenes e Maldonado (2020c) conduziram um MS com o objetivo de investigar lacunas de pesquisa em estudos que apresentem abordagens de avaliação de comunidades no âmbito da aprendizagem social. Um MS é um estudo secundário que permite identificar, avaliar e interpretar dados extraídos de estudos primários relacionados a questões de pesquisa predefinidas (KITCHENHAM, 2004). No contexto deste MS, as comunidades *online* são referenciadas como comunidades virtuais.

De acordo com Kitchenham e Charters (2007), um MS é composto por três fases: planejamento, condução e produção de relatórios. A Figura 7 mostra as três fases executadas no MS, enfatizando as atividades e a quantidade de estudos analisadas na fase de condução. Na fase de planejamento foram definidas a *string* de busca, as bibliotecas digitais a serem pesquisadas, as Questões de Pesquisa (QP) e os critérios de inclusão e exclusão. Entre as QP definidas, destacam-se as descritas a seguir.

Figura 7 – Fases do mapeamento sistemático



Fonte: Adaptada de Silva, Gimenes e Maldonado (2020c).

- **QP1. Uso de bases teóricas**: quais bases teóricas foram aplicadas nos estudos primários? O objetivo desta questão de pesquisa é investigar se os autores utilizaram bases teóricas para conceber a abordagem descrita nos estudos.

- **QP2. Estratégias para análise de dados**: quais são as técnicas e estratégias utilizadas para a análise de dados estruturados e não estruturados? No contexto do MS, os dados estruturados são representados por meio de grafos e evidenciam a interação entre os participantes, como acontece na análise de redes sociais. Por sua vez, os dados não estruturados representam o conteúdo produzido pelos participantes, como o texto gerado nas discussões e investigado na análise do discurso.

A *string* de busca foi aplicada em cinco bibliotecas digitais: *ACM Digital Library*[1], *Engineering Village*[2], *IEEE Xplore*[3], *ScienceDirect*[4] e *Scopus*[5]. Inicialmente, após a realização da busca inicial foram encontrados 1417 estudos. Na sequência, a atividade de eliminação de duplicados excluiu 415 estudos. A atividade de primeira seleção envolveu a leitura do

[1] http://dl.acm.org
[2] http://www.engineeringvillage.com
[3] http://ieeexplore.ieee.org
[4] http://www.sciencedirect.com
[5] https://www.scopus.com

título, resumo e palavras-chave de cada estudo primário e aplicação dos critérios de exclusão e inclusão, e em alguns casos também foram lidas a introdução e as conclusões. Esta atividade considerou que 39 estudos primários eram relevantes para a leitura completa. Após a leitura do texto completo desses estudos, na segunda seleção, os critérios de exclusão e inclusão foram novamente aplicados, o que levou a 28 estudos primários relevantes que foram analisados para responder às questões de pesquisa estabelecidas. As próximas subseções descrevem as respostas das principais QP.

3.4.1 QP1: uso de bases teóricas

Existe uma extensa literatura disponível para apoiar a investigação da aprendizagem social em comunidades *online*. Contudo, a maioria dos estudos analisados não utiliza efetivamente essas bases teóricas para conceber as suas abordagens, como mostrado no gráfico da Figura 8:

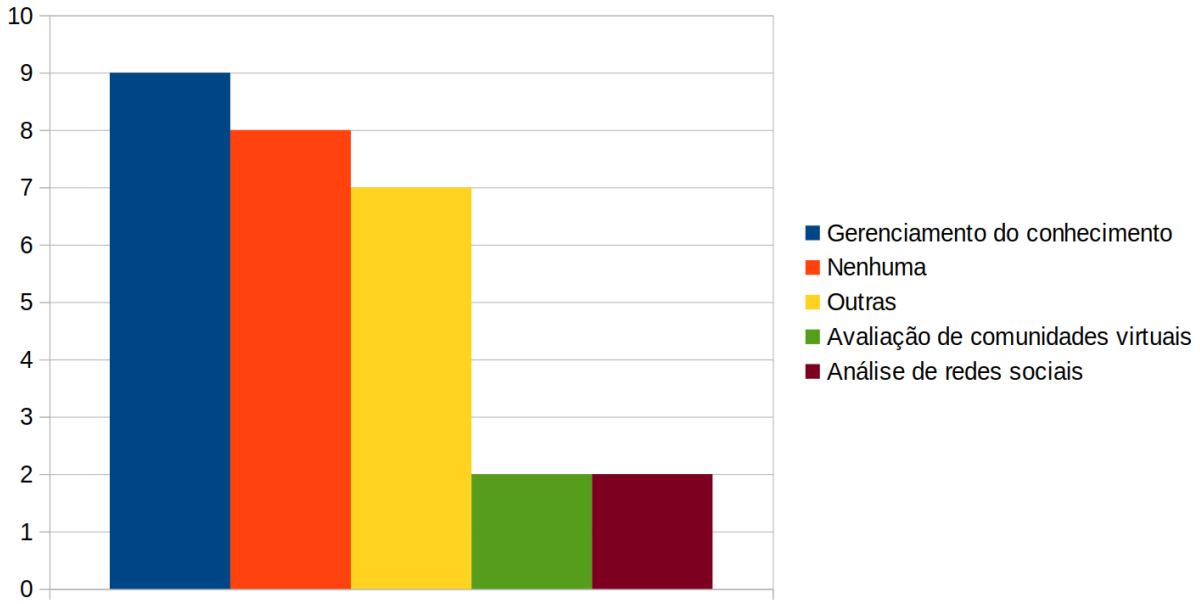
- 17 estudos referenciaram superficialmente teorias relacionadas ao gerenciamento do conhecimento em comunidades *online*, ou simplesmente não utilizaram bases teóricas para conceber as suas abordagens;
- sete estudos referenciaram bases teóricas relacionadas ao domínio da comunidade que estava sendo investigada, como a análise envoltória de dados, do inglês *Data Envelopment Analysis* (DEA) (VENKATESH; THONG; XU, 2012), e a teoria de aceitação e uso de tecnologias (CHARNES; COOPER; RHODES, 1978);
- dois estudos basearam-se nas teorias de análise de redes sociais (SHUM; FERGUSON, 2012) para avaliar a interação entre os membros da comunidade;
- por fim, apenas dois estudos utilizaram o *framework* teórico de criação de valor (WENGER; TRAYNER; De Laat, 2011) para apoiar a avaliação da aprendizagem social em comunidades *online*.

3.4.2 QP2: estratégias para análise de dados

As técnicas utilizadas para análise de dados estruturados e não estruturados são mostradas na Figura 9. Sete estudos primários utilizaram métodos qualitativos como observações, entrevistas e questionários para analisar tanto as interações sociais quanto o conteúdo produzido pelos participantes de comunidades *online*. Tais métodos são mais onerosos e ineficientes quando comparados ao uso de estratégias automatizadas.

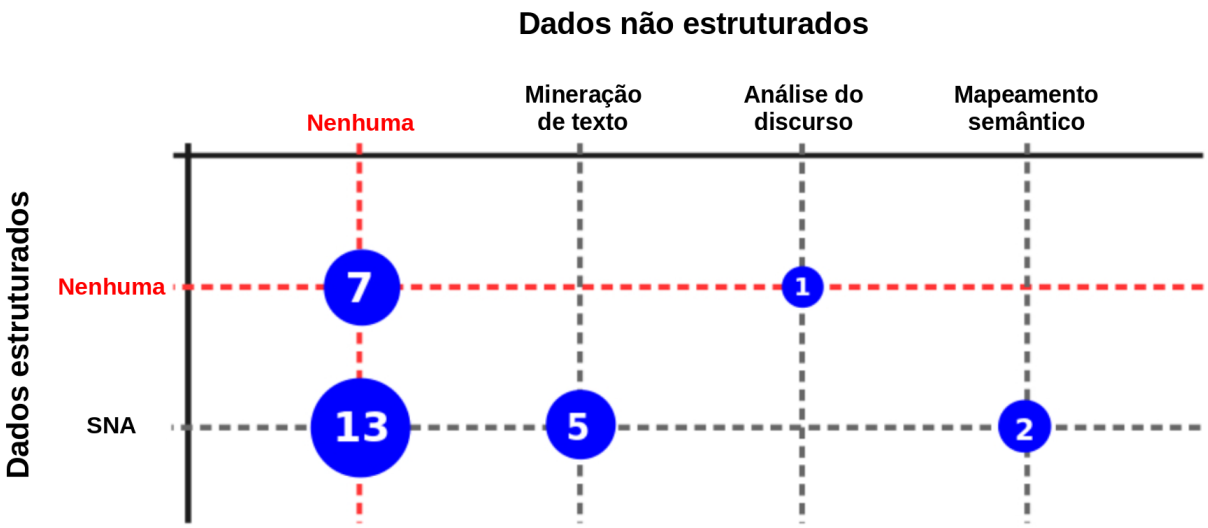
SNA é o método predominantemente aplicado para analisar dados estruturados, sendo utilizado em 20 estudos primários. Ele é frequentemente referenciado como uma técnica eficiente para analisar o comportamento e a interação entre os membros de comunidades *online*. No entanto, sua aplicação não é suficiente para fornecer uma visão completa do desempenho da

Figura 8 – Bases teóricas referenciadas nos estudos primários



Fonte: Adaptada de Silva, Gimenes e Maldonado (2020c).

Figura 9 – Métodos para análise de dados utilizados nos estudos primários



Fonte: Adaptada de Silva, Gimenes e Maldonado (2020c).

comunidade em relação à sua evolução, criação e compartilhamento de conhecimento. Assim, SNA precisa ser combinada com métodos de análise não estruturada para gerar resultados mais próximos da realidade.

Técnicas de SNA são combinadas com a análise de dados não estruturada em sete dos estudos primários, a fim de se obter um diagnóstico mais preciso sobre o conteúdo das discussões das comunidades *online*. Cinco destes estudos aplicaram estratégias relacionadas à mineração de texto, como *Latent Dirichlet Allocation* (LDA) e *Latent Semantic Analysis* (LSA) (BILBRO; OJEDA; BENGFORT, 2019). Dois estudos combinaram SNA e técnicas de mapeamento semântico para relacionar os termos léxicos dos domínios investigados aos tópicos mais importantes discutidos pelos membros da comunidade. Por fim, um único estudo utilizou estratégias de análise do discurso para investigar a qualidade dos diálogos produzidos na comunidade *online* de uma universidade americana, sem combiná-las com técnicas de SNA.

Nesse contexto, as conclusões do MS evidenciam que as estratégias utilizadas para analisar dados estruturados e não estruturados nas abordagens de avaliação de comunidades *online* são consideradas complementares, pois podem fornecer métricas de diferentes aspectos das comunidades. Por um lado, a maioria dos estudos selecionados utilizou técnicas de SNA para a análise dos dados estruturados. Contudo, apenas 25% deles combinaram técnicas estruturadas e não estruturadas para analisar a interação entre os membros e o conteúdo discutido por seus membros.

3.5 Revisão de literatura complementar ao MS

Após a condução do MS foi realizada uma RL com o objetivo de complementar as conclusões descritas nas subseções 3.4.1 e 3.4.2. A RL analisou estudos relacionados à investigação de discussões *online* no contexto da aprendizagem social, tanto em ambientes formais quanto informais, e permitiu evidenciar três lacunas de pesquisa complementares àquelas identificadas no MS: ***(i)*** a falta de estudos que realizem análise de dados em larga escala; ***(ii)*** a falta de aplicação de métodos estatísticos de rede para identificar padrões de comportamento entre os participantes de discussões; e ***(iii)*** a escassez de abordagens específicas para a investigação de ambientes informais de aprendizagem. Assim, as lacunas de pesquisa abordadas neste trabalho de doutorado podem ser descritas como segue.

I. **Uso de bases teóricas**: os estudos analisados apresentam abordagens que não estão explicitamente fundamentadas nas teorias de aprendizagem, conforme descrito na subseção 3.4.1.

II. **Combinação de métodos analíticos**: os estudos apresentam abordagens que não associam métodos ou técnicas de análise das interações (dados estruturados) e do discurso (dados não estruturados) dos participantes, conforme descrito na subseção 3.4.2.

III. **Escala da análise de dados**: uma limitação comum descrita nos estudos que investigam contextos de aprendizagem social *online* é o tamanho (ou escala) do conjunto de dados analisado. O número relativamente pequeno de participantes e interações resulta em um conjunto de dados igualmente pequeno para ser usado como fonte para a abordagem descrita nos estudos (FERREIRA *et al.*, 2020; DASCALU *et al.*, 2018; JAN; VLACHOPOULOS; PARSELL, 2019; SWIECKI; SHAFFER, 2020; SILVA; GIMENES; MALDONADO, 2020a; SILVA; GIMENES; MALDONADO, 2020b).

IV. **Aplicação de métodos estatísticos de redes**: embora o uso de métodos estatísticos que investiguem a ocorrência de efeitos comuns em redes seja utilizado com sucesso em diversas áreas do conhecimento, a sua aplicação em contextos de aprendizagem *online* é relativamente nova (FINCHAM; GAŠEVIĆ; PARDO, 2018; MAMAS *et al.*, 2020).

V. **Abordagens específicas para ambientes informais**: contextos de aprendizagem informal, como as OLC, são sub-representados na área de pesquisa educacional, conforme descrito na seção 3.4.

A partir da revisão de literatura, foram identificados os 21 estudos primários listados na Tabela 3. Após a extração de dados e síntese, as lacunas semânticas identificadas em cada um dos estudos são apresentadas na Tabela 4. A coluna intitulada "**III. Escala**" mostra o número de participantes envolvidos na validação da abordagem. Assim é possível evidenciar a escala da análise de dados dos estudos. Nas demais colunas, um X indica a presença da lacuna de pesquisa.

Tabela 3 – Estudos identificados na revisão de literatura

ID	Estudo	Referência
E01	*What can we learn from Facebook activity? Using social learning analytics to observe new media literacy skills*	Ahn (2013)
E02	*Learning analytics for online discussions: embedded and extracted approaches*	Wise, Zhao e Hausknecht (2014)
E03	*Applying social learning analytics to message boards in online distance learning: a case study*	Hernández-García *et al.* (2015)
E04	*Re-thinking scientific literacy out-of-school: arguing science issues in a niche Facebook application*	Greenhow, Gibbins e Menzer (2015)
E05	*Examining micro-level knowledge sharing discussions in online communities*	Wang *et al.* (2015)

Continua na próxima página

Tabela 3 – continuação da página anterior

ID	Estudo	Referência
E06	*Towards automated content analysis of discussion transcripts: A cognitive presence case*	Kovanović *et al.* (2016)
E07	*Analyzing social media and learning through content and social network analysis: a faceted methodological approach*	Gruzd, Paulin e Haythornthwaite (2016)
E08	*Interactivity in online discussions and learning outcomes*	Kent, Laslo e Rafaeli (2016)
E09	*Epistemic network analysis and topic modeling for chat data from collaborative learning environment*	Cai *et al.* (2017)
E10	*How effective is your facilitation? Group-level analytics of MOOC forums*	Poquet, Dawson e Dowell (2017)
E11	*Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea*	Kim *et al.* (2018)
E12	*Cohesion network analysis of CSCL participation*	Dascalu *et al.* (2018)
E13	*From social ties to network processes: do tie definitions matter?*	Fincham, Gašević e Pardo (2018)
E14	*Fostering student engagement in online discussion through social learning analytics*	Chen *et al.* (2018)
E15	*SENS: network analytics to combine social and cognitive perspectives of collaborative learning*	Gašević *et al.* (2019)
E16	*Comprehensive analysis of discussion forum participation: from speech acts to discussion dynamics and course outcomes*	Joksimovic *et al.* (2019)
E17	*Exploring presence in online learning through three forms of computer-mediated discourse analysis*	Zhu, Herring e Bonk (2019)
E18	*Towards automatic content analysis of social presence in transcripts of online discussions*	Ferreira *et al.* (2020)

Continua na próxima página

Tabela 3 – continuação da página anterior

ID	Estudo	Referência
E19	*Lurkers versus posters: investigation of the participation behaviors in online learning communities*	Speily *et al.* (2020)
E20	*Predicting newcomer integration in online learning communities: automated dialog assessment in blogger communities*	Nistor *et al.* (2020)
E21	*Friendship and support networks among students with disabilities in middle school*	Mamas *et al.* (2020)

Fonte: Elaborada pelo autor.

Tabela 4 – Lacunas de pesquisa identificadas nos estudos

Estudo	I. Teorias	II. Métodos	III. Escala	IV. Mét. de rede	V. Amb. informais
E01	X	X	99	X	
E02		X	N.E.	X	X
E03	X	X	656	X	X
E04		X	346	X	
E05	X	X	N.E.	X	
E06			81	X	X
E07	X		467	X	
E08		X	231	X	X
E09	X		844	X	X
E10		X	22791		X
E11		X	284	X	X
E12			47	X	X
E13	X	X	227		X
E14			39	X	X
E15			1989		X
E16			1797		X
E17		X	62	X	X
E18			81	X	X
E19		X	740	X	
E20			2491	X	
E21	X	X	32		X

Fonte: Elaborada pelo autor.

Nota – N.E.: não especificado.

A seguir apresenta-se uma discussão sobre as lacunas de pesquisa evidenciadas nos estudos analisados na RL. Em relação ao **uso de bases teóricas** para concepção das abordagens, um terço dos estudos não associou de forma explícita as teorias educacionais aos resultados apresentados. Entre as bases teóricas utilizadas nos demais estudos destacam-se a aprendizagem

social e teorias relacionadas, como as teorias de capital social, aprendizagem social construtivista e aprendizagem colaborativa apoiada por computador, do inglês *Computer-Supported Collaborative Learning* (CSCL), entre outras.

Sobre a **combinação de métodos analíticos**, 12 estudos (ou 57,14%) não realizaram a combinação entre a análise das interações e análise do conteúdo das discussões. Sete estudos realizaram apenas a análise das interações, sendo que SNA foi a técnica mais utilizada. Dois estudos concentraram-se em investigar somente o conteúdo das discussões. Estes dados estão de acordo com os argumentos apresentados na seção 3.4.2.

Em relação à **escala da análise de dados** nos estudos, é possível examinar a coluna "III. Escala" da Tabela 4 que informa a quantidade de participantes envolvidos das discussões *online* investigadas:

- dois estudos não informaram a quantidade de participantes envolvidos nas discussões *online*;
- 15 estudos possuem menos de 1.000 estudantes envolvidos nas discussões.;
- três estudos possuem entre 1.000 e 2.500 participantes envolvidos;
- por fim, apenas um estudo possui mais de 20.000 estudantes envolvidos nas discussões.

A maioria dos estudos, mais precisamente 71,42%, envolve a análise das discussões de menos de 1.000 participantes; três estudos envolvem uma quantidade intermediária, entre 1.000 e 2.500 participantes; e por fim, apenas um estudo realiza análise de dados em larga escala, em que são investigadas as discussões de mais de 20.000 participantes. Com base nestes dados, pode-se concluir que a maior parte dos estudos analisa os dados produzidos por uma quantidade de participantes relativamente pequena. Três revisões sistemáticas que investigam discussões *online* em contextos de aprendizagem corroboram com esta evidência (CELA; SICILIA; SÁNCHEZ, 2015; DADO; BODEMER, 2017; JAN, 2019), pois a mediana da quantidade de participantes envolvidos nos estudos analisados em tais revisões é inferior a 1.000. A estatística descritiva do tamanho do conjunto de dados dos estudos apresentados na revisão de literatura descrita neste capítulo e nas revisões sistemáticas publicadas por Cela, Sicilia e Sánchez (2015), Dado e Bodemer (2017) e Jan (2019) é mostrada na Tabela 5. Assim, conclui-se que as informações apresentadas sobre a escala da análise de dados estão de acordo com o argumento descrito anteriormente, em que a maioria dos estudos que investigam discussões e interações em contextos de aprendizagem social *online* envolvem um pequeno número de participantes (FERREIRA *et al.*, 2020; DASCALU *et al.*, 2018; JAN; VLACHOPOULOS; PARSELL, 2019; SWIECKI; SHAFFER, 2020; SILVA; GIMENES; MALDONADO, 2020a).

A maioria dos estudos não aplica **métodos estatísticos de redes** para investigar a interação entre os participantes de discussões *online*. A maior parte das abordagens utiliza apenas

Tabela 5 – Estatística descritiva para a quantidade de participantes envolvidos nos estudos analisados na revisão de literatura descrita neste capítulo e em três revisões sistemáticas

Referência	N	Mín.	1º Quartil	Mediana	3º Quartil	Máx.
Revisão de literatura	19*	32	81	284	792	1.797
Cela, Sicilia e Sánchez (2015)	33**	5	20	36	82	120
Dado e Bodemer (2017)	14	16	23	34	76	143
Jan (2019)	12	43	119	838	3.652	7.233

Fonte: Elaborada pelo autor.

Nota – **N** representa a quantidade de estudos analisados.

Observações – * Dois estudos não especificaram a quantidade de participantes envolvidos, portanto, foram desconsiderados.
** Quatro estudos não especificaram a quantidade de participantes envolvidos, portanto, foram desconsiderados.

técnicas como a análise de regressão ou a análise de correlação de *Pearson* para validar suas conclusões. Contudo, os dados obtidos por meio da análise de redes sociais não possuem a premissa da interdependência exigida por essas técnicas, pois os nós e as conexões que determinam o valor das métricas de rede não são estatisticamente independentes (ZHU *et al.*, 2016; MAMAS *et al.*, 2020; ANTONE *et al.*, 2019). Portanto, é necessário utilizar métodos como o ERGM para analisar dados obtidos de redes sociais e testar hipóteses relacionadas a efeitos estruturais de rede, como discutido na seção 2.4.1.3. Nesse contexto, apenas cinco estudos utilizam métodos estatísticos de rede para validar suas conclusões.

Em relação ao contexto dos estudos, apenas seis abordagens analisam as interações e discussões ocorridas em **ambientes informais de aprendizagem**. Apesar do crescente interesse de educadores e pesquisadores em investigar as experiências de aprendizagem que ocorrem em tais ambientes, este contexto permanece sub-representado na área de pesquisa educacional, como discutido na seção 3.4.

3.6 Considerações finais

Neste capítulo, foi apresentado o contexto de avaliação da aprendizagem em ambientes informais e algumas lacunas de pesquisa importantes desta área. A condução de um MS identificou duas lacunas importantes: a falta de bases teóricas da área educacional na concepção dos estudos e a falta de combinação de métodos de análise das interações e do discurso em discussões *online*. Adicionalmente, uma revisão de literatura sobre discussões no contexto da aprendizagem social *online* evidenciou outras três lacunas de pesquisa: a falta de análise de dados em larga escala, a falta de aplicação de métodos estatísticos para redes e a falta de abordagens específicas para investigação dos ambientes informais. No próximo capítulo, é apresentado o processo proposto neste trabalho de doutorado para cobrir tais lacunas de pesquisa.

CAPÍTULO

4

SLIM: UM PROCESSO PARA AVALIAÇÃO DA APRENDIZAGEM SOCIAL EM AMBIENTES INFORMAIS

4.1 Considerações iniciais

A motivação deste trabalho de doutorado consiste em avaliar condições e comportamentos associados à aprendizagem social *online* em OLC estabelecidas em ambientes informais de aprendizagem. Processos da área educacional, baseados em LA, podem orientar a análise de dados produzidos em larga escala pelos participantes de ambientes informais. Neste contexto, esta seção visa a descrever o processo SLIM. Ele define um conjunto de atividades e diretrizes cujo objetivo é guiar a análise de dados na avaliação da aprendizagem social em OLC. Inicialmente, a seção 4.2 oferece uma visão geral do processo. Na sequência, a seção 4.3 descreve as atividades do processo SLIM. Por fim, a seção 4.4 faz as considerações finais do capítulo.

4.2 Visão geral

SLIM baseia-se no modelo de referência para LA proposto por Chatti *et al.* (2012) e no processo analítico para investigação de dados educacionais sugerido por Chatti *et al.* (2012) e Keim *et al.* (2008), ambos descritos na seção 2.3. O processo SLIM é composto por um conjunto de atividades criadas com base nas lacunas de pesquisa evidenciadas no capítulo 3. A Figura 10 apresenta os conceitos-chave do processo de acordo com as quatro dimensões do modelo de referência para LA (CHATTI *et al.*, 2012). Tais conceitos são descritos nas próximas subseções.

Figura 10 – Descrição dos conceitos-chave do processo SLIM

Fonte: Elaborada pelo autor.

4.2.1 Dimensão "O quê?"

Esta dimensão descreve o espaço de métricas que apoia o processo SLIM e o conjunto de dados que são analisados ao longo da sua execução.

4.2.1.1 O espaço de métricas

As métricas representam os aspectos da OLC que serão mensurados durante as atividades do processo SLIM. Elas foram propostas com base nas diretrizes sugeridas pelos *frameworks* teóricos de criação de valor (WENGER; TRAYNER; De Laat, 2011) e na dimensão presença social das comunidades de inquérito (GARRISON; ANDERSON; ARCHER, 1999), descritos na seção 2.5. Tais *frameworks* sugerem um conjunto de indicadores que permitem avaliar as interações e discussões em comunidades *online*. Contudo, seus autores não oferecem detalhes sobre como utilizá-los na prática. Assim, as métricas propostas nesta subseção são usadas para operacionalizar a avaliação da aprendizagem social, conforme descrito a seguir.

- **Métricas estruturadas**: são calculadas com base nas estruturas de dados que representam os tópicos de discussões assíncronas criados nas OLC, conforme apresentado na Figura 2. Estas métricas podem ser de dois tipos: ***(i)*** as **métricas de SNA** avaliam a rede social e as conexões relacionais formadas a partir das interações dos participantes; elas são calculadas com base na rede que representa tais interações, por meio do método de SNA,

conforme descrito na seção 2.4.1; e ***(ii)*** as **métricas de estrutura** avaliam os padrões de comunicação evidenciados a partir da análise da estrutura das discussões, conforme descrito na subseção 2.4.1.2.

- **Métricas não estruturadas** ou **métricas do discurso**: apoiam a compreensão das dimensões sociais do texto gerado pelos participantes das OLC durante as discussões; elas são calculadas por meio do *framework* LIWC e do método de análise do discurso, como apresentado na seção 2.4.2.

A Tabela 6 apresenta as métricas estruturadas que apoiam o processo SLIM. As colunas Indicadores e Elementos são baseadas nas diretrizes para avaliação de comunidades e redes sugeridas pelo *framework* de criação de valor, conforme apresentado na subseção 2.5.1. A coluna Origem informa a estrutura de dados utilizada para o cálculo da métrica, de acordo com a representação na Figura 2.

A Tabela 7 apresenta as métricas do discurso que apoiam o processo SLIM. O objetivo de tais métricas é evidenciar a dimensão presença social do *framework* teórico das comunidades de inquérito, conforme descrito na seção 2.5.2. Na primeira coluna da tabela são apresentadas as três categorias da presença social. A coluna Métrica apresenta os atributos do *framework* LIWC, e a coluna Referências descreve os estudos que evidenciam a importância do uso da referida métrica para identificação da presença social em discussões *online*.

4.2.1.2 Definição do formato dos dados de aprendizagem

O processo SLIM analisa os dados gerados pelos usuários em ambientes informais de aprendizagem. Tais dados compreendem informações relacionadas aos participantes, aos tópicos de discussão, aos comentários publicados como resposta aos tópicos de discussão e ao processo de avaliação pelos pares executado durante a interação entre os participantes. Os dados analisados ao longo do processo SLIM são especificados no formato *JavaScript Object Notation* (JSON). A seguir é apresentada uma descrição sintetizada de tais dados.

- **Dados dos participantes**
  - *Id*: número inteiro auto-incremental que identifica o participante de forma única.
  - *Label*: rótulo textual que identifica o participante de forma única na comunidade *online*.
  - *Threads*: número inteiro que representa a quantidade de tópicos de discussão criada pelo participante.
  - *Reputation*: número inteiro que representa a reputação do participante, de acordo com o sistema de avaliação pelos pares.

Tabela 6 – Métricas estruturadas

Indicadores	Elementos	Origem	Métrica*
Nível de participação	Características dos participantes	Rede	Grau de entrada
		Rede	Grau de saída
		Rede	Centralidade de intermediação
		Rede	Centralidade de proximidade
		Rede	Centralidade de autovetor
		Rede	Excentricidade
		Rede	*HITS Hub*
		Rede	*HITS Authority*
		Rede	*PageRank*
	Participantes ativos	Árvore	Número de participantes
		Árvore	Número de discussões
Nível de atividade	Quantidade de respostas	Árvore	Tamanho das discussões
	Tempos das Respostas	Árvore	Tempo da primeira resposta
Nível de engajamento	Intensidade das discussões	Rede	Densidade
	Comprimento das discussões	Árvore	Comprimento
		Árvore	Profundidade
	Reciprocidade	Rede	Coeficiente de reciprocidade
	Intensidade do engajamento	Árvore	Intensidade das discussões
	Duração do engajamento	Árvore	Duração do engajamento
Qualidade das interações	Qualidade das discussões	Árvore	*Score* das respostas
Formato estrutural da rede	Componentes	Rede	Número de componentes
	Coesão estrutural	Rede	Caminho médio
		Rede	Coeficiente de agrupamento
		Rede	Diâmetro
		Rede	Triângulos
		Rede	Transitividade
	Gargalos	Rede	Número de gargalos

Fonte: Adaptada de Silva, Gimenes e Maldonado (2020b).

Nota – * A definição das métricas e a referência dos estudos que as utilizam para avaliação em contextos de aprendizagem são apresentadas nas seções 2.4.1.1 (métricas da origem Rede) e 2.4.1.2 (métricas da origem Árvore).

Tabela 7 – Métricas do discurso

Categoria	Métrica	Descrição	Referências
Afetiva	liwc.pronoun	Número de pronomes	Ferreira *et al.* (2020)
	liwc.ppron	Pronomes pessoais	Ferreira *et al.* (2020) Lin, Yu e Dowell (2020)
	liwc.i	Pron. 1ª pessoa do singular	Lin, Yu e Dowell (2020) Zhu, Herring e Bonk (2019)
	liwc.we	Pron. 1ª pessoa do plural	Lin, Yu e Dowell (2020) Barbosa *et al.* (2021)
	liwc.ipron	Pronomes impessoais	Ferreira *et al.* (2020)
	liwc.affect	Processos afetivos	Ferreira *et al.* (2020)
	liwc.posemo	Emoção positiva	Ferreira *et al.* (2020) Zhu, Herring e Bonk (2019)
	liwc.negemo	Emoção negativa	Ferreira *et al.* (2020) Zhu, Herring e Bonk (2019)
	liwc.work	Relação com trabalho	Ferreira *et al.* (2020)
	liwc.power	Relação com poder	Ferreira *et al.* (2020)
	liwc.drives	Direcionamentos	Ferreira *et al.* (2020)
	liwc.percept	Processos perceptivos	Ferreira *et al.* (2020)
	liwc.negate	Negações	Ferreira *et al.* (2020) Barbosa *et al.* (2021)
Interativa	liwc.interrog	Interrogações	Ferreira *et al.* (2020) Barbosa *et al.* (2021)
	liwc.focuspresent	Foco no presente	Ferreira *et al.* (2020)
	liwc.auxverb	Verbos auxiliares	Ferreira *et al.* (2020) Barbosa *et al.* (2021)
	liwc.you	Pronomes em 2ª pessoa	Ferreira *et al.* (2020)
	liwc.assent	Consentimento	Ferreira *et al.* (2020) Barbosa *et al.* (2021)
	liwc.focuspast	Foco no passado	Ferreira *et al.* (2020)
Coesiva	liwc.affiliation	Afiliações	Ferreira *et al.* (2020) Barbosa *et al.* (2021)
	liwc.social	Processos sociais	Lin, Yu e Dowell (2020) Zhu, Herring e Bonk (2019)

Fonte: Adaptada de Silva, Gimenes e Maldonado (2020a).

  - *Metrics*: *array* com as métricas de SNA egocêntricas que avaliam o comportamento do participantes (métricas do elemento "Características dos participantes" apresentado na Tabela 6).

- **Dados dos tópicos de discussão**

  - *Id*: rótulo textual que identifica o tópico de discussão de forma única na comunidade *online*.
  - *User*: participante criador do tópico de discussão.
  - *Date creation*: data e hora de criação do tópico de discussão.

  - *Last reply*: data e hora da última última resposta do tópico de discussão.
  - *Title*: título do tópico de discussão.
  - *Text*: texto do tópico de discussão;
  - *URL*: *Uniform Resource Locator* (URL) do tópico de discussão.
  - *WC*: quantidade de palavras do texto do tópico de discussão.
  - *Score*: pontuação do tópico de discussão, de acordo com o sistema de avaliação pelos pares.
  - *Positive points ratio*: proporção dos pontos positivos atribuídos ao tópico de discussão, em relação ao total de pontos.
  - *Structured metrics*: *array* com as métricas apresentadas na Tabela 6 (exceto as do elemento "Características dos participantes").
  - *Discourse metrics*: *array* com as métricas apresentadas na Tabela 7, cujo objetivo é avaliar o conteúdo do texto dos comentários que fazem parte do tópico de discussão.

- **Dados dos comentários**
  - *Id*: número inteiro auto-incremental que identifica o comentário de forma única no tópico de discussão.
  - *Discussion Id*: identificador do tópico de discussão ao qual o comentário faz parte.
  - *Date*: Data e hora de criação do comentário.
  - *Source user*: participante criador do comentário.
  - *Target user*: participante criador do comentário ou tópico de discussão ao qual se dirige a resposta.
  - *Score*: pontuação do comentário, de acordo com o sistema de avaliação pelos pares.
  - *WC*: quantidade de palavras do comentário.
  - *Text*: texto do comentário.
  - *Discourse metrics*: *array* com as métricas apresentadas na Tabela 7, cujo objetivo é avaliar o conteúdo do texto do comentário.

4.2.2 Dimensão "Por quê?"

A seção 3.3 descreve o conceito de avaliação no contexto de investigação da qualidade das contribuições dos participantes da comunidade, por meio da análise dos dados gerados no processo de avaliação pelos pares. Nesta subseção, o termo avaliação é descrito no contexto do processo SLIM, como um dos aspectos sob o qual se analisa os dados educacionais coletados. Assim, na dimensão "por quê" são definidos os aspectos avaliação e monitoramento de uma OLC.

- **Avaliação**: ao contrário dos ambientes formais, em que a avaliação está relacionada a mensurar resultados específicos da aprendizagem, avaliar no contexto dos ambientes informais, e portanto do processo SLIM, significa analisar e identificar um conjunto de condições e padrões comportamentais associados à aprendizagem social (GRUZD; PAULIN; HAYTHORNTHWAITE, 2016; SILVA; GIMENES; MALDONADO, 2020b; SILVA; GIMENES; MALDONADO, 2020a), que podem ser descritos da seguinte forma:
  - as **condições** associadas à aprendizagem social são identificadas por meio das métricas de SNA (subseção 2.4.1.1), métricas de estrutura (subseção 2.4.1.2) e métricas do discurso que evidenciam a presença social em discussões *online* (subseção 2.5.2);
  - os **padrões comportamentais** associados à aprendizagem social são reconhecidos por meio de modelos (descritos na subseção 4.2.4.2) e efeitos estruturais proeminentes em redes sociais (subseção 2.4.1.3).
- **Monitoramento**: uma necessidade comumente descrita em contextos de investigação de ambientes informais de aprendizagem refere-se ao distanciamento de análises estáticas que observam uma OLC em um ponto específico no tempo, e a busca por abordagens que monitoram a dinâmica de tais comunidades ao longo do tempo (SCHREURS; De Laat, 2014). Neste cenário, o conceito de monitoramento se refere à análise de períodos maiores que um ano (HUDGINS *et al.*, 2020), com o objetivo de investigar a dinâmica temporal do comportamento dos participantes da OLC.

4.2.3 Dimensão "Quem?"

Esta dimensão descreve os atores envolvidos ou partes interessadas na análise e no resultado das atividades do processo SLIM. Tais atores podem ser descritos como segue.

- **Analistas de dados educacionais:** estes profissionais estão interessados nas conclusões obtidas por meio da análise dos dados e nos resultados das atividades do processo SLIM, cujo objetivo é gerar novos conhecimentos que ampliem o entendimento da aprendizagem social *online* no contexto dos ambientes informais (GARCÍA-PEÑALVO; COLOMO-PALACIOS; LYTRAS, 2012; GREENHOW; GIBBINS; MENZER, 2015; GALANIS *et al.*, 2016; SPEILY *et al.*, 2020).
- **Moderadores ou administradores das OLC:** possuem interesse na análise temporal e na evolução da dinâmica dos participantes das comunidades ao longo do tempo (SCHREURS; De Laat, 2014).
- **Participantes das OLC:** possuem interesse em conhecer o nível de atividade da OLC. A conscientização sobre o grau de interação e colaboração existente entre os participantes e seus pares pode levar à autorregulação da interação social e do compartilhamento de

conhecimento em comunidades *online* (JOKSIMOVIĆ *et al.*, 2015; CHEN *et al.*, 2018; SILVA; GIMENES; MALDONADO, 2020a).

4.2.4 Dimensão "Como?"

Esta dimensão descreve os métodos e modelos que definem como o processo SLIM realiza a análise dos dados de aprendizagem coletados, conforme descrito nas subseções seguintes.

4.2.4.1 Métodos de análise

SLIM combina os dois métodos de análise detalhados na seção 2.4, com outros dois métodos analíticos: ***(i)*** análise de dados visual e exploratória; e ***(ii)*** método estatístico para redes. Esses quatro métodos podem ser descritos como segue.

- **Análise de redes sociais (SNA)**: método baseado na teoria dos grafos que visa a investigar os processos de formação de uma rede social e as propriedades das suas conexões relacionais, com o objetivo de compreender como a criação e manutenção de tal rede pode apoiar a aprendizagem.
- **Análise do discurso**: método que oferece técnicas para compreender a grande quantidade de texto gerado em ambientes de aprendizagem, com o objetivo de apoiar o entendimento das dimensões sociais e cognitivas das interações *online*.
- **Análise de dados visual e exploratória**: este método propõe a investigação e análise visual dos dados por meio de gráficos, com o objetivo de identificar padrões e tendências, detectar anomalias (*outliers*) e observar a validade de hipóteses propostas com base nos dados (LANGER; MEISEN, 2021). Assim, o método apoia o analista de dados educacionais a inferir conclusões sobre as características dos dados coletados e, especialmente, sobre o domínio associado a eles, no caso deste trabalho de doutorado, a aprendizagem social *online*. A análise visual e exploratória não mostra, necessariamente, resultados com significância estatística proeminentes do conjunto de dados. Contudo, a visualização dos dados de forma gráfica permite criar novos *insights*, construídos a partir da perspectiva de um especialista no domínio (MA *et al.*, 2017).
- **Método estatístico para redes**: este método descreve relações entre variáveis, revela características importantes e identifica padrões que ocorrem em uma rede. Seu objetivo é comprovar hipóteses geradas a partir dos dados obtidos da rede de interações.

4.2.4.2 Criação de modelos

No contexto do processo SLIM, um modelo pode ser definido como o resultado da aplicação de um algoritmo de aprendizagem de máquina, do inglês *Machine Learning* (ML), cujo objetivo é reconhecer padrões inerentes ou ocultos em conjuntos de dados (SCHELTER *et al.*,

2018). Neste trabalho de doutorado, os modelos de ML são criados com o intuito de identificar padrões comportamentais presentes nos dados educacionais coletados. Um modelo pode ser formado por uma ou mais estruturas de dados, como conjuntos de regras, e vetores ou matrizes de coeficientes que representam os padrões reconhecidos. Os modelos que apoiam o processo SLIM são divididos em duas categorias que podem ser definidas como descrito a seguir.

- **Modelos seccionais (ou transversais)**: apoiam a identificação de padrões comportamentais no aspecto Avaliação do processo SLIM (apresentado na subseção 4.2.2). Os modelos seccionais analisam todo o conjunto de dados, ou representam determinada característica de uma entidade em um ponto específico do tempo, ou seja, a variável *tempo* não desempenha um papel significativo na análise (PAL; PRAKASH, 2017, p. 8). No contexto deste trabalho de doutorado, os modelos seccionais podem ser de dois tipos, descritos a seguir.

  - Modelos de regressão linear: são criados por meio de algoritmos de ML supervisionados, baseados em técnicas de regressão linear. Regressão é o processo de estimar a relação entre variáveis independentes (ou de entrada) e a variável dependente (ou de saída), sendo que tais variáveis são numéricas e de valor contínuo (ARTASANCHEZ; JOSHI, 2020, p. 117).

  - Modelo de clusterização: é criado por meio de algoritmos de ML não supervisionados, baseados em técnicas de clusterização ou agrupamento. Tal modelo é usado para analisar dados e reconhecer *clusters* (ou grupos) ocultos nesses dados, com o objetivo de formar subgrupos cujos elementos são semelhantes entre si. Para identificar tais *clusters* são usadas métricas de similaridade, como a distância euclidiana (ARTASANCHEZ; JOSHI, 2020, p. 154).

- **Modelos temporais**: apoiam o aspecto Monitoramento do processo SLIM (apresentado na subseção 4.2.2). Os modelos temporais são compostos de inúmeras observações quantitativas sobre uma ou mais características mensuráveis de uma entidade individual, tomadas em pontos diferentes no tempo (PAL; PRAKASH, 2017, p. 16). Tais modelos são representados por meio do componente de tendências das séries temporais, cujo objetivo é capturar a tendência de longo prazo em conjuntos de dados organizados em ordem cronológica (PAL; PRAKASH, 2017, p. 23).

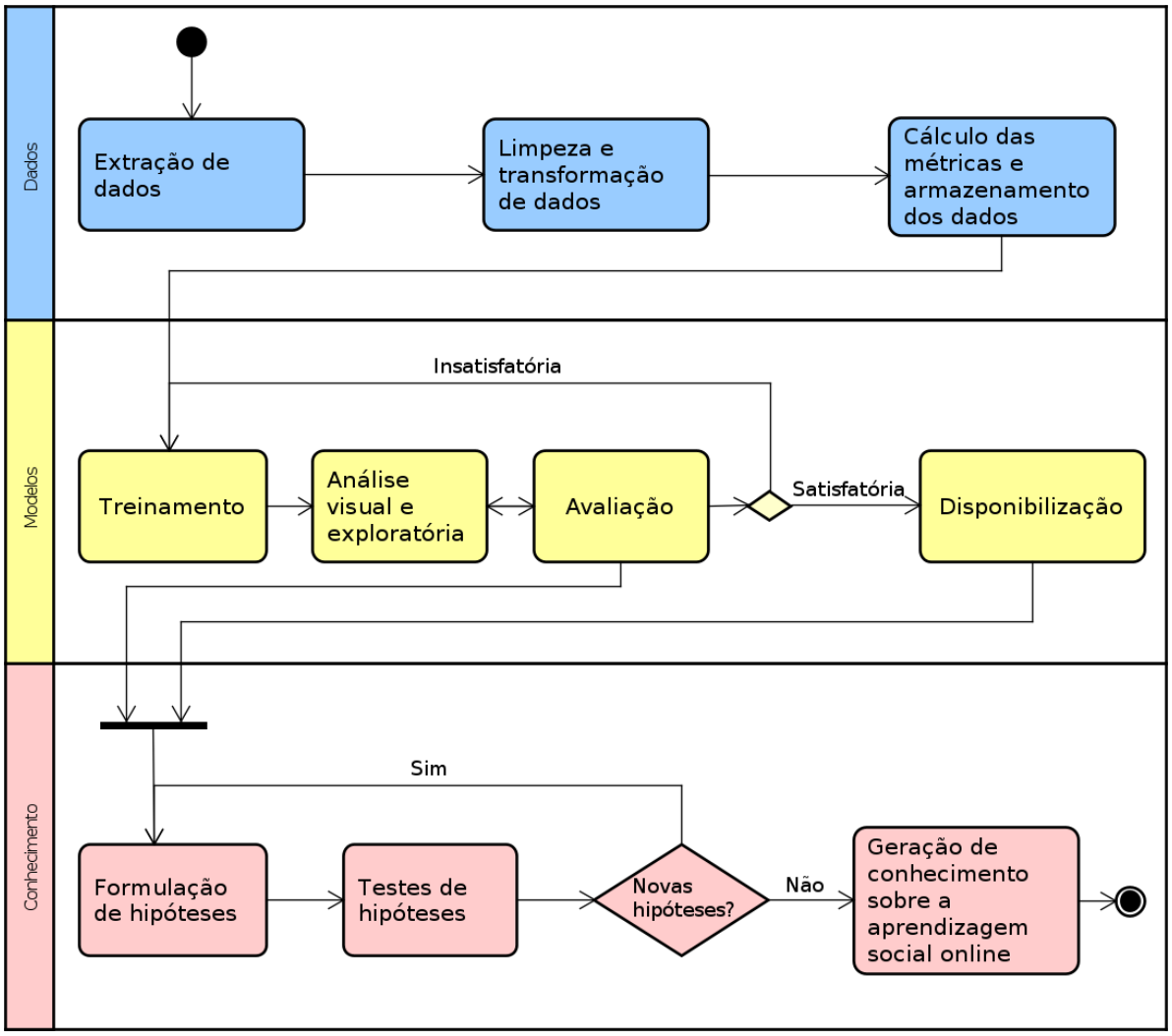
Além dos modelos de ML, o processo SLIM utiliza o ERGM, apresentado na subseção 2.4.1.3, como modelo de probabilidade para criar inferências generalizáveis sobre os padrões de comportamento de dados organizados em redes, apoiando assim o aspecto avaliação do processo SLIM.

4.3 Atividades do processo SLIM

A Figura 11 mostra um diagrama de atividades que representa o processo SLIM. As atividades são divididas em três partições que podem ser descritas como segue.

- A partição **Dados** contém as atividades relacionadas à extração, limpeza e transformação, cálculo das métricas e armazenamento dos dados.
- A partição **Modelos** contém as atividades relacionadas ao treinamento, do inglês *fitting*, avaliação e disponibilização de modelos.
- A partição **Conhecimento** contém as atividades relacionadas à geração de conhecimento sobre aprendizagem social *online* em ambientes informais.

Figura 11 – Representação do processo SLIM por meio de um diagrama de atividades



Fonte: Elaborada pelo autor.

As próximas subseções descrevem com mais detalhes as atividades do processo SLIM.

4.3.1 Atividades relacionadas aos dados

Esta subseção descreve as atividades relacionadas aos dados. Ao final da última atividade, os dados coletados e transformados são armazenados para que eles possam ser utilizados nas atividades subsequentes.

4.3.1.1 Extração de dados

A extração de dados é a primeira atividade do processo SLIM. Ela se refere à coleta e disponibilização dos dados das comunidades de aprendizagem para que eles sejam utilizados ao longo da execução do processo. A extração não realiza qualquer tipo de processamento com os dados, portanto, nesta atividade, eles são chamados de dados educacionais brutos. Diversas técnicas de extração podem ser utilizadas, como a leitura de grandes arquivos de dados, *web scrapping* (SINGRODIA; MITRA; PAUL, 2019) e a utilização de Interfaces de Programação de Aplicações, do inglês *Application Programming Interfaces* (API). A avaliação de comunidades de aprendizagem hospedadas em plataformas tecnológicas, como as redes sociais, pode utilizar as API disponibilizadas por tais plataformas, pois elas possibilitam a extração de dados em formatos hierárquicos e semi-estruturados como o JSON. A seguir descreve-se a especificação da atividade extração de dados.

- **Entrada**: URL da comunidade de aprendizagem, data inicial e data final da extração de dados.
- **Diretrizes**:
  - **D.1.1** conectar-se à plataforma tecnológica da comunidade de aprendizagem ou à fonte de dados por meio da URL especificada na entrada da atividade;
  - **D.1.2** extrair e armazenar em uma coleção de documentos JSON os **dados dos tópicos de discussão** criados na comunidade *online* entre a data inicial e a data final especificada na entrada da atividade;
  - **D.1.3** extrair e armazenar em uma coleção de documentos JSON os **dados dos comentários** dos tópicos de discussão extraídos na diretriz D.1.2;
  - **D.1.4** extrair e armazenar em uma coleção de documentos JSON os **dados dos participantes** envolvidos nos tópicos de discussão e comentários extraídos nas diretrizes D.1.2 e D.1.3.
- **Saída**: três coleções de documentos no formato JSON com os dados dos participantes, tópicos de discussão e dos comentários, de acordo o formato apresentado na subseção 4.2.1.2. Os dados referentes às métricas não devem ser calculados, pois eles são processados na atividade de cálculo das métricas.

4.3.1.2 Limpeza e transformação de dados

Dados coletados da Internet, na maioria dos casos, estão em formato bruto (PENTREATH, 2015). Assim, a atividade de limpeza e transformação possui dois objetivos principais: ***(i)*** pré-processar os dados de modo a identificar e corrigir possíveis inconsistências; e ***(ii)*** transformar os dados para uma representação que facilite a execução das atividades subsequentes do processo. Com base nestes conceitos, a seguir descreve-se a especificação da atividade de limpeza e transformação de dados.

- **Entrada**: coleções de documentos JSON com os dados brutos produzidos na atividade de extração.
- **Diretrizes**:
  - **D.2.1** identificar dados faltantes ou inconsistentes; dados faltantes no formato numérico devem ser preenchidos com zero; dados faltantes no formato texto devem ser preenchidos com uma *string* vazia; dados inconsistentes devem ser removidos;
  - **D.2.2** identificar os *outliers* dos atributos *Reputation* dos participantes e *Score* dos tópicos de discussão; de acordo com os argumentos descritos por Aguinis, Gottfredson e Joo (2013), tais valores devem ser identificados (e não removidos) por meio da técnica de *box plot*, pois eles podem influenciar a atividade de treinamento e avaliação dos modelos;
  - **D.2.3** excluir os tópicos de discussão que não possuem comentários (respostas), pois eles podem influenciar a atividade de treinamento e avaliação dos modelos;
  - **D.2.4** realizar a normalização e correção ortográfica das palavras nos atributos *Text* referentes aos dados dos tópicos de discussão e comentários; o objetivo é minimizar erros de digitação cometidos pelos participantes.
- **Saída**: três coleções de documentos no formato JSON com os dados educacionais transformados.

4.3.1.3 Cálculo das métricas e armazenamento dos dados

O objetivo desta atividade é calcular e armazenar as métricas que apoiam o processo SLIM. Sua especificação pode ser descrita como segue.

- **Entrada**: coleções de documentos JSON com os dados educacionais processados na atividade de limpeza e transformação.
- **Diretrizes**:

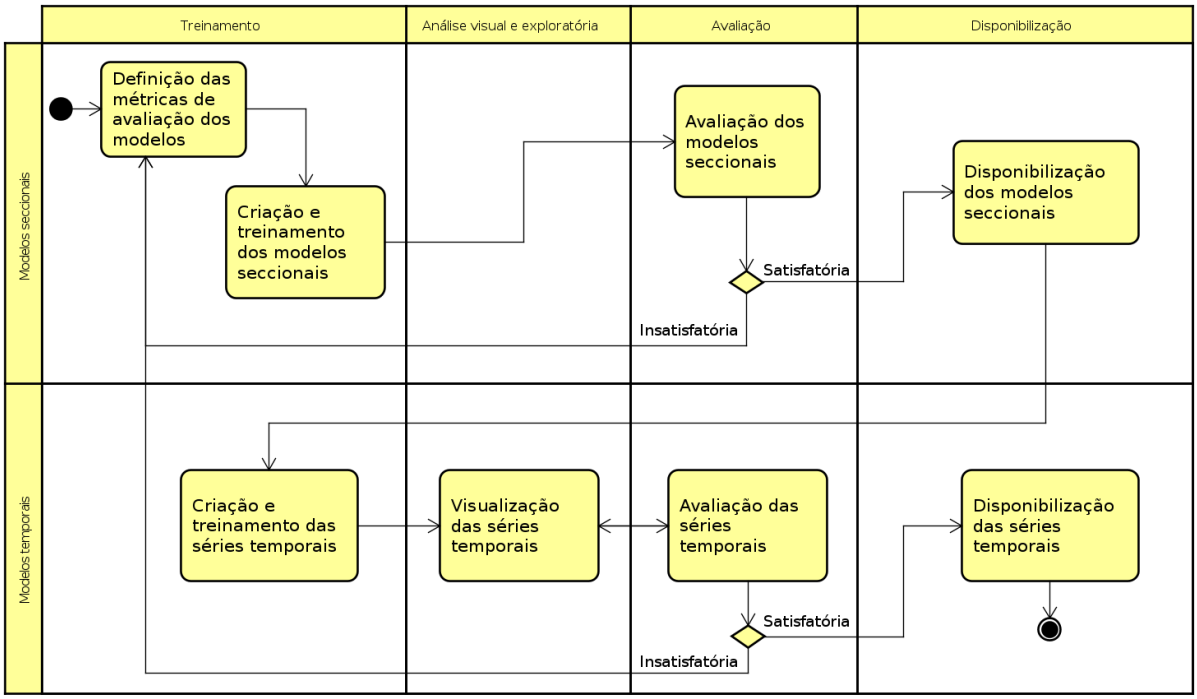
- **D.3.1** com base nos atributos *Source user* e *Target user* da coleção de documentos JSON com os comentários publicados pelos participantes, criar a estrutura de dados do tipo rede que representa todas as interações ocorridas no período sob análise (grafo direcionado conforme apresentado na Figura 2);
- **D.3.2** com base na rede criada na diretriz D.3.1, calcular as métricas de SNA egocêntricas que avaliam o comportamento dos participantes (métricas do elemento "Características dos participantes" apresentadas na Tabela 6); elas devem ser armazenadas no atributo *Metrics* da coleção de documentos JSON referente aos participantes;
- **D.3.3** com base no atributo *Text* da coleção de documentos JSON com os comentários publicados pelos participantes, calcular as métricas do discurso apresentadas na Tabela 7; elas devem ser armazenadas no atributo *Discourse metrics*.
- **D.3.4** com base na coleção de documentos JSON que contêm os tópicos de discussão extraídos da comunidade, criar as estruturas de dados do tipo Árvore e Rede para cada tópico de discussão;
- **D.3.5** a partir das estruturas de dados criadas na diretriz D.3.4, calcular as métricas estruturadas apresentadas na Tabela 6 (exceto as métricas de SNA egocêntricas, calculadas na diretriz D.3.2); elas devem ser armazenadas no atributo *Structured metrics*;
- **D.3.6** para cada tópico de discussão, recuperar os comentários pertencentes ao mesmo; calcular a média de cada métrica do discurso dos comentários pertencentes ao tópico; este valor agregado (média) deve ser armazenado no *array Discourse metrics* do tópico de discussão;
- **D.3.7** armazenar as coleções de documentos JSON que representam os dados e métricas dos participantes, tópicos de discussão e comentários em um banco de dados não-relacional.

• **Saída**: três coleções de documentos no formato JSON com os dados processados e métricas calculadas, armazenados em um banco de dados não-relacional.

4.3.2 Atividades relacionadas aos modelos

A partição denominada Modelos do diagrama apresentado na Figura 11 representa as atividades responsáveis pela criação, treinamento, avaliação e disponibilização dos modelos de ML que reconhecem os padrões existentes nos dados. A Figura 12 mostra um diagrama com o detalhamento de tais atividades. Cada atividade da partição Modelos na Figura 11 é representada por uma partição vertical na Figura 12, e as partições horizontais representam os modelos seccionais e temporais, respectivamente, conforme descrito na seção 4.2.4.2. As atividades relacionadas aos modelos são descritas nas próximas subseções.

Figura 12 – Diagrama que apresenta o detalhamento das atividades relacionadas aos modelos



Fonte: Elaborada pelo autor.

4.3.2.1 Definição das métricas de avaliação dos modelos

Esta atividade visa a definir as métricas de avaliação que serão utilizadas para analisar a qualidade dos modelos criados. Os **modelos seccionais** podem ser modelos de regressão linear ou de clusterização; os **modelos temporais** são representados pelas séries temporais. A seguir, descreve-se a especificação da atividade.

- **Diretrizes**:
  - **D.4.1** definir as métricas de avaliação dos modelos de regressão linear da seguinte forma:
    - **D.4.1.1** métrica: coeficiente de determinação ajustado (*adjusted R-square*); valor: maior ou igual a $0,75$ (LANTZ, 2019, p. 193);
    - **D.4.1.2** métrica: coeficiente de correlação de *Pearson*; valor: menor ou igual a $-0,5$ ou maior ou igual a $0,5$ (LANTZ, 2019, p. 176);
    - **D.4.1.3** métrica: significância estatística (*p-value*); valor: menor que $0,05$;
  - **D.4.2** definir a métrica de avaliação do modelo de clusterização da seguinte forma:
    - métrica: coeficiente de silhueta (*silhouette coefficient*); valor: maior que 0 (os valores possíveis estão no intervalo $[-1,1]$), pois valores positivos indicam

que provavelmente o algoritmo criou uma divisão satisfatória entre os *clusters* (KASSAMBARA, 2017, p. 141);

- **D.4.3** definir as métricas de avaliação dos modelos temporais da seguinte forma:
  - **D.4.3.1** métrica: coeficiente do teste *Augmented Dickey-Fuller* (ADF); valor: menor que 0, pois valores negativos indicam que a série temporal é estacionária (PAL; PRAKASH, 2017, p. 65);
  - **D.4.3.2** métrica: significância estatística (*p-value*); valor: menor que $0,05$;

- **Saída**: dados tabulados com as métricas de avaliação dos modelos.

4.3.2.2 Criação e treinamento dos modelos seccionais

Esta atividade visa a criar e treinar os modelos seccionais, com o objetivo de reconhecer os padrões presentes nos dados educacionais processados. Os modelos podem ser treinados e avaliados iterativamente nas atividades subsequentes, com diferentes atributos de entrada e hiper-parâmetros, até que as métricas de avaliação atinjam os valores determinados na atividade de definição das métricas. A atividade de criação e treinamento é responsável pela produção de três modelos seccionais: ***(i)*** modelo dos participantes especialistas; ***(ii)*** modelos dos tópicos de discussão bem avaliados; e ***(iii)*** modelo do estilo de discurso dos participantes. A especificação para criação e treinamento de tais modelos pode ser descrita como segue.

(i) Modelo dos participantes especialistas

- **Objetivo**: identificar as métricas de SNA egocêntricas fortemente associadas à *expertise* (atributo *Reputation*) dos participantes.

- **Tipo do modelo**: regressão linear múltipla.

- **Entrada**: coleção de documentos no formato JSON com os dados processados e métricas dos participantes.

- **Diretrizes**:
  - **D.5.1.1** definir o atributo *Reputation* como a variável dependente do modelo;
  - **D.5.1.2** definir os atributos do *array Metrics* (métricas de SNA egocêntricas) como as variáveis independentes do modelo;
  - **D.5.1.3** treinar o modelo com os atributos definidos anteriormente.

- **Saída**: modelo treinado dos participantes especialistas.

(ii) Modelo dos tópicos de discussão bem avaliados

- **Objetivo**: identificar as métricas estruturadas fortemente associadas aos tópicos de discussão mais bem avaliados pelos participantes (atributo *Score*).
- **Tipo do modelo**: regressão linear múltipla.
- **Entrada**: coleção de documentos no formato JSON com os dados educacionais processados e métricas dos tópicos de discussão.
- **Diretrizes**:
  - **D.5.2.1** definir o atributo *Score* como a variável dependente do modelo;
  - **D.5.2.2** definir os atributos do *array Structured metrics* como as variáveis independentes do modelo;
  - **D.5.2.3** treinar o modelo com os atributos definidos anteriormente.
- **Saída**: modelo treinado dos tópicos de discussão bem avaliados.

(iii) Modelo do estilo de discurso dos participantes

- **Objetivo**: identificar as métricas do discurso associadas aos tópicos de discussão mais bem avaliados pelos participantes.
- **Tipo do modelo**: clusterização por meio do algoritmo de ML *Kmeans* (HARTIGAN; WONG, 1979) ou uma de suas variantes.
- **Entrada**: coleção de documentos no formato JSON com os dados processados e métricas dos tópicos de discussão.
- **Diretrizes**:
  - **D.5.3.1** encontrar o valor adequado do hiper-parâmetro k, que determina número de *clusters* usado no algoritmo, por meio do método *Elbow* (BHOLOWALIA; KUMAR, 2014);
  - **D.5.3.2** definir o hiper-parâmetro *número máximo de iterações* com o valor 10 (KASSAMBARA, 2017, p. 39);
  - **D.5.3.3** definir o hiper-parâmetro *número de partições aleatórias iniciais* com o valor 50 (KASSAMBARA, 2017, p. 41);
  - **D.5.3.4** criar uma matriz com as variáveis do atributo *Discourse metrics* dos tópicos de discussão; tais variáveis são equivalentes às variáveis independentes do modelo;

  - **D.5.3.5** treinar o modelo com os hiper-parâmetros definidos e a matriz criada na diretriz D.5.3.4;
  - **D.5.3.6** atribuir para cada tópico de discussão o número do seu *cluster* correspondente, de acordo com o treinamento realizado na diretriz D.5.3.5.

- **Saída**: modelo treinado do discurso dos participantes.

4.3.2.3 Avaliação dos modelos seccionais

O objetivo desta atividade é avaliar a qualidade dos modelos seccionais criados. A especificação da atividade de avaliação dos modelos seccionais é descrita como segue.

- **Entrada**: tabela criada na atividade de definição das métricas de avaliação dos modelos.
- **Diretrizes**:
  - **D.6.1** criar uma tabela com o resultado da avaliação dos modelos seccionais;
  - **D.6.2** comparar as métricas do resultado da avaliação com os valores definidos na tabela descrita na entrada da atividade;
  - **D.6.3** se a qualidade dos modelos for satisfatória:
    * **D.6.3.1** identificar as métricas estruturadas (variáveis independentes) significativas dos modelos de regressão linear por meio dos valores do coeficiente de correlação de *Pearson* e *p-value*, definidos nas diretrizes D.4.1.2 e D.4.1.3;
    * **D.6.3.2** identificar o *cluster* mais representativo do modelo do estilo de discurso dos participantes; tal *cluster* possui os valores médios mais altos para as métricas significativas identificadas na diretriz D.6.3.1;
    * **D.6.3.3** utilizar o valor dos centróides (LANTZ, 2019, p. 294) do *cluster* identificado na diretriz D.6.3.2 para determinar as métricas do discurso (variáveis independentes) significativas;
  - **D.6.4** se a qualidade dos modelos não for satisfatória, retornar à atividade de definição das métricas de avaliação dos modelos seccionais; tais métricas podem ser redefinidas, caso necessário.
- **Saída**: dados tabulados com o resultado da avaliação, contendo: ***(i)*** métricas de avaliação dos modelos seccionais; ***(ii)*** métricas estruturadas significativas dos modelos de regressão linear; ***(iii)*** métricas do discurso significativas do modelo de clusterização.

4.3.2.4 Disponibilização dos modelos seccionais

O objetivo desta atividade é disponibilizar os modelos seccionais para que eles possam ser utilizados pelas atividades relacionadas aos modelos temporais. A especificação da atividade pode ser descrita da seguinte forma.

- **Entrada**: modelos treinados e avaliados que representam os modelos seccionais.
- **Diretriz**:
  - **D.7.1** salvar no formato adequado os modelos que representam os participantes especialistas, os tópicos de discussão bem avaliados e o estilo de discurso dos participantes.
- **Saída**: arquivos que representam os modelos seccionais e identificam as suas variáveis independentes significativas.

4.3.2.5 Criação e treinamento das séries temporais

Esta atividade visa a criar e treinar os modelos de séries temporais, cujo objetivo é reconhecer a dinâmica temporal das comunidades *online*. A atividade não considera o modelo dos participantes especialistas, pois tal modelo não leva em consideração o aspecto temporal. A atividade de criação e treinamento das séries temporais é responsável pela produção de dois modelos: ***(i)*** modelo de séries temporais das métricas estruturadas; e ***(ii)*** modelos de séries temporais das métricas do discurso. A especificação da atividade pode ser descrita como segue.

(i) Modelo de séries temporais das métricas estruturadas

- **Objetivo**: identificar tendências de longo prazo e reconhecer a dinâmica temporal das métricas estruturadas.
- **Tipo do modelo**: séries temporais.
- **Entrada**: ***(i)*** métricas estruturadas (variáveis independentes) significativas dos modelos que representam os tópicos de discussão bem avaliados; e ***(ii)*** coleção de documentos no formato JSON com os dados processados dos tópicos de discussão.
- **Diretrizes**:
  - **D.8.1.1** definir os atributos descritos no item ***(i)*** da **Entrada** da atividade como as variáveis independentes das séries temporais; a variável dependente será representada pelo *tempo*.
  - **D.8.1.2** considerando períodos de análise superiores a um ano, definir o valor de 30 dias para as médias móveis das séries temporais, com o objetivo de obter um efeito de suavização das séries originais e reduzir o efeito de possíveis ruídos dos dados (PAL; PRAKASH, 2017, p. 55);
  - **D.8.1.3** treinar o modelo com os dados processados dos tópicos de discussão.
- **Saída**: componente *tendência* da série temporal que representa a dinâmica das métricas estruturadas ao longo do tempo.

(ii) Modelo de séries temporais das métricas do discurso

- **Objetivo**: identificar tendências de longo prazo e reconhecer a dinâmica temporal das métricas do discurso.
- **Tipo do modelo**: séries temporais.
- **Entrada**: ***(i)*** métricas do discurso (variáveis independentes) significativas dos modelos que representam o estilo do discurso dos participantes; e ***(ii)*** coleção de documentos no formato JSON com os dados processados dos tópicos de discussão.
- **Diretrizes**:
  - **D.8.2.1** definir os atributos descritos no item ***(i)*** da **Entrada** da atividade como variáveis independentes das séries temporais; a variável dependente será representada pelo *tempo*.
  - **D.8.2.2** considerando períodos de análise superiores a um ano, definir o valor de 30 dias para as médias móveis das séries temporais, com o objetivo de obter um efeito de suavização das séries originais e eliminar possíveis ruídos dos dados (PAL; PRAKASH, 2017, p. 55);
  - **D.8.2.3** treinar o modelo com os dados processados dos tópicos de discussão.
- **Saída**: componente *tendência* da série temporal que representa a dinâmica das métricas do discurso ao longo do tempo.

4.3.2.6 Visualização das séries temporais

Esta atividade visa a guiar o uso de técnicas de visualização interativas a fim de ampliar a cognição, identificar padrões, obter *insights* e criar questões de pesquisa específicas que possibilitem uma compreensão mais ampla da aprendizagem social em comunidades *online*. A especificação da atividade de visualização das séries temporais pode ser descrita como segue.

- **Entrada**: componentes de *tendência* das séries temporais que representam a dinâmica das métricas estruturadas e do discurso ao longo do tempo;
- **Diretrizes**:
  - **D.9.1** aplicar um método de padronização nos componentes de tendência das séries temporais, a fim de apresentar os valores das métricas na mesma escala;
  - **D.9.2** disponibilizar uma interface gráfica que possibilite combinar e visualizar diferentes componentes de tendência das séries temporais;
  - **D.9.3** possibilitar a aplicação de filtros para analisar iterativamente diferentes períodos de tempo dos componentes de tendência das séries temporais;

  - **D.9.4** criar questões de pesquisa específicas sobre a aprendizagem social em comunidades *online* com base nos modelos seccionais e temporais analisados.

- **Saída**: questões de pesquisa específicas criadas sobre a aprendizagem social em comunidades *online*.

4.3.2.7 Avaliação das séries temporais

O objetivo desta atividade é avaliar a qualidade das séries temporais criadas. A especificação da atividade é descrita como segue.

- **Entrada**: tabela criada na atividade de definição das métricas de avaliação dos modelos.
- **Diretrizes**:
  - **D.10.1** criar uma tabela com o resultado da avaliação das séries temporais;
  - **D.10.2** comparar as métricas do resultado da avaliação com os valores da tabela descrita na entrada da atividade;
  - **D.10.3** para cada uma das séries temporais (métricas estruturadas e do discurso significativas), analisar se a qualidade do modelo é satisfatória; excluir da atividade de visualização dos modelos as séries temporais cuja qualidade não é satisfatória;
  - **D.10.4** se todas as séries temporais possuírem qualidade não satisfatória, retornar à atividade de definição das métricas de avaliação dos modelos; tais métricas podem ser redefinidas, caso necessário.
- **Saída**: dados tabulados com o resultado da avaliação, contendo: ***(i)*** métricas de avaliação das séries temporais; ***(ii)*** relação das séries temporais das métricas estruturadas com qualidade satisfatória; ***(iii)*** relação das séries temporais das métricas do discurso com qualidade satisfatória.

4.3.2.8 Disponibilização das séries temporais

O objetivo desta atividade é disponibilizar os modelos de séries temporais para que eles possam ser utilizados pelas partes interessadas na avaliação das comunidades de aprendizagem *online*. A especificação da atividade de disponibilização das séries temporais pode ser descrita como segue.

- **Entrada**: modelos treinados que representam as séries temporais.
- **Diretriz**:
  - **D.11.1** salvar no formato adequado as séries temporais que representam a dinâmica ao longo do tempo das métricas estruturadas e do discurso.
- **Saída**: arquivos que representam as séries temporais.

4.3.3 Atividades relacionadas ao conhecimento

A partição denominada Conhecimento do diagrama apresentado na Figura 11 representa as atividades responsáveis pela formulação e teste de hipóteses que visam a gerar conhecimento e proporcionar um entendimento mais amplo da aprendizagem social em ambientes informais. As atividades relacionadas à geração de conhecimento são descritas nas próximas subseções.

4.3.3.1 Formulação de hipóteses

O objetivo desta atividade é formular um conjunto de hipóteses, com base nas questões de pesquisa criadas na atividade de visualização dos dados. A especificação da atividade de formulação de hipóteses pode ser descrita como segue.

- **Entrada**: ***(i)*** dados tabulados dos modelos seccionais e temporais; e ***(ii)*** questões de pesquisa específicas criadas sobre a aprendizagem social em comunidades *online*.
- **Diretriz**:
  - **D.12.1** formular um conjunto de hipóteses com base nos elementos do item **Entrada** da especificação desta atividade.
- **Saída**: conjunto de hipóteses.

4.3.3.2 Testes de hipóteses

O objetivo desta atividade é testar o conjunto de hipóteses criado anteriormente. Elas são testadas por meio de modelos do tipo *Exponential Random Graph* (ERG) descritos na subseção 4.2.4.2. A especificação da atividade de testes de hipóteses pode ser descrita como segue.

- **Entrada**: conjunto de hipóteses.
- **Diretrizes**:
  - **D.13.1** criar um modelo ERG com os parâmetros adequados para testar cada hipótese ou conjunto de hipóteses;
  - **D.13.2** treinar os modelos ERG;
  - **D.13.3** aceitar ou rejeitar as hipóteses com base nos parâmetros dos modelos ERG; uma hipótese é aceita se o coeficiente de um parâmetro possui significância estatística, considerando-se os seguintes critérios (LUKE, 2015, p. 179):
    * o coeficiente do parâmetro possui *p-value*<0.05;
    * o valor absoluto de *z value* do coeficiente é maior ou igual a 2; isso que significa que o coeficiente do parâmetro é, no mínimo, duas vezes maior que o seu erro padrão;

  - **D.13.4** analisar a necessidade de formular novas hipóteses com base nos resultados da diretriz D.13.3; caso afirmativo, retornar à atividade de Formulação de hipóteses.

- **Saída**: tabelas com o resultado do teste de hipóteses.

4.3.3.3 Geração de conhecimento

O objetivo desta atividade é inferir conclusões e gerar conhecimento sobre a aprendizagem social *online* em ambientes informais, com base nos resultados dos testes de hipóteses. A especificação da atividade de geração de conhecimento pode ser descrita como segue.

- **Entrada**: tabelas com o resultado dos testes de hipóteses.
- **Diretriz**:
  - **D.14.1** a partir dos resultados dos testes de hipóteses, inferir conclusões e gerar conhecimento sobre a aprendizagem social *online* em ambientes informais.
- **Saída**: conclusões e conhecimento sobre a aprendizagem social *online* em ambientes informais.

4.4 Considerações finais

Este capítulo apresentou o processo SLIM, a especificação das suas atividades e diretrizes, cujo objetivo é sistematizar a análise de dados educacionais na avaliação de comunidades *online*. SLIM visa a evidenciar como a aprendizagem social *online* ocorre nos ambientes informais de aprendizagem. Assim, o próximo capítulo apresenta a caracterização das duas comunidades que serão utilizadas como estudo de caso para o processo SLIM.

CAPÍTULO

5

CARACTERIZAÇÃO DAS COMUNIDADES DE APRENDIZAGEM

5.1 Considerações iniciais

Este capítulo apresenta a caracterização das comunidades de aprendizagem *online* avaliadas neste trabalho de doutorado, seus participantes, moderadores, nível de atividade e principais temas discutidos. Tal caracterização é importante para reconhecer efetivamente o grupo de usuários investigado como uma comunidade de aprendizagem. São analisadas duas comunidades da rede social Reddit[1]: *learnprogramming*[2] e *MachineLearning*[3]. A rede social Reddit foi escolhida devido ao seu uso massivo em diferentes países do mundo. Por sua vez, as comunidades *learnprogramming* e *MachineLearning* foram selecionadas, pois seus membros se concentram na discussão de dois temas específicos: aprendizagem de programação e conceitos relacionados a *machine learning*, respectivamente. Inicialmente, a seção 5.2 apresenta a rede social Reddit. A seção 5.3 descreve as características gerais das comunidades investigadas. A seção 5.4 analisa os tópicos de discussão, chamados de submissões, e os seus comentários. A seção 5.5 apresenta uma análise das atividades dos participantes. Por fim, a seção 5.6 faz as considerações finais do capítulo.

5.2 A rede social Reddit

Reddit é um site de fóruns com características de redes sociais. É composto por comunidades, também chamadas de *subreddits*, em que os participantes compartilham seus conhecimentos, suas experiências e dúvidas sobre os mais diversos assuntos. Esta rede social permite que os

[1] http://www.reddit.com
[2] http://www.reddit.com/r/learnprogramming
[3] http://www.reddit.com/r/MachineLearning

usuários criem e compartilhem conteúdos em torno de um determinado tema. Em uma espécie de fórum, um participante inicia um tópico de discussão ao publicar uma submissão. Esta submissão pode ser respondida por comentários de outros participantes, ou mesmo por aquele que a criou. Comentários, por sua vez, podem ser respondidos por outros comentários, criando uma hierarquia de respostas. Outra característica importante é que os participantes podem permanecer anônimos, por meio da definição de identidades temporárias, o que os encoraja a discutir sobre os mais diversos assuntos (CAPLAN; TURNER, 2007). Dados de 2020 mostram o uso massivo da rede social Reddit. Ela é constituída por 52 milhões de usuários ativos por dia. Esta métrica é contabilizada por meio dos usuários que efetuam *login* na rede social. Portanto, pessoas que acessam a rede somente para visualizar alguma mensagem ou assistir um vídeo não são consideradas como usuários ativos. Ao todo são 303 milhões de submissões e 2 bilhões de comentários publicados anualmente[4].

No Reddit, os participantes precisam se registrar e efetuar *login* para criar ou responder uma submissão. Os participantes, conhecidos como *redditors*, podem avaliar positivamente ou negativamente as submissões, criando o seu *score*. Além disso, eles também podem associar pontos positivos (*upvote*) ou negativos (*downvote*) às respostas dos outros participantes. Esses pontos são chamadas de *karma*. Tanto o *score* da submissão quanto a pontuação do *karma* fazem parte do sistema de avaliação por pares do Reddit. Em geral, os votos positivos associados a uma submissão indicam a opinião da comunidade sobre ela. Assim, as submissões com as melhores respostas normalmente possuem as pontuações mais altas (HUDGINS *et al.*, 2020). Da mesma forma, a pontuação do *karma* define o nível de atividade e popularidade dos participantes. Portanto, aqueles com as maiores pontuações são reconhecidos como especialistas ou *experts* pela comunidade (SILVA; GIMENES; MALDONADO, 2020a). A Figura 13 apresenta a estrutura de uma submissão publicada na comunidade *learnprogramming* e alguns dos seus principais elementos.

A comunidade *learnprogramming* foi criada em 24 de setembro de 2009 e possuía 3.440.477 participantes inscritos à época da coleta de dados, em março de 2021. Seu objetivo é discutir questões relacionadas à aprendizagem de qualquer linguagem de programação. A comunidade *MachineLearning* foi criada em 29 de julho de 2009 e possuía 1.935.702 participantes inscritos à época da coleta de dados. Seu objetivo é discutir questões relacionadas à aprendizagem de máquina. Ambas as comunidades apresentam inúmeras diretrizes sobre como publicar e responder perguntas da maneira correta. As submissões devem fornecer todas as informações contextuais da questão a ser discutida, na forma de uma boa descrição e um título curto e intuitivo. Existem recomendações para que os participantes verifiquem as mensagens mais antigas antes de publicar uma nova questão. As próximas seções apresentam a caracterização destas comunidades.

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

Figura 13 – Submissão em que um usuário publica uma dúvida sobre a sintaxe da linguagem C

Fonte: Adaptada de REDDIT (2021).

5.3 Características gerais das comunidades

Para a investigação apresentada neste trabalho de doutorado, foram coletados os dados dos participantes, submissões, comentários e informações relacionadas à avaliação por pares, no período entre 01 de Janeiro de 2019 e 31 de Dezembro de 2020. Este intervalo foi considerado pois a análise de longos períodos, superiores a um ano, são capazes de fornecer mais detalhes sobre o engajamento dos participantes, as atividades desenvolvidas e o ciclo de vida das comunidades *online* (HUDGINS *et al.*, 2020).

A Figura 14 apresenta o volume mensal de submissões e comentários em cada uma das comunidades no período investigado. Ao observar a reta pontilhada dos gráficos que representam a tendência de crescimento, observa-se que tanto o número de comentários quanto submissões aumentaram entre 2019 e 2020 em ambas as comunidades.

Figura 14 – Evolução do número de submissões e comentários em cada comunidade

Fonte: Elaborada pelo autor.

A Tabela 8 apresenta o número total e algumas proporções referentes aos dados coletados. A comunidade *learnprogramming* apresenta uma quantidade maior de submissões, participantes únicos ativos e comentários. Contudo, em termos proporcionais, os valores são próximos para

Tabela 8 – Detalhes dos dados coletados nas comunidades entre 01/01/2019 e 31/12/2020

	learnprogramming	***MachineLearning***	**TOTAL**
Submissões	69.447	22.124	91.571
Participantes únicos ativos	95.335	35.702	131.037
Comentários	442.243	152.625	594.868
Participantes / submissões	1,37	1,61	1,43
Comentários / participantes	4,63	4,27	4,53
Comentários / submissões	6,37	6,89	6,49

Fonte: Elaborada pelo autor.

ambas as comunidades. Considera-se um participantes ativo aquele que publicou, no mínimo, uma submissão ou comentário no período investigado. Existe uma intersecção de 2.402 participantes ativos nas duas comunidades. Esse valor corresponde a 2,51% e 6,72% dos participantes ativos das comunidades *learnprogramming* e *MachineLearning*, respectivamente. Neste contexto, considera-se que foram coletados dados educacionais em larga escala, pois a quantidade total de participantes envolvidos, 131.037 considerando a soma dos participantes ativos das comunidades *learnprogramming* e *MachineLearning*, é superior à quantidade apresentada nos estudos da Tabela 5 (seção 3.5).

5.4 Análise das submissões e comentários

Nesta etapa, foram analisadas as submissões e os comentários publicados pelos participantes de cada comunidade em relação ao comprimento do texto (número de palavras) e ao seu *score*. Assim como as submissões, os comentários também podem ser avaliados pelos participantes, portanto, também possuem um *score*. A Tabela 9 apresenta a média, a mediana e o Desvio Padrão (DP) do *score* e do comprimento do texto. É possível observar que o *score* de uma submissão é, em média, 5,04 vezes maior que o *score* de um comentário. Uma possível interpretação é que os autores das submissões recebem mais suporte social do que os autores dos comentários. Outra interpretação possível é que as submissões recebem mais votos positivos pois naturalmente atraem mais atenção que os comentários associados a elas. Além disso, nota-se que os valores do *score* são, na média e mediana, sempre positivos, indicando que geralmente os participantes incentivam uns aos outros com votos positivos, ao invés de inibi-los com votos negativos.

Em relação ao comprimento médio do texto (em quantidade de palavras), é possível observar que na comunidade *learnprogramming* as submissões são aproximadamente duas vezes maiores que os comentários, enquanto na *MachineLearing* a média de quantidade de palavras das submissões e comentários são similares. Além disso, 50% das submissões possuem pelo menos 85 palavras na *learnprogramming*, enquanto na *MachineLearning* a mediana do comprimento do texto é de quatro palavras. O alto valor de DP mostra que há grande variabilidade tanto no *score*

Tabela 9 – Estatística descritiva para o *score* e quantidades de palavras nas comunidades

	learnprogramming			***MachineLearning***		
	Média	Mediana	DP	Média	Mediana	DP
Score da submissão	19,11	1	167,79	24,97	1	118,01
Score dos comentários	3,57	1	22,11	5,28	2	17,05
Palavras da submissão	118,81	85	143,10	64,55	4	136,31
Palavras dos comentários	59,27	35	79,48	61,43	37	81,78

Fonte: Elaborada pelo autor.

quanto no comprimento de submissões e comentários, em ambas as comunidades.

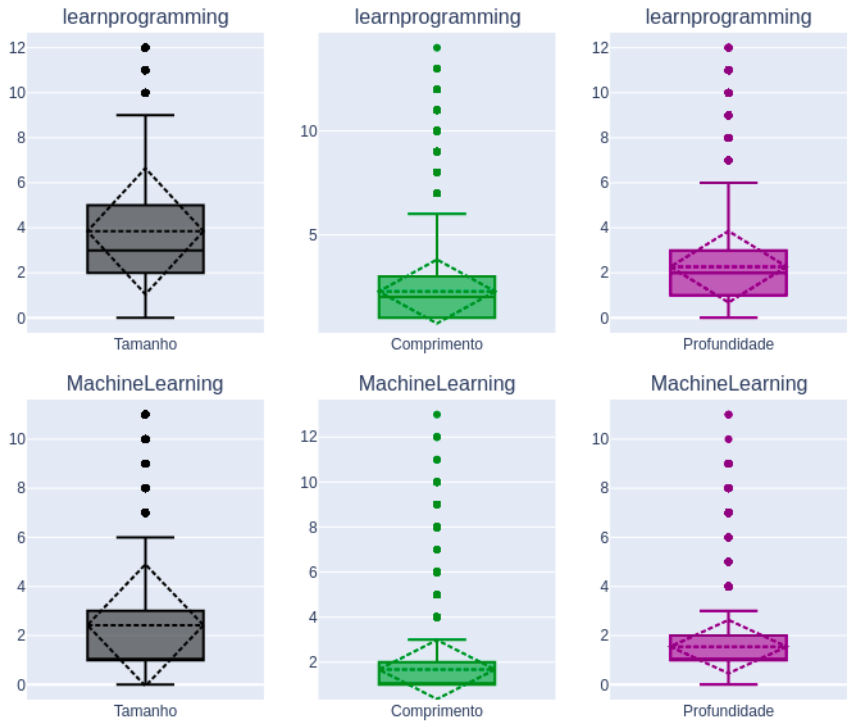
A Figura 15 apresenta a distribuição para o tamanho, comprimento e profundidade das submissões. Estes três conceitos foram definidos na subseção 2.4.1.2. A linha pontilhada representa a média aritmética e o quadrilátero pontilhado representa o desvio padrão. O resultado é bastante similar para ambas as comunidades. Nesta distribuição foram excluídos os *outliers*, pois algumas submissões eram bastante longas e possuíam tamanho máximo de 1.233 e 1.517 para as comunidades *learnprogramming* e *MachineLearning*, respectivamente. O alto valor do desvio padrão revela que há grande variabilidade no tamanho, e pouca variação no comprimento e profundidade das submissões.

5.4.1 Diversidade e intensidade das interações

A partir de um grafo direcionado com pesos nas arestas que representam o número de interações entre os vértices (como representado na Figura 2-**c**), foram calculadas as distribuições dos graus de entrada e de saída. O grau de entrada de um vértice é a quantidade de comentários que o participante correspondente recebeu em suas publicações. O grau de saída de um vértice é a quantidade de comentários publicados por um participante. Portanto, os graus de entrada e saída são métricas de diversidade de interação. O resultado foi bastante similar para as duas comunidades e pode ser observado na Tabela 10. Uma inspeção minuciosa do grau de entrada revela que aproximadamente 70% dos participantes receberam pelo menos um comentário em suas publicações, e 6,50% (*learnprogramming*) ou 7,13% (*MachineLearning*) receberam 10 ou mais comentários, o que sugere altos níveis de troca de informações nas comunidades. A análise do grau de saída mostra que 87,09% (*learnprogramming*) ou 83,65% (*MachineLearning*) dos participantes comentaram pelo menos uma vez as publicações dos seus pares, e 5,23% (*learnprogramming*) ou 6,16% (*MachineLearning*) comentaram 10 ou mais vezes.

Os graus de entrada e saída são métricas de diversidade de interação. No entanto, também é importante investigar a sua intensidade. Nesse caso, foram analisados os pesos das arestas do grafo, que representam o número de comentários publicados ou recebidos por uma mesma dupla de participantes. O resultado é apresentado na Tabela 11 e foi similar nas duas comunidades: 13,12% (*learnprogramming*) ou 11,30% (*MachineLearning*) dos participantes

Figura 15 – Distribuição dos valores de Tamanho, Comprimento e Profundidade das submissões



Fonte: Elaborada pelo autor.

Nota – Desvio padrão para as métricas da comunidade *learnprogramming* - Tamanho: 16,75; Comprimento: 8,35; Profundidade: 2,41. Desvio padrão para as métricas da comunidade *MachineLearning* - Tamanho: 23,34; Comprimento: 7,76; Profundidade: 2,05.

Tabela 10 – Diversidade na interação entre participantes

	learnprogramming	***MachineLearning***
Grau de entrada ≥ 1	70,17%	70,73%
Grau de entrada ≥ 2	42,95%	35,47%
Grau de entrada ≥ 3	28,99%	24,91%
Grau de entrada ≥ 5	16,04%	15,09%
Grau de entrada ≥ 10	6,50%	7,13%
Grau de saída ≥ 1	87,09%	83,65%
Grau de saída ≥ 2	40,14%	38,21%
Grau de saída ≥ 3	24,56%	24,34%
Grau de saída ≥ 5	12,75%	13,69%
Grau de saída ≥ 10	5,23%	6,16%

Fonte: Elaborada pelo autor.

interagiram um com o outro duas ou mais vezes, e apenas 0,74% (*learnprogramming*) ou 0,56% (*MachineLearning*) dos participantes interagiram entre si cinco ou mais vezes. Este resultado corrobora a filosofia da rede social Reddit, em que o objetivo principal é o engajamento em torno de conteúdos em que os participantes possuem interesse, independentemente de quais deles

interagem na discussão, pois a identidade dos participantes sequer é revelada, em muitos casos.

Tabela 11 – Intensidade na interação entre participantes

	learnprogramming	***MachineLearning***
Intensidade ≥ 2	13,12%	11,30%
Intensidade ≥ 3	3,86%	3,04%
Intensidade ≥ 5	0,74%	0,56%

Fonte: Elaborada pelo autor.

Também foram calculadas algumas métricas estruturais a partir da rede contendo todas as interações de cada comunidade no período analisado. Mais especificamente, foi calculado o diâmetro e a transitividade da rede, além da média, mediana e desvio padrão de três métricas egocêntricas: grau, excentricidade e o número de triângulos. A Tabela 12 mostra os resultados obtidos para cada métrica. No geral, observa-se que os valores são semelhantes para as duas comunidades. A baixa transitividade combinada com alto diâmetro e excentricidade dos nós corrobora o fato de que as interações dos usuários são baseadas principalmente no conteúdo das submissões e comentários, independentemente dos participantes que os geram. Também foi observada uma grande variação (alto DP) na distribuição do número de triângulos, indicando que esse número varia substancialmente em diferentes nós. Essa é uma consequência da grande variação no número de submissões e comentários realizados por diferentes participantes.

Tabela 12 – Métricas estruturais de rede e egocêtricas

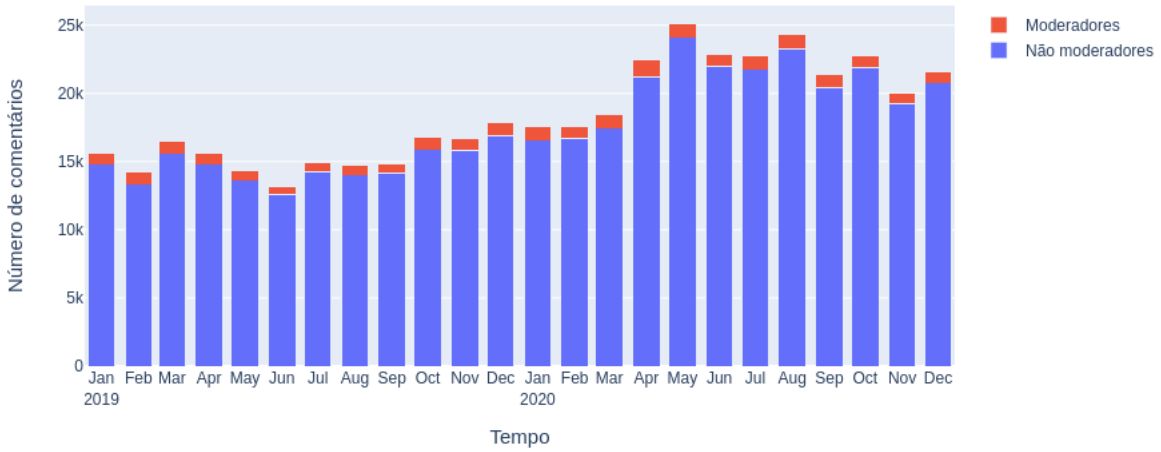
	Métricas de rede		**Métricas egocêntricas**								
			Grau			**Excentricidade**			**Triângulos**		
	Diâm.	Transit.	Média	M_d	DP	Média	M_d	DP	Média	M_d	DP
LP	67	0,0023	9,32	3	78,36	10,11	10	1,26	3,76	0	109,68
ML	82	0,0015	8,55	2	67,03	8,98	9	1,14	3,65	0	49,52

Fonte: Elaborada pelo autor.

Observações – **LP** refere-se à comunidade *learnprogramming*; **ML** refere-se à comunidade *MachineLearning*; M_d refere-se à Mediana.

Em resumo, a análise realizada nesta subseção mostra que nas duas comunidades o modelo de interação é centrado em torno do conteúdo das submissões e comentários, e não nos participantes que os criam. Esse comportamento é positivo e está de acordo com os argumentos descritos por Jenkins (2009) e Haythornthwaite (2018), em que os autores descrevem que uma característica fundamental para realizar a transição de uma comunidade *online* para uma comunidade de aprendizagem *online* passa da simples busca por informações pessoais para práticas coletivas associadas à cultura de troca de informações e organizada em torno de um domínio de conhecimento.

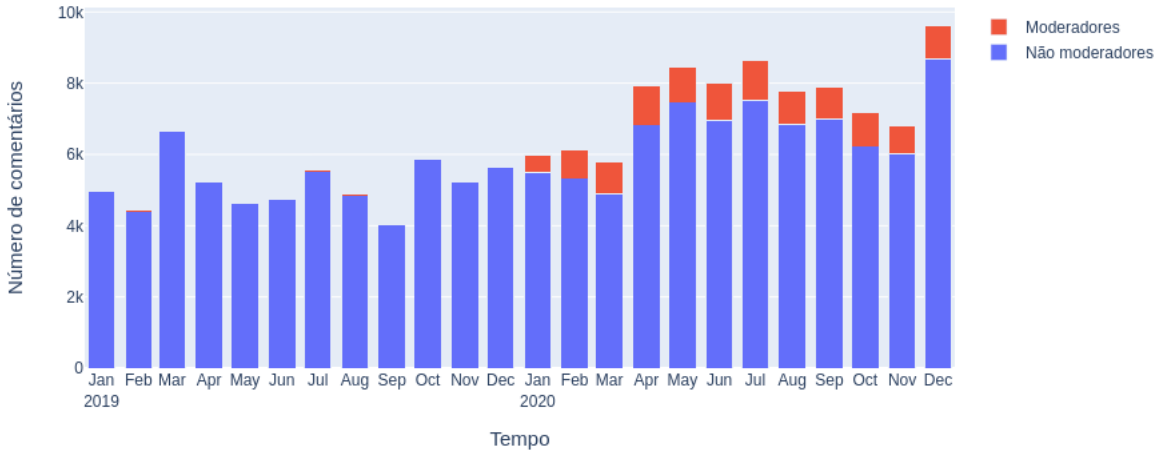
Figura 16 – Número de comentários publicados mensalmente por moderadores e não moderadores na comunidade *learnprogramming*



Fonte: Elaborada pelo autor.

Nota – Estatística descritiva para o total de comentários publicados mensalmente - Média: 18.410,16; Mediana: 17.521; Desvio Padrão: 3.632,24. Estatística descritiva para o total de comentários publicados mensalmente por moderadores - Média: 869,12; Mediana: 855; Desvio Padrão: 144,95

Figura 17 – Número de comentários publicados mensalmente por moderadores e não moderadores na comunidade *MachineLearning*



Fonte: Elaborada pelo autor.

Nota – Estatística descritiva para o total de comentários publicados mensalmente - Média: 6331,37; Mediana: 5932; Desvio Padrão: 1530,09. Estatística descritiva para o total de comentários publicados mensalmente por moderadores - Média: 460,25; Mediana: 254; Desvio Padrão: 465,84

5.4.2 O papel dos moderadores

De forma geral, os participantes de uma comunidade de aprendizagem *online* consideram benéfica a presença de moderadores, pois eles acompanham as contribuições dos participantes e intervêm regularmente para orientá-los ou para monitorar a adesão dos comentários ao tópico de discussão (HAYTHORNTHWAITE *et al.*, 2018). A atividade de moderação no Reddit requer uma reputação, mensurada em uma determinada pontuação de *karma*. À época da análise dos dados, as comunidades *learnprogramming* e *MachineLearning* possuíam oito e nove moderadores, respectivamente. A Figura 16 e a Figura 17 mostram o número de comentários publicados mensalmente por moderadores e não moderadores durante o período investigado. A comunidade *learnprogramming* apresentou um número consistente de comentários de moderadores, variando de 3,76% (Dezembro de 2020) a 5,93% (Fevereiro de 2019) do total de comentários. A comunidade *MachineLearning* teve poucos comentários publicados por moderadores em 2019, cerca de 0,37% do total. No entanto, essa taxa aumentou para 12% no ano de 2020. Estes dados demonstram que as duas comunidades possuem um grupo de moderadores que costuma publicar conteúdos regularmente.

5.4.3 Análise do conteúdo das discussões

Nesta subseção serão realizados dois tipos de análise para investigar o conteúdo publicado pelos participantes das duas comunidades: ***(i)*** uma análise qualitativa para descrever as características das submissões mais bem avaliadas pelos participantes; e ***(ii)*** a aplicação de uma técnica de mineração de textos para identificar os tópicos e os termos mais frequentes das submissões e seus comentários. Tais análises são importantes para fornecer um panorama geral sobre o conteúdo criado e compartilhado pelos participantes das duas comunidades no período sob investigação.

5.4.3.1 Submissões de maior repercussão (ou mais bem avaliadas)

A análise qualitativa visa a descrever as principais características das submissões de maior repercussão nas duas comunidades. O Quadro 4 descreve as dez submissões mais bem avaliadas na comunidade *learnprogramming*. A coluna tamanho refere-se à quantidade de comentários da submissão. Ao analisar o quadro, percebe-se que o título criado pelos usuários para identificar a discussão é longo e descreve com clareza o problema que será debatido. Em média, cada uma das submissões mais bem avaliadas possui 249 comentários. Elas se referem a temas variados, como experiências de iniciantes na área de programação, dicas de programadores mais experientes para os iniciantes ou indicação de cursos de programação gratuitos. O Quadro 5 descreve as dez submissões mais bem avaliadas na comunidade *MachineLearning*. É possível perceber uma diferença entre os títulos das submissões desta comunidade e *learnprogramming*, pois eles são mais curtos e normalmente apenas sinalizam, de forma sucinta, o assunto que será discutido. As submissões de maior repercussão possuem, em média, 135 comentários. Elas se referem a

projetos ou temas de pesquisa da área de *machine learning*, com a exceção de um tópico que discute o possível plágio do artigo de um pesquisador da área.

Quadro 4 – As dez submissões mais bem avaliadas na comunidade *learnprogramming*

Score	**Data**	**Tamanho**	**Título**
8472	14/08/2020	339	*I wrote a syllabus for learning Python and Django. Four people have gone through it, two are interviewing and one got a job. It's based on using a somewhat even mix of coding challenges, personal projects and books.*
6907	04/09/2020	89	*Please do not downvote someone who is asking for help in solving a code that looks stupid.*
6378	13/03/2020	144	*The Massachusetts Institute of Technology has a class called 'The missing semester of your computer science education' It is a collection of things that most developers and data scientists typically teach themselves on the job.*
6151	10/06/2019	325	*The online course for "Automate the Boring Stuff with Python"is free to sign up this week.*
5262	04/11/2019	290	*"Automate the Boring Stuff with Python"Udemy course is free to sign up for with code NOV2019FREE*
4912	08/06/2020	138	*After 2 years of learning programing, I knew how to write code, but didn't know how to make an application. So now, I'm writing the guide that I wish I had at that time. Here's part 1.*
4833	27/04/2020	334	*Spent the last 100 days learning to code, and told no one. Hey Reddit, this is my first post ever. Seems like a cool community so I thought I'd share my story.*
4656	19/06/2020	93	*Had to teach my "Learn Python for Research"PhD course remotely and decided to make the recordings and material publicly available for those interested. Feel free to check it out!*
4490	22/11/2019	217	*If you are learning programming(newbie), these may be your treasures on the internet!*
4482	29/10/2020	522	*EVERYTHING JUST CLICKED. dude everything is just an object. LITERALLY ITS ALL OBJECTS. THATS WHY ITS OOP. I AM SCREAMING.*

Quadro 5 – As dez submissões mais bem avaliadas na comunidade *MachineLearning*

Score	Data	Tamanho	Título
6732	10/05/2020	167	*[Project] From books to presentations in 10s with AR + ML*
4501	25/04/2020	92	*[Research] First Order Motion Model applied to animate paintings.*
3272	27/09/2020	110	*[Project] Using oil portraits and First Order Model to bring the paintings back to life.*
2817	27/12/2020	90	*[Project] Doing a clone of Rocket League for AI experiments. Trained an agent to air dribble the ball.*
2589	02/05/2020	101	*[Research] Consistent Video Depth Estimation (SIGGRAPH 2020) - Links in the comments.*
2533	20/06/2020	100	*[Research] Wolfenstein and Doom Guy upscaled into realistic faces with PULSE*
2483	12/10/2019	425	*[Discussion] Siraj has a new paper: 'The Neural Qubit'. It's plagiarised*
2476	15/11/2020	135	*[Research] [RIFE: 15FPS to 60FPS] Video frame interpolation , GPU real-time flow-based method*
2417	01/11/2020	33	[Project] *A little seasonal homage...*
2214	25/04/2020	97	*[Research] Adversarial Latent Autoencoders (CVPR2020 paper + code)*

5.4.3.2 Mineração de textos e temas mais discutidos

Para identificar os temas mais discutidos nas duas comunidades durante o período sob investigação, foi aplicada LDA: uma técnica de mineração de textos da família de modelos probabilísticos, cujo objetivo é detectar tópicos a partir de um conjunto de documentos. Nesse contexto, de acordo com as definições de Blei, Ng e Jordan (2003), pode-se apresentar os conceitos descritos a seguir.

- Uma *palavra*, ou *termo*, é a unidade discreta de análise, definida como um item de um vocabulário indexado no formato $\{1,2,3,...,V\}$. Os termos são representados por meio de elementos de um vetor em que um único elemento é igual a um e os demais são iguais a zero. Assim, o enésimo termo no vocabulário é representado por w em um vetor V, em que $w_n = 1$ e $w_u = 0$, sendo que $n \neq u$.

- Um *documento* é uma sequência de N termos denotada por $\mathbf{w} = (w_1, w_2, w_3, ..., w_N)$, em que w_N é o enésimo termo da sequência.
- Um *corpus* é uma coleção de M documentos denotada por $\mathbf{D} = (d_1, d_2, d_3, ..., d_M)$.
- Um tópico T_k é uma distribuição de w termos únicos, em que cada um tem a proporção P_{kw}, ou seja, P_{kw} representa a importância relativa do termo w para a definição do tópico k.

No contexto deste capítulo, um *documento* representa uma submissão e seus comentários, e um *corpus* representa o conjunto de submissões criadas em cada uma das comunidades, de acordo com a Tabela 8. Assim, a aplicação de LDA possibilita identificar os tópicos que representam os temas mais relevantes discutidos pelos participantes. Tal técnica foi aplicada ao *corpus* das comunidades *learnprogramming* e *MachineLearning*, considerando k (número de tópicos) igual a 10, 20 e 30. Após análise qualitativa verificou-se que o valor de $k = 10$ produziu os tópicos mais distintos entre si. A Figura 18 e a Figura 19 apresentam os termos mais frequentes e os três tópicos com a maior frequência de termos das duas comunidades. Os termos são ordenados de acordo com o seu índice de saliência (CHUANG; MANNING; HEER, 2012). Tal índice é calculado com base nos conceitos descritos a seguir.

- Para cada termo w é calculada a sua probabilidade condicional $P(T|w)$: a probabilidade de que o termo observado w foi gerado pelo tópico latente T.
- Em seguida, é calculada a probabilidade marginal $P(T)$: a probabilidade de que qualquer termo w' selecionado aleatoriamente tenha sido gerado pelo tópico latente T.
- A partir desses conceitos, é calculado o índice de *distinção* do termo w, que é definido como a divergência entre $P(T|w)$ e $P(T)$:

$$distincao(w) = \sum_T P(T|w) \times log \frac{P(T|w)}{P(T)}$$

- O índice de *distinção* descreve, no sentido da teoria da informação, o quanto o termo w é informativo para determinar o tópico gerador T, em comparação com um termo w' selecionado aleatoriamente. Assim, o índice de saliência é definido pelo seguinte produto:

$$saliencia(w) = P(w) \times distincao(w)$$

Ao analisar a Figura 18, é possível inferir as conclusões descritas a seguir.

- Entre os **termos mais frequentes** utilizados pelos participantes da comunidade *learnprogramming* (Fig. 18-**A**) estão palavras genéricas usadas no domínio da área de tecnologia, como *project*, *tutorial*, *resource* e *language*, sendo esta última a mais frequente em toda a comunidade. Também destacam-se palavras relacionadas a atividades sociais ou recursos utilizados pelos usuários (palavras como *find*, *share*, *tip* e *guideline*), e referências à própria comunidade (palavras como *team* e *community*).

Figura 18 – Termos mais frequentes e os três tópicos mais relevantes para a comunidade *learnprogramming*

Fonte: Elaborada pelo autor.

- No **primeiro tópico** (Fig. 18-**B**) a palavra mais frequente é *language*, o que sugere discussões em torno de diferentes linguagens de programação. Neste tópico, aparecem o nome de três linguagens, indicando que estas são as mais discutidas pelos participantes da comunidade: *python*, *java* e *javascript*. Também é possível identificar alguns termos que indicam discussões relacionadas às grandes áreas de tecnologias de implementação: *web*, *data*, *programming*, *game*, *app* e *database*. Os termos *course*, *job* e *work* indicam discussões relacionadas ao mercado de trabalho e cursos de aperfeiçoamento. Por fim, é importante destacar o uso frequente da palavra *thanks*, que foi citada mais de 20.000 vezes. Mensagens com referências a palavras de agradecimento podem indicar que um participante aprendeu algo que foi demonstrado por outro participante (GRUZD; PAULIN; HAYTHORNTHWAITE, 2016).

- O **segundo tópico** (Fig. 18-**C**) basicamente se refere a diversos conceitos ou palavras-chaves que são comuns a maioria das linguagens de programação, como *array*, *function*, *loop*, *string*, *int*, *return*, *recursion*, *list*, *method*, *class* e *pointer*, entre outras. Portanto, este tópico possui termos que sugerem discussões sobre conceitos básicos das linguagens de programação.

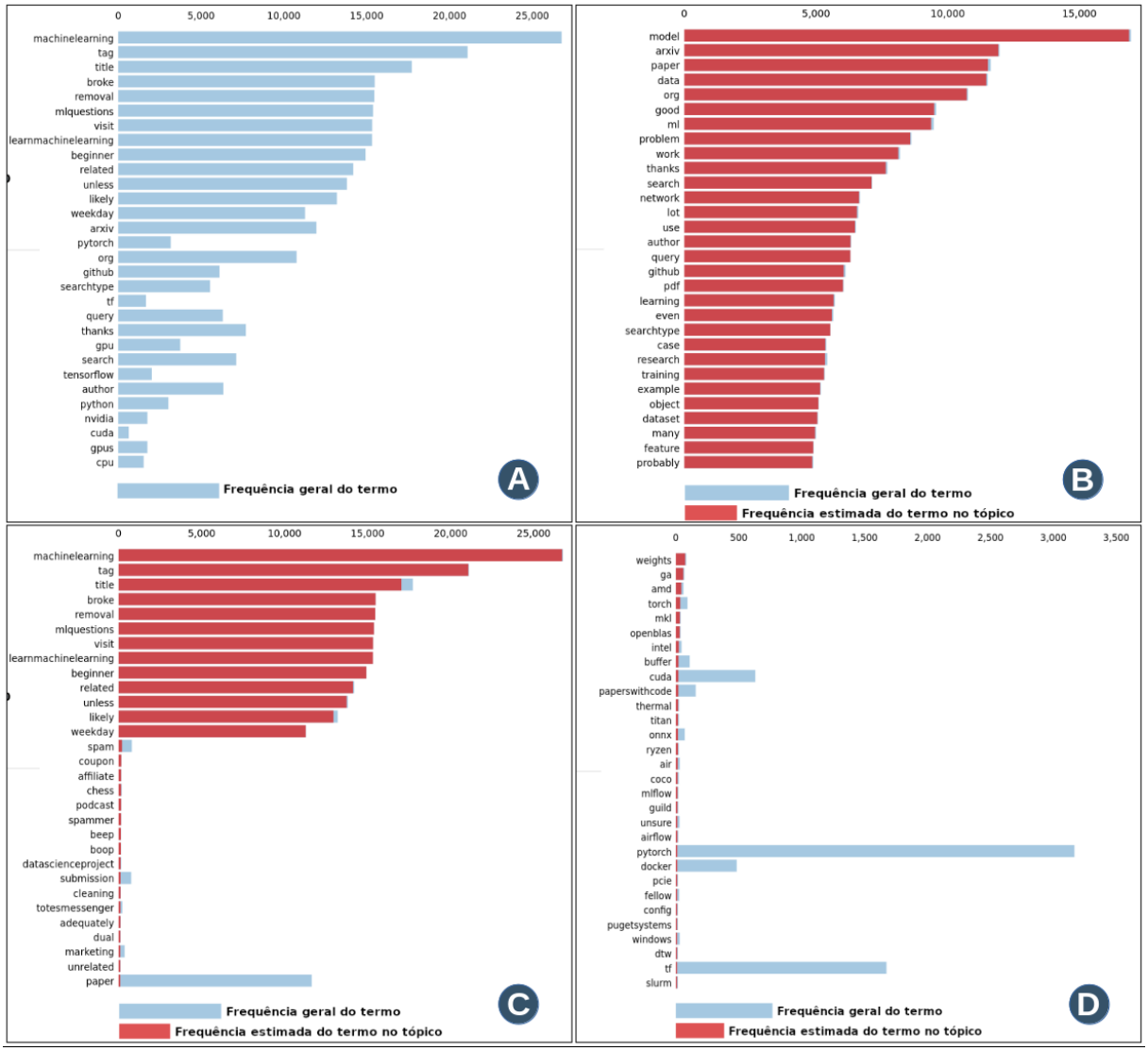
- Por fim, o **terceiro tópico** (Fig. 18-**D**) possui termos relacionados à discussão de linguagens de programação ou *frameworks* tecnológicos mais específicos, como *selenium*[5], *flutter*[6], *kotlin*[7] e *scala*[8]. Estes termos são citados com menor frequência, portanto, eles aparecem em um número menor de discussões, em contraponto ao primeiro tópico que apresenta os termos mais citados e se refere a tecnologias de propósito geral.

De forma análoga, ao analisar a Figura 19 é possível inferir as conclusões descritas a seguir sobre o conteúdo das discussões na comunidade *MachineLearning*.

- Entre os **termos mais frequentes** utilizados pelos usuários da comunidade (Fig. 19-**A**) estão as palavras *visit*, *beginner* e *learnmachinelearning*, sendo que esta última faz referência a uma comunidade[9] do *Reddit* para aprendizagem de temas relacionados à *machine learning*; este pode ser um indicativo de que tal comunidade é sugerida (*visit*) para iniciantes (*beginners*) que publicam na comunidade *MachineLearning*, cujos participantes preferem discutir temas mais avançados. Alguns conceitos e tecnologias da área de *machine learning* aparecem entre os termos mais frequentes, como *pytorch*[10], *tf* ou *tensorflow*[11] e *python*. Também existem termos que se referem ao poder de processamento necessário

[5] http://www.selenium.dev
[6] http://flutter.dev
[7] http://kotlinlang.org
[8] http://www.scala-lang.org
[9] http://www.reddit.com/r/learnmachinelearning
[10] http://pytorch.org
[11] http://www.tensorflow.org

Figura 19 – Termos mais frequentes e os três tópicos mais relevantes para a comunidade *MachineLearning*



Fonte: Elaborada pelo autor.

para lidar com grandes conjuntos de dados, como *gpu* (sigla de *Graphics Processing Unit*), *nvidia*, *gpus* e *cpu*. Por fim, a palavra *thanks* é citada quase 10.000 vezes, e assim como na comunidade *learnprogramming* pode indicar que um participante aprendeu algo que foi demonstrado.

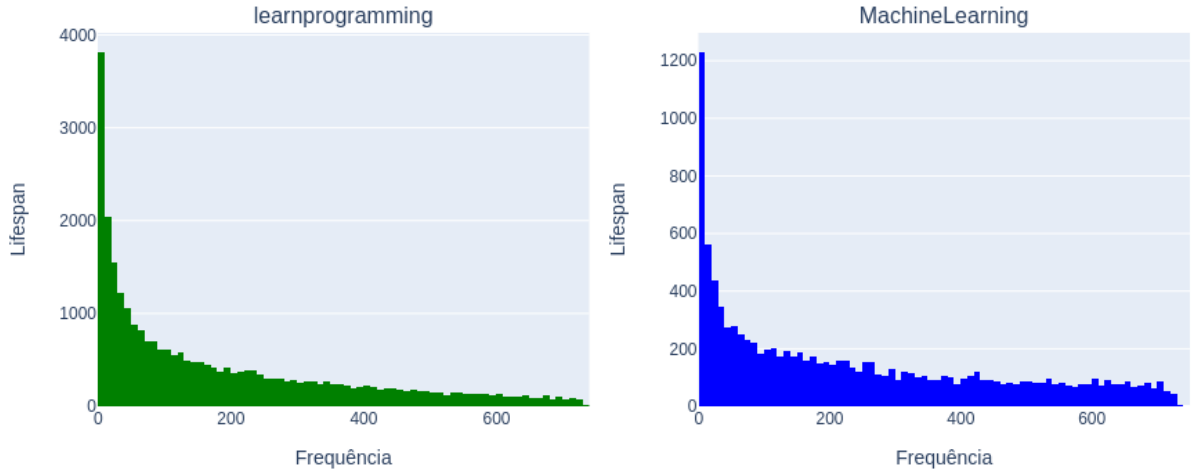
- No **primeiro tópico** (Fig. 19-**B**) existem muitos termos referentes ao processo de treinamento de modelos de *machine learning*, como *model*, *data*, *learning*, *training*, *example*, *dataset* e *feature*. Corroborando com o argumento que a comunidade *MachineLearning* visa a discutir temas mais avançados da área, existem muitos termos relacionados à publicação de artigos científicos, como *arxiv* (arquivo para *preprints* eletrônicos de artigos científicos), *paper*, *problem*, *author* e *research*.

- Dos 30 termos que compõem o **segundo tópico** (Fig. 19-**C**), 13 possuem alta frequência (acima de 10.000) e possuem palavras como *visit*, *beginner* e *learnmachinelearning*. Os demais termos do segundo tópico e a maioria dos termos do **terceiro tópico** (Fig. 19-**D**) são palavras com baixa frequência, muitas delas relacionadas ao processo de treinamento de modelos ou tecnologias da área de *machine learning*, como *weights*, *amd*, *torch*, *intel*, *mlflow*, *airflow*, *pytorch* e *tf*.

5.5 Análise das atividades dos participantes

As atividades dos participantes das comunidades se referem às submissões e aos comentários publicados por eles, além da execução do processo de avaliação por pares, em que as submissões e os comentários dos outros participantes são votados. A Figura 20 apresenta o *lifespan* dos participantes em cada comunidade. O *lifespan* representa a diferença, em dias, entre a primeira e a última publicação do participante no período sob investigação. Consequentemente, aqueles que tiveram apenas uma atividade na comunidade foram descartados. A distribuição do *lifespan* é bastante semelhante em ambas as comunidades. Como os dados analisados correspondem ao intervalo de dois anos, entre 2019 e 2020, o *lifespan* mais longo possível é 731 dias. Observa-se que há poucos participantes com *lifespan* maior que 600 em cada uma das comunidades, e que a maioria dos participantes permanece ativo por menos de 200 dias. A média aritmética do *lifespan* em dias da comunidade *learnprogramming* é 137,25 e em *MachineLearning* é 188,06.

Figura 20 – Distribuição do *lifespan* em dias.

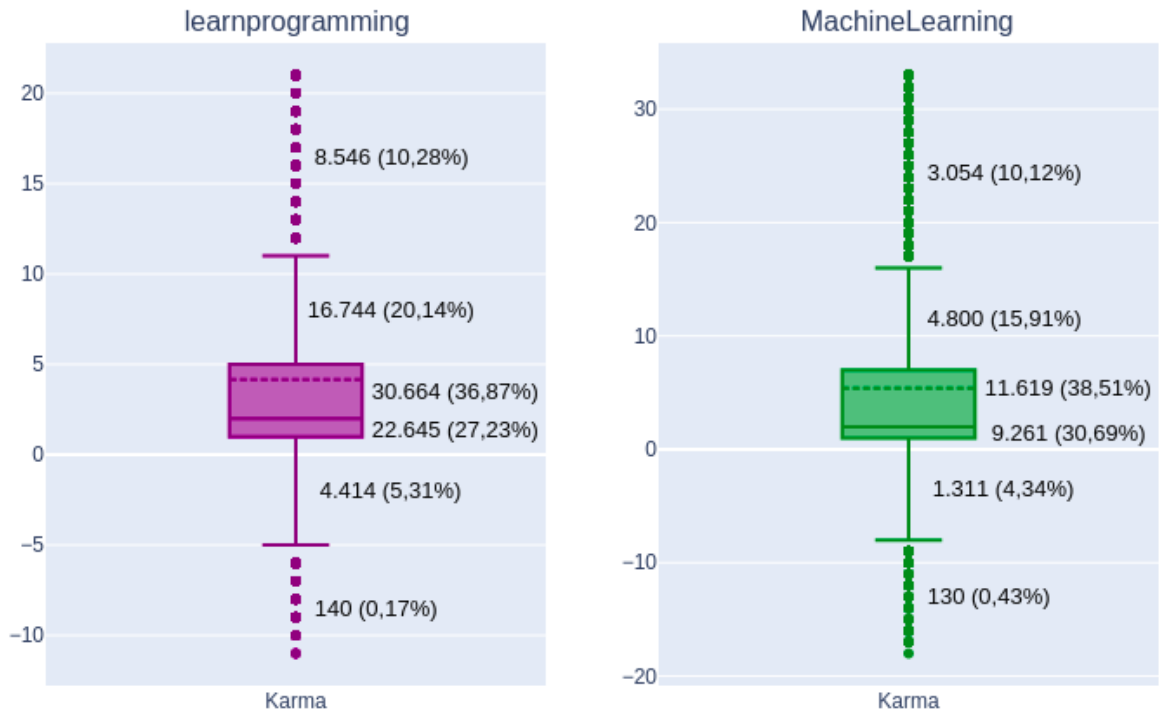


Fonte: Elaborada pelo autor.

A análise da pontuação do *karma* dos participantes também é importante, pois tal análise mostra como eles avaliam a participação dos seus pares. Foram excluídos os participantes considerados *outliers*, pois alguns deles possuíam pontuação de *karma* com valores discrepantes,

como -385 e 4.158 na comunidade *learnprogramming*. Assim, foram excluídos 12.182 e 5.527 participantes nas comunidades *learnprogramming* e *MachineLearning*, respectivamente. A Figura 21 mostra a distribuição do *karma* nas duas comunidades. Os números mostrados em cada um dos *boxplots* representam a quantidade de participantes em cada intervalo e seu percentual em relação ao total, entre parênteses. O resultado foi bastante similar em ambas as comunidades. Ao realizar uma análise mais detalhada na comunidade *learnprogramming*, é possível utilizar a distribuição do valor do *karma* para dividir os seus participantes em seis grupos, como descrito a seguir (Q1 e Q3 significam primeiro quartil e terceiro quartil, respectivamente).

Figura 21 – Distribuição da pontuação de *karma* e a quantidade de participantes em cada intervalo.



Fonte: Elaborada pelo autor.

- Grupo 1: 140 participantes (ou 0,17% do total) possuem *karma* abaixo do limite inferior; este grupo representa os participantes que foram avaliados majoritariamente de forma negativa, ou seja, as suas publicações não foram reconhecidas como contribuições de qualidade pelos seus pares.
- Grupo 2: 4.414 participantes (ou 5,31% do total) possuem *karma* entre o limite inferior e Q1; este grupo possui participantes que foram avaliados tanto de forma negativa quanto positiva.
- Grupo 3: 22.645 participantes (ou 27,23% do total) possuem *karma* entre Q1 e a mediana; a partir deste grupo, todos os participantes foram avaliados majoritariamente de forma positiva.

- Grupo 4: 30.664 participantes (ou 36,87% do total) possuem *karma* entre a mediana e Q3; este representa o maior dos seis grupos.
- Grupo 5: 16.744 participantes (ou 20,14% do total) possuem *karma* entre Q3 e o limite superior.
- Grupo 6: por fim, 8.546 participantes (ou 10,28%) do total possuem *karma* acima do limite superior; este grupo representa os participantes cujas publicações foram mais bem avaliadas pela comunidade.

Com base nesta análise, é possível estabelecer que o **grupo 6** representa os especialistas (ou *experts*) da comunidade, pois o valor e a qualidade das suas contribuições foram devidamente reconhecidas pelos seus pares. Wenger, McDermott e Snyder (2002) argumentam que nas comunidades que visam a criação e o compartilhamento de conhecimento com base em práticas sociais, o grupo central de participantes, que representa aqueles mais ativos e cuja participação é frequentemente legitimada por toda a comunidade, corresponde a um grupo de 10 a 15 porcento do total de participantes (WENGER; MCDERMOTT; SNYDER, 2002, p. 56). Esta evidência corrobora com o argumento que o grupo 6 representa os especialistas da comunidade *learnprogramming*, pois ele representa 10,28% do total de participantes.

A mesma análise pode ser feita com os participantes da comunidade *MachineLearning*, pois a distribuição dos valores de *karma* foi bastante semelhante à de *learnprogramming*. Portanto, em *MachineLearning* o grupo de especialistas (grupo 6) é formado por 3.054 participantes, ou 10,12% do total.

5.6 Considerações finais

Este capítulo apresentou a caracterização das duas comunidades *online* da rede social Reddit investigadas nesta tese de doutorado: *learnprogramming* e *MachineLearning*. Tal caracterização é importante para fornecer um panorama sobre as submissões e comentários publicados, as atividades e o perfil dos participantes, além dos temas mais discutidos no período sob investigação. A análise das características gerais das comunidades indicou que as discussões são realizadas em torno de determinados assuntos, e não dos participantes que as publicam. Tal característica é importante para qualificar as comunidades investigadas como comunidades de aprendizagem *online*. A análise do conteúdo das submissões mostrou que, em cada comunidade, as discussões frequentemente abordam temas relevantes das suas respectivas áreas, aprendizagem de programação e *machine learning*. Por fim, a análise da atividade dos participantes mostrou que os membros das comunidades cujas contribuições são as mais bem avaliadas representam cerca de 10% do total de participantes ativos. No próximo capítulo, as duas comunidades caracterizadas serão avaliadas por meio da aplicação de um estudo de caso para o processo SLIM.

CAPÍTULO

6

UM ESTUDO DE CASO PARA O PROCESSO SLIM

6.1 Considerações iniciais

Neste capítulo é apresentado um estudo de caso para o processo SLIM. O estudo foi conduzido por meio da aplicação do SLIM na avaliação das comunidades de aprendizagem *learnprogramming* e *MachineLearning*, caracterizadas no Capítulo 5. O objetivo geral do estudo de caso é investigar a efetividade da aplicação do processo SLIM na geração de conhecimento sobre a aprendizagem social *online* em ambientes informais, por meio da avaliação de comunidades de aprendizagem. O restante do capítulo está organizado como segue. A seção 6.2 descreve o planejamento para o estudo de caso. A seção 6.3 descreve a aplicação do estudo de caso. Por fim, a seção 6.4 faz as considerações finais sobre o capítulo.

6.2 Um estudo de caso para o processo SLIM

Estudos de caso são métodos de pesquisa observacionais conduzidos pela observação de um projeto ou atividade em andamento, cujo objetivo é investigar uma entidade ou fenômeno em seu contexto real, em um intervalo de tempo específico. Tais estudos utilizam diferentes fontes de evidências, como documentações, registros de arquivos, observação direta, observação participante, artefatos físicos, entre outros (WOHLIN *et al.*, 2012).

Os estudos de caso são a estratégia selecionada quando perguntas com as palavras "como" ou "por que" são propostas, quando o foco está em um fenômeno contemporâneo em contextos da vida real, quando o investigador tem pouco controle sobre os eventos estudados e não é possível, ou é difícil, manipular o comportamento dos indivíduos envolvidos no estudo (HOLLWECK, 2015). Ao conduzir um estudo de caso, existem cinco etapas principais que devem ser seguidas (WOHLIN *et al.*, 2012):

- **planejamento do estudo de caso** - os objetivos são definidos e o estudo é planejado;
- **preparação para a coleta de dados** - são definidos procedimentos e protocolos para a coleta de dados;
- **coleta de dados** - execução da coleta de dados do estudo de caso;
- **análise dos dados coletados** - os dados coletados são analisados;
- **redação de relatórios** - é realizada a documentação do estudo de caso.

As próximas subseções descrevem a execução das etapas referentes ao estudo de caso apresentado neste capítulo.

6.2.1 Planejamento do estudo de caso

A etapa de planejamento é crucial para a execução bem sucedida de um estudo de caso. Um plano deve conter alguns elementos, como a descrição do seu objetivo, a especificação da Questão de Pesquisa (QP) que será investigada, a definição do tipo de estudo e os métodos de coleta de dados (WOHLIN *et al.*, 2012). No estudo de caso apresentado neste capítulo, os elementos do planejamento foram definidos como descritos a seguir.

- **Objetivo**: investigar a efetividade da aplicação do processo SLIM na geração de conhecimento sobre a aprendizagem social *online* em ambientes informais, por meio da avaliação de comunidades de aprendizagem.
- **Questão de pesquisa**: as questões de pesquisa definem o que é necessário investigar para cumprir o objetivo do estudo (RUNESON; HÖST, 2009). A QP que motivou e guiou este trabalho de doutorado foi a seguinte: "como a análise dos dados gerados pelos participantes de comunidades *online* pode contribuir para ampliar a compreensão e gerar conhecimento sobre a aprendizagem social em ambientes informais de aprendizagem?" Esta QP irá nortear a criação de questões mais específicas nas atividades relacionadas à criação de conhecimento do processo SLIM.
- **Tipo do estudo**: um estudo de caso pode pertencer aos tipos exploratório, descritivo ou explicativo (RUNESON; HÖST, 2009). Neste trabalho de doutorado foi desenvolvido um estudo exploratório que visa a formulação e a confirmação de hipóteses baseadas na QP definida anteriormente.
- **Métodos de coleta de dados**: a coleta de dados de um estudo de caso pode ser realizada por diferentes métodos. No método direto, o pesquisador está em contato em tempo real com os sujeitos envolvidos no estudo, por meio do uso de técnicas como entrevistas. No método indireto, o pesquisador coleta dados sem interagir com os sujeitos envolvidos, por

meio de técnicas como a instrumentação de ferramentas e análise de arquivos de dados (RUNESON; HÖST, 2009). No estudo de caso descrito neste capítulo, foi utilizado o método indireto de coleta de dados, pois foram extraídos dados de diferentes aspectos da participação dos usuários em comunidades de aprendizagem *online*.

6.2.2 Preparação para a coleta de dados

Na etapa de preparação foi implementado um *script* na linguagem de programação *Python* para realizar a coleta, limpeza, transformação, cálculo das métricas e armazenamento dos dados das comunidades da rede social Reddit descritas e caracterizadas no capítulo 5. Tal *script* foi implementado de acordo com as diretrizes das atividades do processo SLIM relacionadas aos dados, especificadas na subseção 4.3.1. Detalhes técnicos da implementação de tais atividades são descritas como segue:

- **extração dos dados** (subseção 4.3.1.1) - Reddit disponibiliza uma API que possibilita extrair dados referentes às suas comunidades a partir da especificação de uma data inicial e data final de coleta; assim, na atividade de extração foram utilizadas as bibliotecas para a linguagem *Python* PSAW[1] e PRAW[2] para conectar-se à API Reddit e coletar os dados descritos na subseção 4.2.1.2;

- **limpeza e transformação dos dados** (subseção 4.3.1.2) - foram utilizadas bibliotecas *Python* para limpeza e transformação de dados, como *pandas*[3], *numpy*[4] e *statsmodels*[5]; a normalização e correção das palavras em atributos que representam o texto dos comentários dos participantes foi realizada por meio da biblioteca *ekphrasis*[6];

- **cálculo das métricas e armazenamento dos dados** (subseção 4.3.1.3) - os dados foram armazenados em coleções do banco de dados não-relacional MongoDB[7]; nas diretrizes de cálculo das métricas foram utilizados os seguintes recursos:

  - métricas de SNA: calculadas por meio da biblioteca *Python networkx*[8];
  - métricas de estrutura: calculadas por meio da biblioteca *Python treelib*[9];
  - métricas do discurso: calculadas por meio da biblioteca *Python pyliwc*[10].

[1] http://github.com/dmarx/psaw
[2] http://praw.readthedocs.io
[3] http://pandas.pydata.org
[4] http://numpy.org
[5] http://www.statsmodels.org
[6] http://github.com/cbaziotis/ekphrasis
[7] http://www.mongodb.com
[8] http://networkx.org/
[9] http://treelib.readthedocs.io
[10] https://github.com/dfederschmidt/pyliwc

6.2.3 Coleta de dados

O objetivo desta etapa é efetivamente realizar a coleta dos dados que serão analisados no estudo de caso. Os dados foram coletados a partir da execução do *script* descrito na seção 6.2.2. A descrição dos dados coletados e a caracterização das comunidades de aprendizagem foram apresentadas no capítulo 5.

6.2.4 Análise dos dados coletados

A análise dos dados é conduzida de forma diferente para dados quantitativos e qualitativos, conforme discutido nas próximas subseções.

6.2.4.1 Análise dos dados quantitativos

A análise dos dados quantitativos geralmente inclui etapas como a análise de métricas estatísticas descritivas, análise de correlação e criação de modelos (WOHLIN *et al.*, 2012). Métricas de estatística descritiva, como valores médios, desvio padrão, histogramas e gráficos de dispersão, são usadas para obter uma compreensão dos dados que foram coletados. A análise de correlação e a criação de modelos possibilitam reconhecer padrões inerentes aos dados.

No estudo de caso, a etapa de análise dos dados quantitativos foi realizada em duas fases, como descrito a seguir.

- **Criação e análise dos modelos seccionais**: foi guiada pelas diretrizes das atividades do processo SLIM relacionadas à criação e treinamento (subseção 4.3.2.2), avaliação (subseção 4.3.2.3) e disponibilização (subseção 4.3.2.4) dos modelos seccionais. Tais atividades foram automatizadas por meio da implementação de um *pipeline* de ML. Um *pipeline* é um elemento que representa a automatização da sequência de atividades necessárias à construção de um modelo de ML (GÉRON, 2019, p. 38). Tal elemento é comum em sistemas baseados nas técnicas de ML, pois há muitos dados que precisam ser transformados e processados ao longo do treinamento e avaliação dos modelos. A implementação foi realizada por meio da biblioteca *Python Scikit-learn*[11], cujo objetivo é apoiar a criação de *pipelines* que envolvem tarefas de ML supervisionadas ou não-supervisionadas.

- **Criação e análise dos modelos temporais**: foi guiada pelas diretrizes das atividades do processo SLIM relacionadas à criação e treinamento (subseção 4.3.2.5), visualização (subseção 4.3.2.6), avaliação (subseção 4.3.2.7) e disponibilização (subseção 4.3.2.8) dos modelos temporais. Para apoiar a atividade de visualização dos modelos temporais (diretrizes D.9.1 a D.9.4), foi implementado um *Learning Analytics Dashboard* (LAD). Um LAD é uma ferramenta baseada em interfaces gráficas, geralmente de página única,

[11] http://scikit-learn.org

que apresenta de forma visual o estado atual e tendências detectadas ao longo do tempo de indicadores-chave de dados educacionais, com o objetivo de proporcionar tomadas de decisão rápidas (SCHWENDIMANN *et al.*, 2016). O LAD foi implementado utilizando o *framework Dash*[12]. Assim, ele permite realizar de forma interativa a análise visual e exploratória dos dados educacionais investigados no processo SLIM e apoiar a criação de questões de pesquisa sobre a dinâmica das OLC.

6.2.4.2 Análise dos dados qualitativos

O objetivo básico da análise dos dados qualitativos é inferir conclusões a partir dos dados coletados, mantendo uma cadeia de evidências em que o pesquisador deve ser capaz de acompanhar tais conclusões. Duas técnicas podem ser aplicadas durante a análise de dados qualitativos: a formulação de hipóteses e a confirmação de hipóteses (RUNESON; HÖST, 2009). A técnica de formulação visa a encontrar hipóteses a partir dos dados coletados e analisados. Por sua vez, as técnicas de confirmação de hipóteses podem ser usadas para confirmar se uma hipótese é realmente verdadeira, por meio de métodos estatísticos apropriados ou a análise de mais dados (WOHLIN *et al.*, 2012). As duas técnicas podem ser usadas iterativamente e em combinação. Assim, hipóteses primeiramente são formuladas e depois confirmadas. Uma vez confirmadas, novas hipóteses podem ser geradas dando início a um novo ciclo de confirmação.

A análise dos dados qualitativos foi realizada de acordo com as diretrizes das atividades do processo SLIM relacionadas à produção de conhecimento: formulação de hipóteses (subseção 4.3.3.1), teste de hipóteses (subseção 4.3.3.2) e geração de conhecimento (subseção 4.3.3.3). Os modelos do tipo ERG construídos na atividade de teste de hipóteses foram treinados por meio da biblioteca para o ambiente R *ergm*[13].

6.2.5 Redação de relatórios

A redação de relatórios é a etapa que tem o objetivo fundamental de comunicar os resultados do estudo de caso. Assim, ela produz a principal fonte de informação para analisar a qualidade do estudo. Os relatórios produzidos podem ter diferentes formatos e ser direcionados para diferentes públicos (WOHLIN *et al.*, 2012). A comunicação dos resultados do estudo de caso apresentado neste capítulo será realizada na próxima seção desta tese de doutorado.

6.3 SLIM: aplicação do estudo de caso

Esta seção apresenta os resultados do estudo de caso em que as atividades do processo SLIM foram aplicadas, com o objetivo de investigar a sua efetividade durante a avaliação da aprendizagem social *online* nas comunidades *learnprogramming* e *MachineLearning*. Após

[12] http://plotly.com/dash
[13] http://cran.r-project.org/web/packages/ergm

a execução das etapas de planejamento (subseção 6.2.1) e preparação para a coleta de dados (subseção 6.2.2) do estudo de caso, as próximas subseções descrevem a execução das etapas de coleta e análise dos dados.

6.3.1 Execução da coleta de dados

A etapa de coleta de dados foi executada pelo *script* descrito na subseção 6.2.2. Os parâmetros de entrada da atividade de **extração de dados** foram definidos da seguinte forma:

- data inicial da coleta: 01/01/2019;
- data final da coleta: 31/12/2020;
- URL das comunidades:
  - *http://www.reddit.com/r/learnprogramming*;
  - *http://www.reddit.com/r/MachineLearning*.

O total de dados coletados das duas comunidades foi apresentado na Tabela 8 (subseção 5.3). Na atividade de **limpeza e transformação dos dados** foram excluídas 6.491 e 4.694 submissões sem respostas das comunidades *learnprogramming* e *MachineLearning*, respectivamente. Por fim, na atividade de **cálculo das métricas e armazenamento dos dados**, as métricas estruturadas e do discurso foram calculadas e armazenadas nas respectivas coleções de documentos JSON no banco de dados MongoDB.

6.3.2 Análise dos dados quantitativos

A etapa de análise dos dados quantitativos foi guiada pelas atividades do processo SLIM relacionadas aos modelos e automatizadas no *pipeline* de ML descrito na subseção 6.2.4.1. Após a definição das métricas de avaliação, os modelos seccionais e temporais foram criados, treinados, avaliados e disponibilizados, conforme descrito nas próximas subseções.

6.3.2.1 Criação e análise dos modelos seccionais

O processo SLIM prevê a criação e análise de três modelos seccionais (conforme descrito na subseção 4.3.2.2): ***(i)*** modelo dos participantes especialistas; ***(ii)*** modelo dos tópicos de discussão bem avaliados; e ***(iii)*** modelo do estilo de discurso dos participantes. Na criação e treinamento destes modelos foram descartados os tópicos de discussão excluídos na atividade de limpeza e transformação dos dados. Assim, foram considerados 62.956 e 17.430 tópicos de discussão nas comunidades *learnprogramming* e *MachineLearning*, respectivamente. A avaliação dos três modelos foi guiada pelas diretrizes D.6.1 a D.6.4.

A criação e o treinamento dos **modelos dos participantes especialistas** foram guiados pelas diretrizes D.5.1.1 a D.5.1.3. A Tabela 13 mostra o resultado dos modelos para as duas comunidades. O coeficiente de determinação ajustado calculado pela atividade de avaliação foi de $0,7511$ e $0,8053$ para as comunidades *learnprogramming* e *MachineLearning*, respectivamente. Portanto, a qualidade dos modelos foi considerada satisfatória. Assim, foi possível identificar quatro métricas significativas (destacadas em negrito) para cada comunidade: ***(i)*** grau de entrada, centralidade de intermediação e *pagerank* foram consideradas significativas para ambas as comunidades; ***(ii)*** grau de saída foi considerada significativa apenas para *learnprogramming*; e ***(iii)*** e a centralidade de autovetor apenas para *MachineLearning*. Apesar desta diferença entre as métricas significativas identificadas nos modelos, é possível evidenciar a importância da quantidade de participação, da popularidade e da mediação do fluxo de informação para que um participante seja reconhecido como especialista pelos seus pares.

Tabela 13 – Coeficientes de correlação de *Pearson* para o modelo dos participantes especialistas

Comunidade *learnprogramming*				
1. Reputação (karma)	*1,00*		6. Centralidade de autovetor	0,16
2. Grau de entrada	**0,80***		7. Excentricidade	-0,04
3. Grau de saída	**0,59***		8. HITS Hub	0,24
4. Centralidade de intermediação	**0,65*****		9. HITS Authority	0,14
5. Centralidade de proximidade	0,01		**10. Pagerank**	**0,75*****
Comunidade *MachineLearning*				
1. Reputação (karma)	*1,00*		**6. Centralidade de autovetor**	**0,76*****
2. Grau de entrada	**0,84*****		7. Excentricidade	-0,03
3. Grau de saída	0,38*		8. HITS Hub	0,22
4. Centralidade de intermediação	**0,52*****		9. HITS Authority	0,01
5. Centralidade de proximidade	0,01		**10. Pagerank**	**0,85*****

Fonte: Elaborada pelo autor.

Nota – *** significa *p-value* $<0,001$; * significa *p-value* $<0,05$;
Coeficiente de determinação ajustado para a comunidade *learnprogramming*: $0,7511$;
Coeficiente de determinação ajustado para a comunidade *MachineLearning*: $0,8053$.

A criação e o treinamento dos **modelos das discussões bem avaliadas** foram guiados pelas diretrizes D.5.2.1 a D.5.2.3. A Tabela 14 mostra o resultado dos modelos para as duas comunidades. O coeficiente de determinação ajustado calculado pela atividade de avaliação foi de $0,8200$ e $0,7980$ para as comunidades *learnprogramming* e *MachineLearning*, respectivamente. Portanto, a qualidade dos modelos foi considerada satisfatória. A aprendizagem em comunidades *online* depende amplamente do nível de atividade e da participação contínua de seus membros (HAYTHORNTHWAITE; LAAT, 2010). Assim, como esperado, os modelos de ambas as comunidades possibilitaram identificar três métricas de estrutura significativas que representam quantidade de interação: número de participantes, tamanho e comprimento das submissões. Em relação às métricas de SNA, o número de gargalos evidencia a importância dos participantes que podem mediar a comunicação entre os diferentes componentes da rede, a fim de facilitar o fluxo

Tabela 14 – Coeficientes de correlação de *Pearson* para o modelo das discussões bem avaliadas

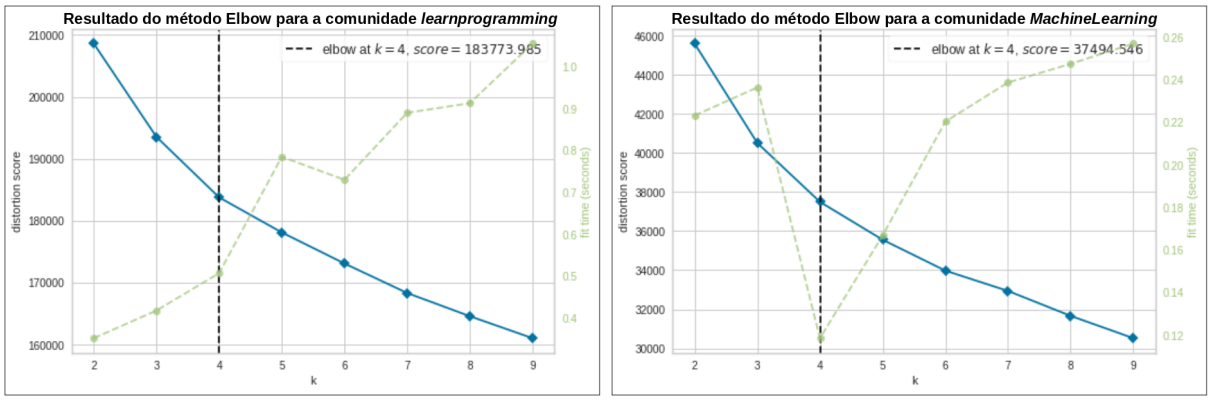
Comunidade *learnprogramming*				
Métricas de estrutura			**Métricas de SNA**	
1. Score da submissão	*1,00*		9. Densidade	-0,18
2. Número de participantes	**0,87*** **		10. Reciprocidade	-0,02
3. Tamanho	**0,84*** **		11. Número de componentes	0,00
4. Tempo da 1ª resposta	0,00		12. Caminho médio	-0,03
5. Comprimento	**0,79*** **		13. Coeficiente de agrupamento	0,01
6. Profundidade	0,25***		14. Diâmetro	0,37***
7. Intensidade das discussões	0,03		**15. Número de triângulos**	**0,64*** **
8. Duração do engajamento	0,23		**16. Número de gargalos**	**0,56*** **
Comunidade *MachineLearning*				
Métricas de estrutura			**Métricas de SNA**	
1. Score da submissão	*1,00*		9. Densidade	-0,26
2. Número de participantes	**0,85*** **		10. Reciprocidade	0,05
3. Tamanho	**0,81** **		11. Número de componentes	0,01
4. Tempo da 1ª resposta	0,00		12. Caminho médio	0,04
5. Comprimento	**0,70*** **		13. Coeficiente de agrupamento	0,05
6. Profundidade	0,45***		**14. Diâmetro**	**0,51*** **
7. Intensidade das discussões	0,20		**15. Número de triângulos**	**0,52*** **
8. Duração do engajamento	0,21		**16. Número de gargalos**	**0,78*** **

Nota – *** significa *p-value* $<0,001$; ** significa *p-value* $<0,01$;
Coeficiente de determinação ajustado para a comunidade *learnprogramming*: $0,8200$;
Coeficiente de determinação ajustado para a comunidade *MachineLearning*: $0,7980$.

das informações (SILVA; GIMENES; MALDONADO, 2020b). Além disso, a importância do número de triângulos é uma característica que leva à transitividade, um indicativo da existência de uma comunidade *online* não restritiva em que a informação pode fluir livremente (JAN, 2019). Por fim, a identificação do diâmetro como uma métrica significativa na comunidade *MachineLearning* permite sugerir que as informações compartilhadas pelos membros de tal comunidade atingem os participantes mais distantes da rede, quando comparada à comunidade *learnprogramming*.

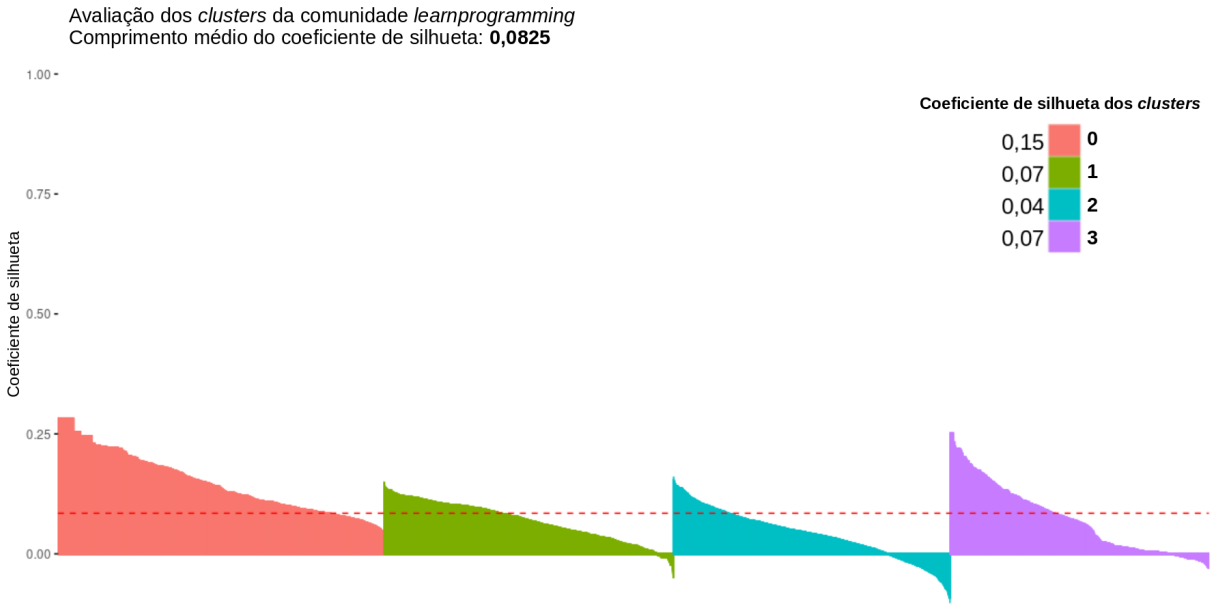
A criação e o treinamento dos **modelos do estilo de discurso dos participantes** foram guiados pelas diretrizes D.5.3.1 a D.5.3.6. Inicialmente, foi utilizado o método *Elbow* para encontrar o valor adequado do hiper-parâmetro k do algoritmo *Kmeans*. O resultado é apresentado na Figura 22. A linha tracejada na vertical mostra que o número adequado de *clusters* para ambas as comunidades é igual a quatro. Após o treinamento dos modelos foram geradas a Figura 23 e a Figura 24, que representam a qualidade dos modelos do estilo de discurso das comunidades *learnprogramming* e *MachineLearning*, respectivamente. Nos quatro *clusters* de ambos os modelos, o coeficiente de silhueta foi positivo, sendo que o seu comprimento médio foi de $0,0825$ e $0,0725$. Isso significa que o método de clusterização foi capaz de separar as submissões das comunidades de forma adequada, ou seja, a qualidade dos modelos é satisfatória.

Figura 22 – Resultado da aplicação do método *Elbow*



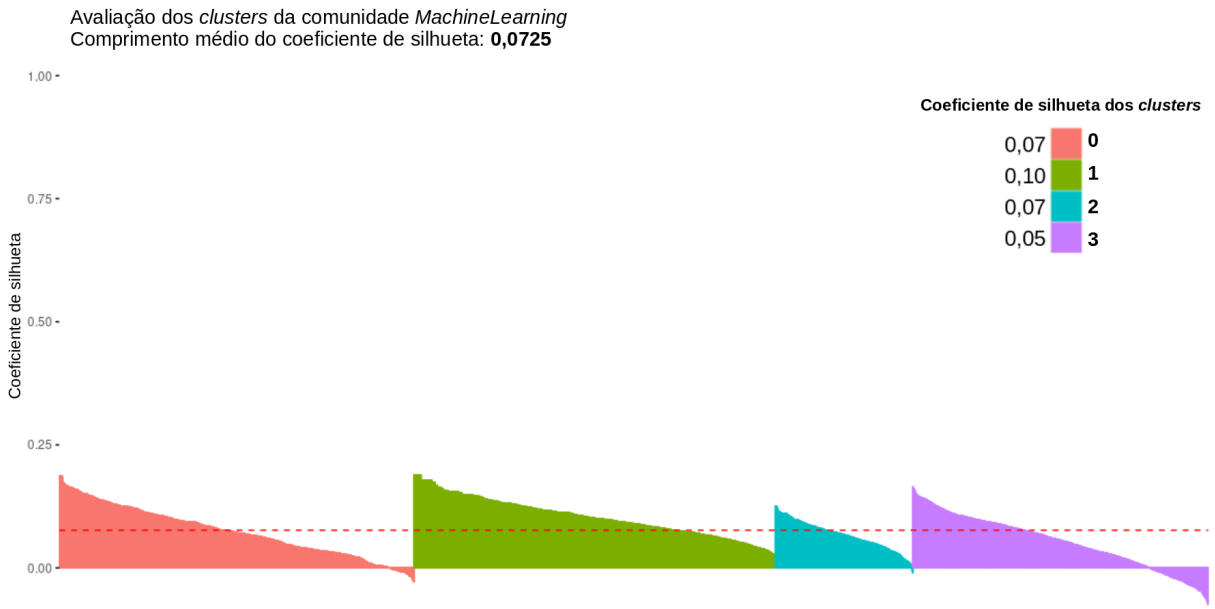
Fonte: Elaborada pelo autor.

Figura 23 – Resultado da avaliação do modelo de clusterização para a comunidade *learnprogramming*



Fonte: Elaborada pelo autor.

Figura 24 – Resultado da avaliação do modelo de clusterização para a comunidade *MachineLearning*



Fonte: Elaborada pelo autor.

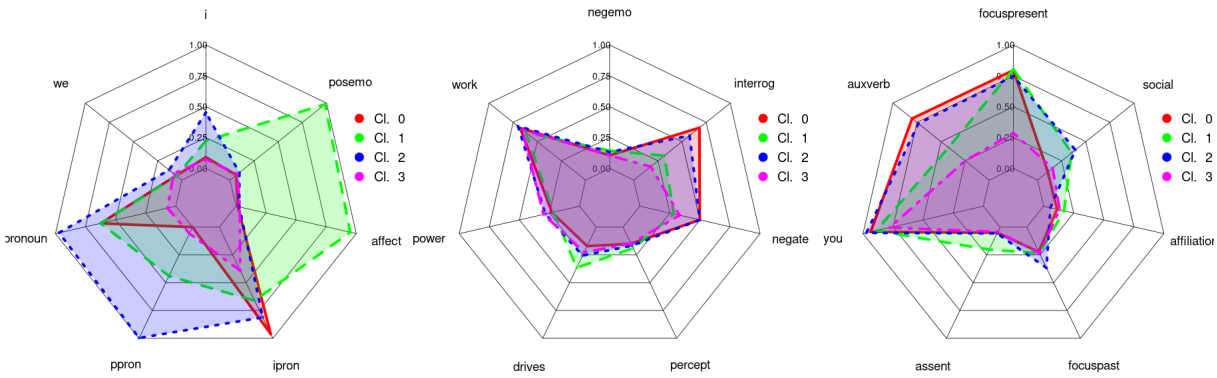
A determinação do *cluster* mais representativo gerado pelo método de clusterização foi guiada pelas diretrizes D.6.3.2 e D.6.3.3. A Tabela 15 mostra os valores médios das métricas estruturadas significativas identificadas durante a avaliação dos modelos das discussões bem avaliadas. Considera-se que os *clusters* com os valores mais altos (destacados em negrito) possuem as discussões mais bem avaliadas pelos participantes. Ao analisar a Tabela 15, é possível evidenciar que o *cluster* 1 apresentou a maior média para todas as métricas na comunidade *learnprogramming*, exceto para o número de gargalos. O motivo pelo qual essa métrica não possui a média mais alta no *cluster* 1 precisa de mais investigação. O *cluster* 2 apresentou a maior média para todas as métricas na comunidade *MachineLearning*. Portanto, o *cluster* 1 e o *cluster* 2 foram identificados como os mais representativos das comunidades *learnprogramming* e *MachineLearning* respectivamente, ou seja, eles possuem os tópicos de discussão mais bem avaliados pelos participantes.

Para identificar as métricas do discurso significativas, foram analisadas as coordenadas normalizadas dos centróides do *cluster* mais representativo de cada comunidade. O centróide é um valor no intervalo $[0, 1]$ que indica o valor médio de cada métrica no *cluster*. Os valores mais significativos são os mais próximos de um. Os resultados, apresentados na Figura 25 e na Figura 26, foram bastante semelhantes para ambas as comunidades. Eles evidenciam oito métricas que podem ser consideradas importantes para identificar um estilo de discurso relacionado às discussões mais bem avaliadas pelos participantes das comunidades: *liwc.posemo liwc.affect*, *liwc.negemo*, *liwc.percept*, *liwc.drives*, *liwc.focuspresent*, *liwc.affiliation* e *liwc.assent*.

Tabela 15 – Valores médios das métricas estruturadas significativas para cada um dos *clusters*

Comunidade *learnprogramming*				
Cluster	*Cluster* 0	***Cluster* 1**	*Cluster* 2	*Cluster* 3
Tamanho do 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
Número de participantes	5,23	**5,85**	4,74	3,69
Tamanho	7,22	**8,63**	6,75	3,88
Comprimento	3,51	**4,17**	3,28	2,50
Diâmetro	2,35	**2,40**	2,29	1,60
Número de triângulos	0,14	**0,15**	0,10	0,05
Número de gargalos	1,89	2,23	2,05	**2,30**
Comunidade *MachineLearning*				
Cluster	*Cluster* 0	*Cluster* 1	***Cluster* 2**	*Cluster* 3
Tamanho do cluster	5.411	3.118	5.636	3.265
% of Total	31,04%	17,89%	32,34%	18,73%
Score	71,58	10,25	**116,77**	48,78
Número de participantes	6,34	2,36	**8,54**	3,82
Tamanho	8,32	1,66	**13,04**	4,46
Comprimento	3,91	1,22	**5,32**	1,92
Diâmetro	2,17	1,10	**2,94**	1,33
Número de triângulos	0,30	0,02	**0,44**	0,39
Número de gargalos	2,45	2,02	**2,50**	2,27

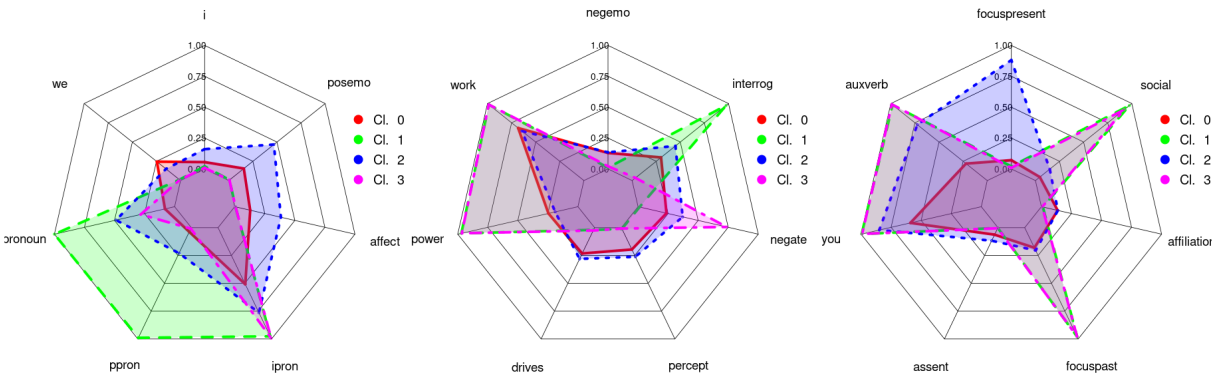
Figura 25 – Resultado da análise dos centróides para o modelo do estilo de discurso na comunidade *learnprogramming*



Fonte: Elaborada pelo autor.

Nota – O ***cluster 1*** apresentou as discussões mais bem avaliadas pelos participantes; as métricas do discurso cujos valores dos centróides no *cluster* 1 foram superiores quando comparados aos demais *clusters* foram as seguintes: *liwc.posemo liwc.affect*, *liwc.negemo*, *liwc.percept*, *liwc.drives*, *liwc.focuspresent*, *liwc.affiliation* e *liwc.assent*.

Figura 26 – Resultado da análise dos centróides para o modelo do estilo de discurso na comunidade *MachineLearning*



Fonte: Elaborada pelo autor.

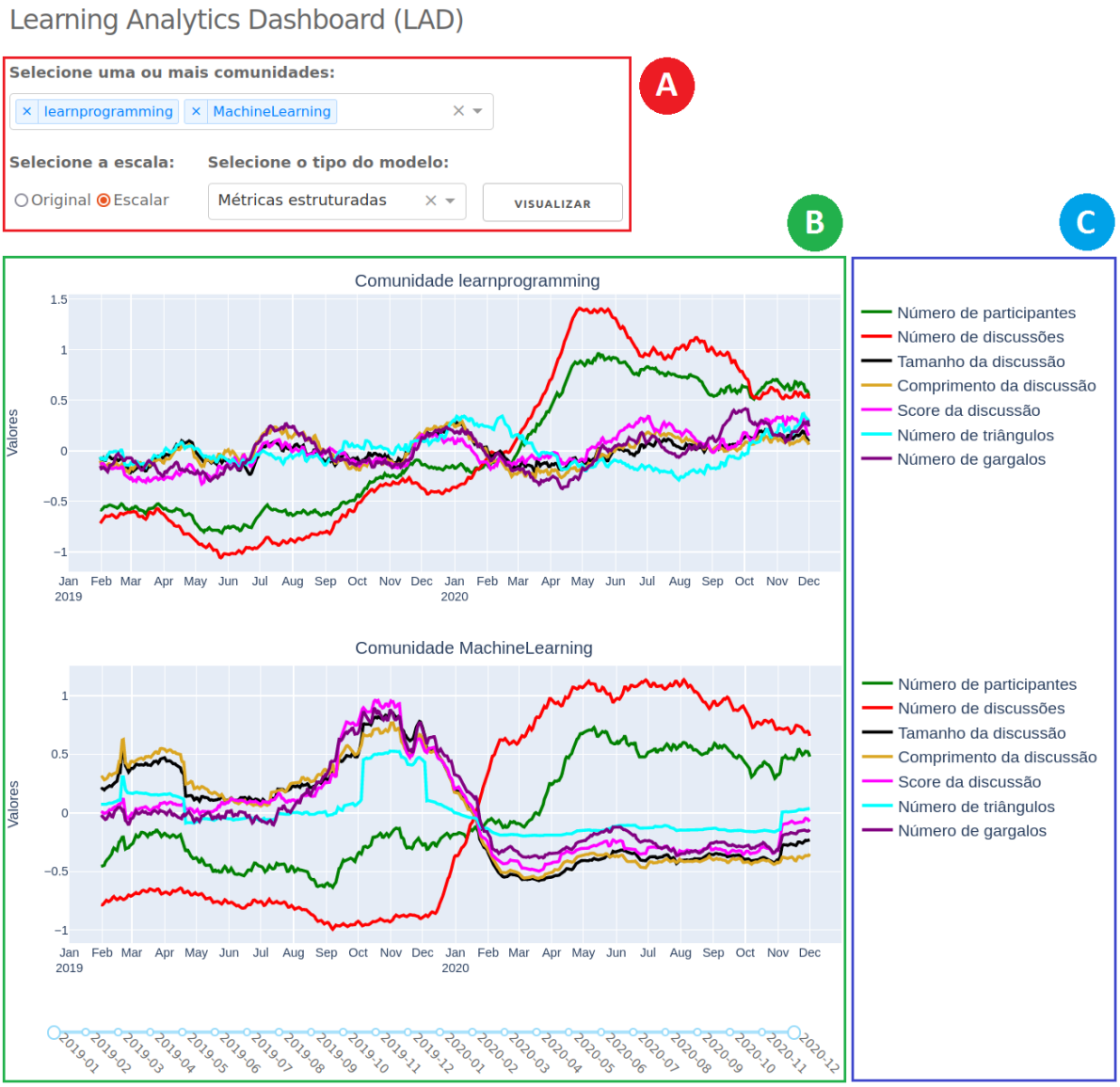
Nota – O ***cluster 2*** apresentou as discussões mais bem avaliadas pelos participantes; as métricas do discurso cujos valores dos centróides no *cluster* 2 foram superiores quando comparados aos demais *clusters* foram as seguintes: *liwc.i*, *liwc.posemo liwc.affect*, *liwc.negemo*, *liwc.percept*, *liwc.drives*, *liwc.focuspresent*, *liwc.affiliation* e *liwc.assent*.

6.3.2.2 Criação e análise dos modelos temporais

O processo SLIM prevê a criação e análise de dois modelos temporais (conforme descrito na subseção 4.3.2.5): ***(i)*** modelo de séries temporais das métricas estruturadas; e ***(ii)*** modelo de séries temporais das métricas do discurso. Após a atividade de criação e treinamento, guiada pelas diretrizes D.8.1.1 a D.8.2.3, o LAD descrito na subseção 6.2.4.1 foi utilizado para visualizar os modelos temporais. O aspecto visual do LAD é apresentado na Figura 27. Seus principais componentes podem ser descritos como segue.

- Definição de parâmetros (Figura 27-**A**) - refere-se aos seguintes parâmetros que configuram a visualização de dados:
  - selecione uma ou mais comunidades - permite selecionar os dados de uma ou mais comunidades, com o objetivo de comparar as suas similaridades e diferenças;
  - selecione a escala - permite aplicar um método para padronizar as métricas que compõem as séries temporais, a fim de apresentar os seus valores na mesma escala;
  - selecione o tipo do modelo - permite selecionar a exibição dos dados das métricas estruturadas ou métricas do discurso.
- Visualização das séries temporais (Figura 27-**B**) - exibe o comportamento ao longo do tempo das métricas estruturadas e de discurso significativas, identificadas por meio da análise dos modelos seccionais. O período de visualização das métricas pode ser alterado para investigar diferentes intervalos de tempo.

Figura 27 – Componentes visuais do LAD



Fonte: Elaborada pelo autor.

- Selecionar métricas (Figura 27-**C**) - permite desabilitar determinadas métricas, com o objetivo de enfatizar diferentes grupos de métricas ao longo da análise.

Antes de prosseguir com a análise de dados visual, as séries temporais foram avaliadas de acordo com as diretrizes D.10.1 a D.10.4. O resultado é apresentado na Tabela 16. Na comunidade *learnprogramming* todas as séries temporais apresentaram qualidade satisfatória, portanto, serão consideradas na atividade de análise visual e exploratória. Na comunidade *MachineLearning* quatro métricas (em itálico) não mostraram qualidade satisfatória, pois apesar do coeficiente ADF possuir valor negativo, elas não apresentaram *p-value* $<0,05$. Assim, as séries temporais relativas às métricas diâmetro, processos perceptivos, foco no presente e afiliações serão desconsideradas na análise visual e exploratória da comunidade *MachineLearning*.

Tabela 16 – Coeficientes do teste ADF para avaliação das séries temporais

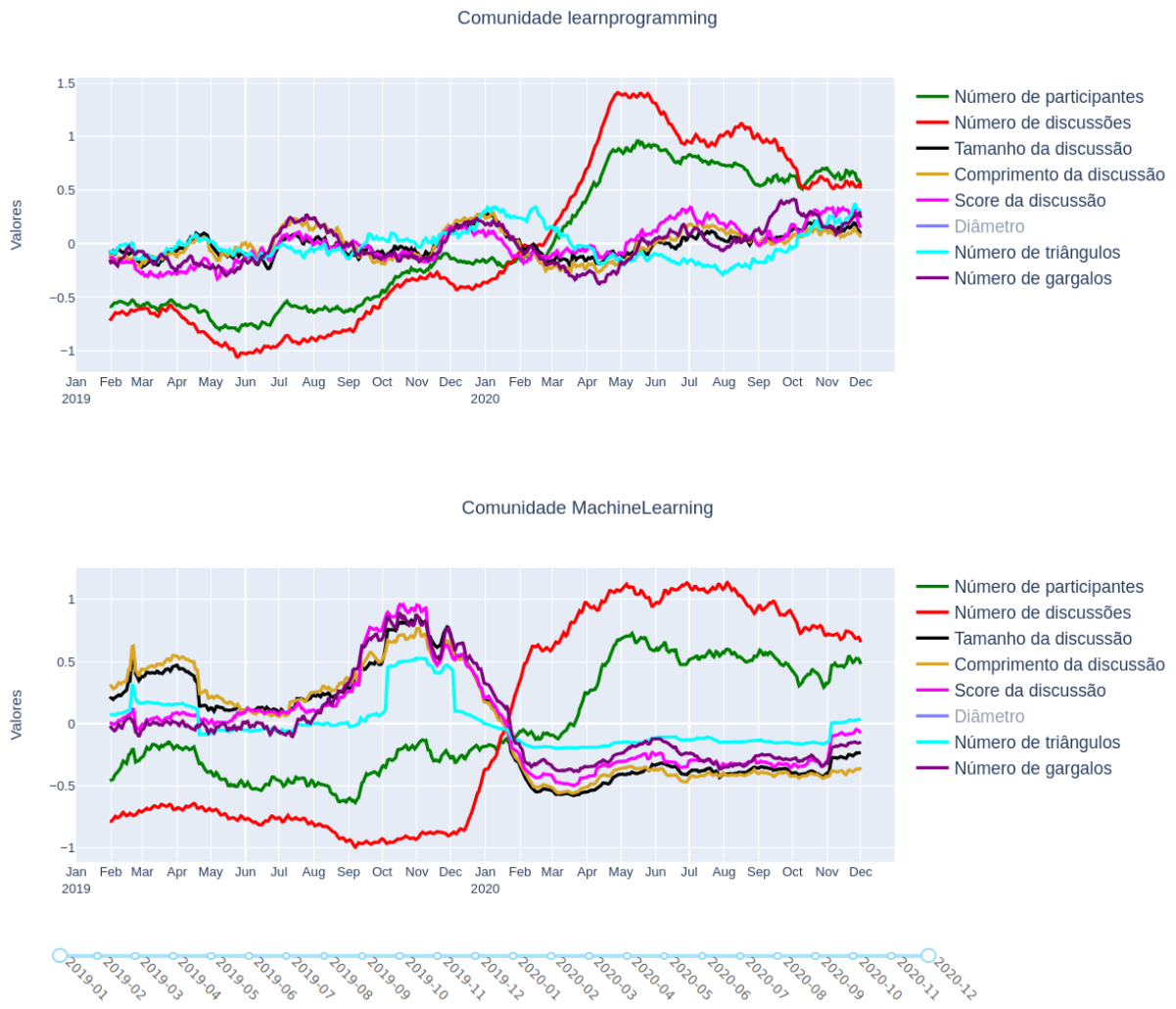
Comunidade *learnprogramming*				
Métricas estruturadas			**Métricas do discurso**	
Score da discussão	-28.34***		Processos afetivos	-5.72***
Número de participantes	-1.81*		Emoção positiva	-11.07***
Número de discussões	-1.25*		Emoção negativa	-26.41***
Tamanho da discussão	-28.61***		Direcionamentos	-5.61***
Comprimento da discussão	-28.86***		Processos perceptivos	-27.48***
Diâmetro	-24.18***		Foco no presente	-7.98***
Número de triângulos	-27.46***		Consentimento	-24.88***
Número de gargalos	-9.84***		Afiliações	-25.35***
Comunidade *MachineLearning*				
Métricas estruturadas			**Métricas do discurso**	
Score da discussão	-3.01***		Processos afetivos	-6.95***
Número de participantes	-2.59*		Emoção positiva	-5.46***
Número de discussões	-1.18*		Emoção negativa	-6.52***
Tamanho da discussão	-3.79***		Direcionamentos	-3.02*
Comprimento da discussão	-3.69***		*Processos perceptivos*	*-2.34*
Diâmetro	*-2.15*		*Foco no presente*	*-1.72*
Número de triângulos	-17.67***		Consentimento	-7.17***
Número de gargalos	-4.94***		*Afiliações*	*-2.43*

Fonte: Elaborada pelo autor.

Nota – *** significa *p-value* $<0,001$; * significa *p-value* $<0,05$.

A Figura 28 mostra as séries temporais que representam o comportamento das métricas estruturadas ao longo do tempo em ambas as comunidades. A métrica diâmetro foi desabilitada no gráfico de *learnprogramming* para facilitar a comparação entre as duas comunidades. A comunidade *learnprogramming* apresentou tendência de crescimento das métricas referentes ao número de participantes e número de discussões. Este cenário também resultou em uma tendência de crescimento das métricas relacionadas à quantidade de participação (tamanho, comprimento e *score* das discussões), embora ela tenha sido menos intensa. Altos níveis de

Figura 28 – Séries temporais das métricas estruturadas

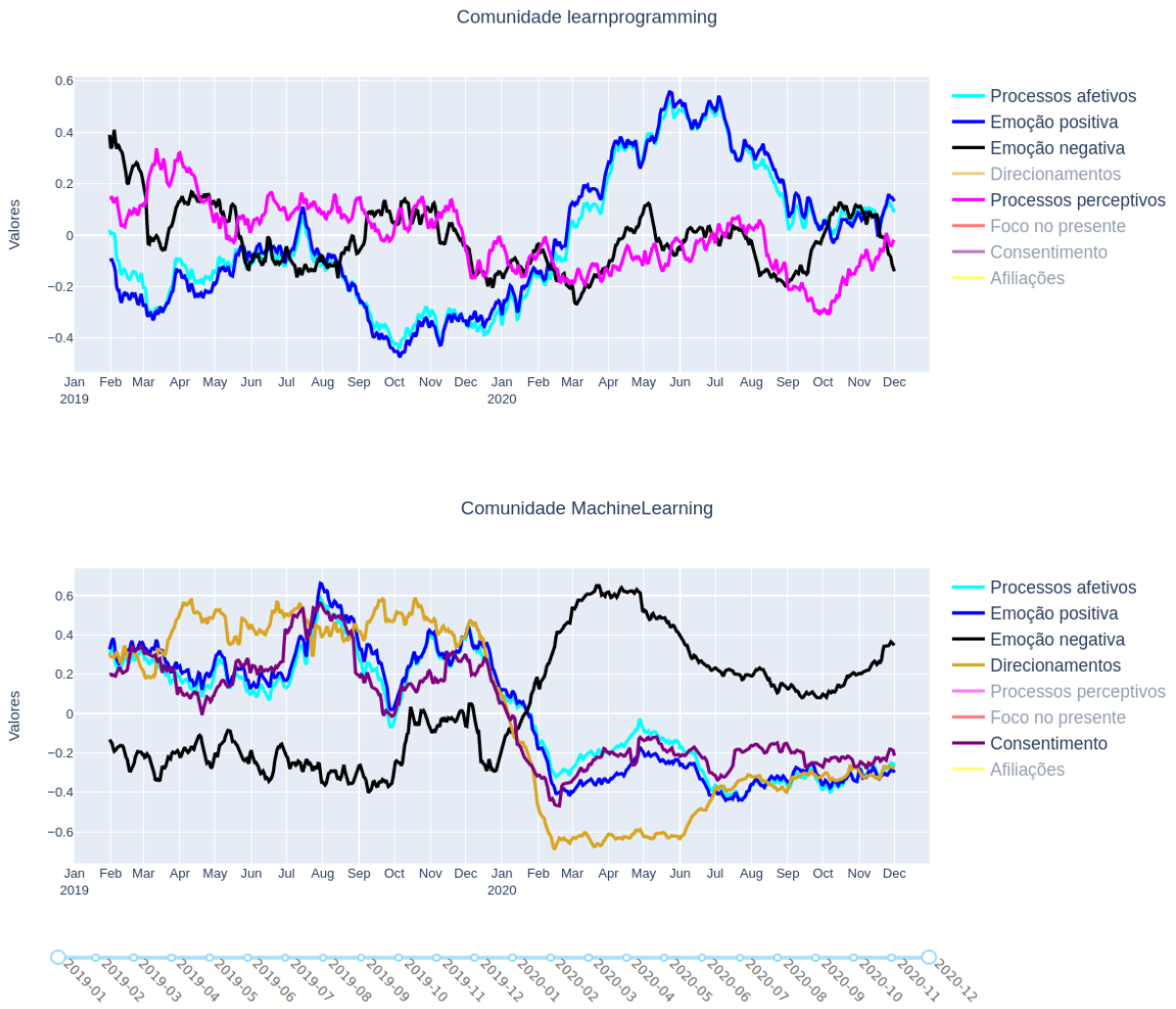


Fonte: Elaborada pelo autor.

atividade e participação em OLC são a chave para o sucesso de tais comunidades. Os participantes são, ao mesmo tempo, produtores e consumidores de informação. Assim, eles têm um papel importante na criação de artefatos de conhecimento e no compartilhamento destes com seus pares (SPEILY *et al.*, 2020). A comunidade *MachineLearning* também apresentou uma tendência de crescimento referente ao número de participantes e o número de discussões. No entanto, esta tendência não resultou em uma maior quantidade de participação, pois as métricas de tamanho, comprimento e *score* das discussões (além do número de triângulos e número de gargalos) apresentaram tendências de diminuição consistentes ao longo do período investigado. A análise da Figura 29, descrita a seguir, pode auxiliar a esclarecer este cenário.

A Figura 29 mostra as séries temporais que representam o comportamento das métricas do discurso ao longo do tempo em ambas as comunidades. Na comunidade *learnprogramming* foram desabilitadas as métricas direcionamentos, foco no presente, consentimento e afiliações pois

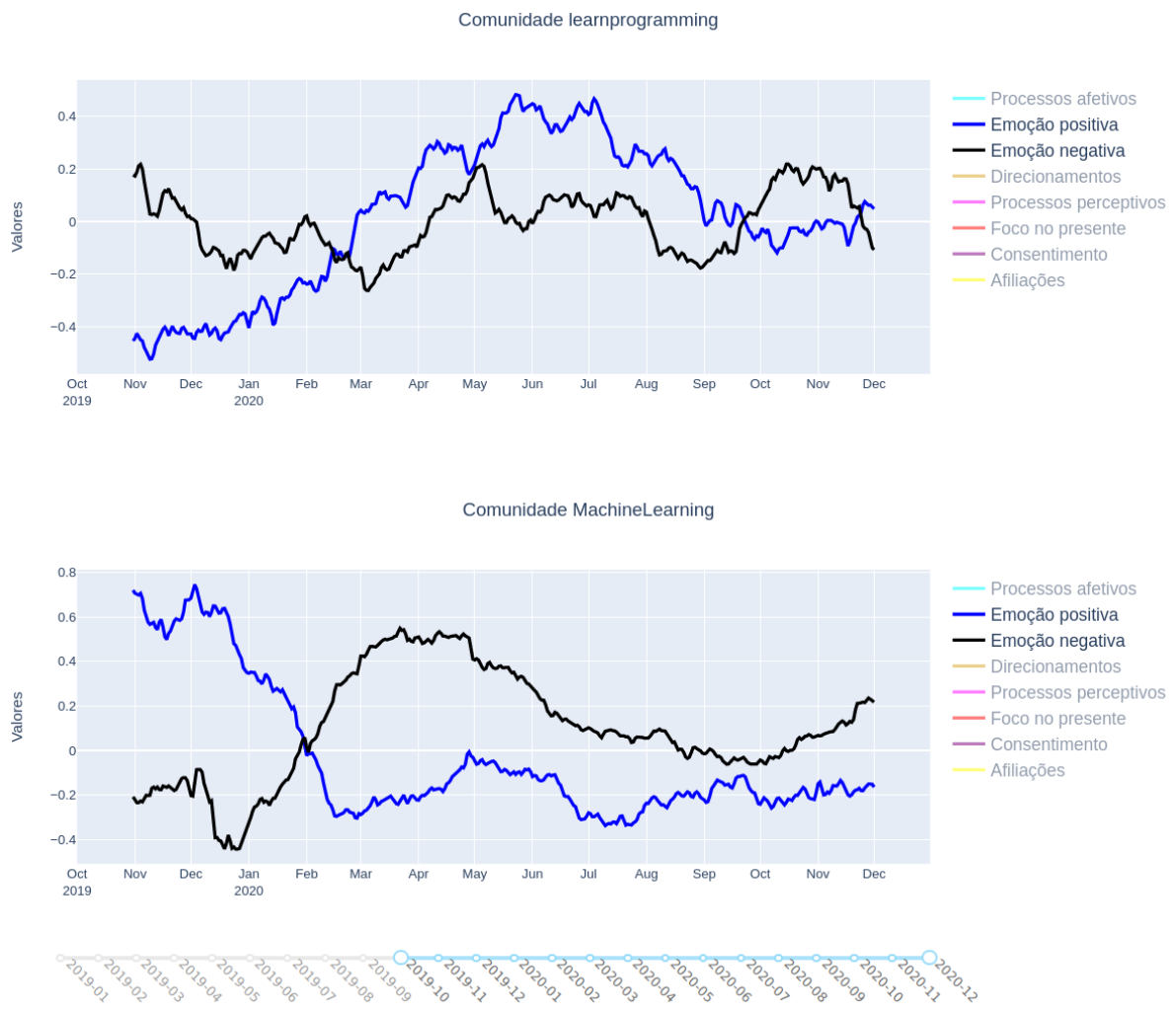
Figura 29 – Séries temporais das métricas do discurso



Fonte: Elaborada pelo autor.

elas não apresentaram tendências crescentes ou decrescentes significativas ao longo do tempo. As métricas emoção positiva (palavras como *love*, *good* e *nice*) e processos afetivos (palavras como *admire*, *dear* e *laugh*) apresentaram tendência de crescimento, enquanto as métricas processos perceptivos (palavras como *feeling*, *hear* e *look*) e emoção negativa (palavras como *angry*, *bad* e *nasty*) apresentaram suave tendência de diminuição. Por outro lado, na comunidade *MachineLearning*, todas as métricas do discurso apresentaram tendência de diminuição, exceto emoção negativa, que manteve-se estável, mas apresentou tendência de crescimento próximo ao início do ano de 2020. Para investigar o papel das emoções positivas e negativas com mais detalhes, apenas estas duas métricas foram consideradas em um intervalo de análise visual encurtado para o período entre outubro de 2019 e dezembro de 2020. O resultado é mostrado na Figura 30. A comunidade *learnprogramming* apresentou alternância nas tendências de crescimento das emoções positiva e negativa, com prevalência da emoção positiva a maior parte do tempo. Por outro lado, *MachineLearning* apresentou uma tendência de crescimento de

Figura 30 – Comportamento das métricas de emoção positiva e emoção negativa no período entre outubro/2019 e dezembro/2020



Fonte: Elaborada pelo autor.

emoção negativa superior à emoção positiva próximo de fevereiro de 2020 até o final do período sob análise.

As emoções desempenham um papel crítico durante o processo de aprendizagem e a resolução de problemas com tecnologias de educacionais (AZEVEDO *et al.*, 2017). Pesquisas experimentais evidenciam que as emoções positivas facilitam o uso de estratégias de aprendizagem flexíveis e criativas, como elaboração, organização e metacognição (ARTINO; JONES, 2012); enquanto a ativação de emoções negativas leva a estratégias mais rígidas, como formas superficiais de processamento de informações (PEKRUN, 2006). Além disso, estudos demonstraram que comentários com a prevalência de emoções negativas em discussões *online* possuem menor probabilidade de receber respostas de outros participantes, e resultaram em menor prestígio para os seus autores (ZOU *et al.*, 2021). Assim, a análise visual e exploratória permitiu identificar um contexto relacionado à expressão de emoções em discussões *online* que poderia levar a

níveis mais baixos de atividade e participação na comunidade *MachineLearning*. Tal contexto foi considerado na elaboração de questões de pesquisa específicas sobre a aprendizagem social *online* em ambientes informais, como descrito a seguir.

Elaboração de questões de pesquisa específicas sobre a aprendizagem social *online* em ambientes informais

Com base nas características das OLC e na análise dos modelos seccionais e temporais, foram elaboradas as quatro questões de pesquisa específicas descritas a seguir. Tais questões apoiaram a formulação de hipóteses, conforme apresentado na próxima subseção.

QP1. Uma característica fundamental das OLC é possibilitar a interação entre participantes com diferentes níveis de conhecimento (SPEILY *et al.*, 2020). Esta característica pode ser investigada por meio da homofilia: um efeito que afirma que os usuários de uma rede interagem majoritariamente com outros semelhantes a si, considerando um determinado atributo (HALBERSTAM; KNIGHT, 2016). Tal efeito reflete a tendência de uma rede criar ligações entre nós que possuem os mesmos atributos. No contexto desta QP, o atributo a ser considerado é o nível de *expertise* do participante (*karma*). Deve-se considerar que uma ligação entre os nós da rede significa interação, ou seja, troca de conhecimento. Portanto, comprovar a inexistência da homofilia significa evidenciar a interação entre participantes especialistas e regulares (não especialistas). Assim, pergunta-se: **as OLC promovem a troca de conhecimento entre participantes especialistas e regulares?**

QP2. O modelo dos participantes especialistas (Tabela 13) correlaciona as métricas de centralidade, como grau, centralidade de intermediação, centralidade de autovetor e *Pagerank* à quantidade de participação, representada pelo grau de saída. Assim, o objetivo desta QP é investigar o processo de formação da rede e identificar se os participantes com os valores mais altos para as métricas de centralidade estão associados à maior quantidade de participação. Neste contexto, questiona-se: **o valor das métricas de centralidade determina a quantidade de participação dos membros da comunidade?**

QP3. O modelo das discussões bem avaliadas (Tabela 14) e o modelo dos participantes especialistas (Tabela 13) evidenciam a importância da participação tanto na avaliação das discussões quanto no reconhecimento de um participante como *expert* pelos seus pares. Assim, pergunta-se: **os participantes especialistas possuem níveis de atividade e popularidade mais altos que os participantes regulares?**

QP4. As séries temporais das métricas do discurso (Figura 29 e Figura 30) evidenciam que a prevalência de emoção negativa está associada ao menor nível de participação e atividade da comunidade *MachineLearning*. Portanto, questiona-se: **períodos com prevalência de emoção negativa produzem menos frequentemente efeitos estruturais de rede associados à aprendizagem social *online*?**

6.3.3 Análise dos dados qualitativos

A análise dos dados qualitativos foi realizada com base nas atividades do processo SLIM relacionadas ao conhecimento, conforme descrito nas próximas subseções.

6.3.3.1 Formulação de hipóteses

As QP específicas sobre aprendizagem social *online* em ambientes informais foram utilizadas para embasar a formulação das hipóteses descritas a seguir.

QP1. As OLC promovem a troca de conhecimento entre participantes especialistas e regulares?

- **H1.1**: Inexiste o efeito de homofilia entre os participantes regulares das OLC.
- **H1.2**: Inexiste o efeito de homofilia entre os participantes especialistas das OLC.
- **H1.3**: Inexiste o efeito de homofilia entre os participantes regulares e especialistas das OLC.

QP2. O valor das métricas de centralidade determina a quantidade de participação dos membros da comunidade?

- **H2.1**: valores mais altos para a centralidade de grau indicam maior quantidade de participação.
- **H2.2**: valores mais altos para a centralidade de intermediação indicam maior quantidade de participação.
- **H2.3**: valores mais altos para a centralidade de autovetor indicam maior quantidade de participação.
- **H2.4**: valores mais altos para a métrica *pagerank* indicam maior quantidade de participação.

QP3. Os participantes especialistas possuem níveis de atividade e popularidade mais altos que os participantes regulares?

- **H3.1**: os participantes especialistas possuem nível de atividade mais alto, ou seja, publicam mais submissões e comentários que os participantes regulares?
- **H3.2**: os participantes especialistas possuem nível de popularidade mais alto, ou seja, recebem mais comentários em suas publicações que os participantes regulares.

QP4. Períodos com prevalência de emoção negativa produzem menos frequentemente efeitos estruturais de rede associados à aprendizagem social *online*?

- **H4.1**: períodos com a prevalência de emoção negativa produzem menos frequentemente o efeito de reciprocidade.

- **H4.2**: períodos com a prevalência de emoção negativa produzem menos frequentemente o efeito de conectividade simples.
- **H4.3**: períodos com a prevalência de emoção negativa produzem menos frequentemente o efeito de propagação da atividade.
- **H4.4**: períodos com a prevalência de emoção negativa produzem menos frequentemente o efeito de propagação da popularidade.
- **H4.5**: períodos com a prevalência de emoção negativa produzem menos frequentemente o efeito de transitividade.
- **H4.6**: períodos com a prevalência de emoção negativa produzem menos frequentemente o efeito de múltipla conectividade.

6.3.3.2 Testes de hipóteses e geração de conhecimento

Nesta subseção são treinados e avaliados os modelos ERG para testar as hipóteses definidas anteriormente. Os critérios para aceitar ou rejeitar uma hipótese são definidos pela diretriz D.13.3. As conclusões dos testes de hipóteses são usadas para gerar conhecimento sobre a aprendizagem social *online* em ambientes informais, de acordo com a diretriz D.14.1. Sobre os modelos ERG apresentados nesta subseção, deve-se fazer as seguintes observações:

(i) as hipóteses relacionadas à **QP1**, **QP2** e **QP3** foram testadas em modelos ERG treinados com base na estrutura do tipo rede que representa as interações dos participantes em todo o período sob investigação; tal rede foi criada de acordo com a diretriz D.3.1; foram atribuídos aos nós o valor da pontuação do *karma* e os valores das métricas de SNA egocêntricas calculadas para cada participante, com o objetivo de avaliar a influência de tais valores na formação da rede;

(ii) foram excluídos da rede descrita no item **(i)** os participantes considerados *outliers*, com base no valor da pontuação do *karma*; assim, de acordo com os critérios apresentados na subseção 5.5, foram excluídos 12.182 e 5.527 participantes na rede das comunidades *learnprogramming* e *MachineLearning*, respectivamente;

(iii) considerando os valores da distribuição da pontuação do *karma* apresentada na Figura 21 (subseção 5.5), foram considerados participantes especialistas aqueles cuja pontuação do *karma* estava acima do limite superior de tal distribuição; assim, foram considerados especialistas 8.546 e 3.054 participantes nas comunidades *learnprogramming* e *MachineLearning*, respectivamente; os especialistas foram identificados por meio do atributo $karma = 1$, e os demais (regulares ou não especialistas) $karma = 0$;

(iv) em todos os modelos ERG foi adicionado um parâmetro do tipo *baseline*, representado pelo termo *edges* da biblioteca para o ambiente R *ergm*; este termo produz um modelo

de rede aleatório que tem o mesmo número de nós da rede observada, e seu objetivo é facilitar a comparação entre diferentes modelos (LUKE, 2015, p. 167).

Com base nestas observações, para cada questão de pesquisa foi criado e treinado um modelo ERG, conforme descrito nos próximos parágrafos.

QP1 [H1.1 a H1.3]: As OLC promovem a troca de conhecimento entre participantes especialistas e regulares?

O resultado do teste de hipóteses da QP1 foi bastante similar para ambas as OLC, o que permitiu inferir as mesmas conclusões para as comunidades *learnprogramming* e *MachineLearning*. A Tabela 17 mostra o modelo ERG que testa as hipóteses de H1.1 a H1.3. Valores de coeficiente positivos indicam que o efeito ocorre mais frequentemente que o esperado em redes aleatórias. Por sua vez, valores negativos indicam que o efeito ocorre menos frequentemente que o esperado. Assim, coeficientes positivos ou negativos podem ser usados para detectar efeitos de rede, no caso da QP1, a homofilia. Todos os parâmetros possuem significância estatística e podem ser interpretados como segue.

Tabela 17 – Modelo ERG que investiga a interação entre participantes especialistas e regulares (**QP1**)

comunidade *learnprogramming*			
Parâmetro ERGM	**Coeficiente**	**Erro padrão**	***z value***
Baseline	-11,41478***	0,01509	-756,48
H1.1 Homofilia (*karma* = 0)	-1,10240***	0,01964	-56,12
H1.2 Homofilia (*karma* = 1)	1,46196***	0,03380	43,25
H1.3 Homofilia (*karma*)	-0,95619***	0,01904	-50,22
comunidade *MachineLearning*			
Parâmetro ERGM	**Coeficiente**	**Erro padrão**	***z value***
Baseline	-10,47511***	0,01463	-715,83
H1.1 Homofilia (*karma* = 0)	-1,04005***	0,01873	-55,53
H1.2 Homofilia (*karma* = 1)	1,42806***	0,03354	42,58
H1.3 Homofilia (*karma*)	-0,91290***	0,01825	-50,03

Fonte: Elaborada pelo autor.

Nota – *** significa *p-value* $<0,001$.

H1.1 O coeficiente do parâmetro é negativo, portanto, não foi detectado o efeito de homofilia entre os participantes regulares. Assim, a hipótese **H1.1 foi aceita**.

H1.2 O coeficiente do parâmetro é positivo, portanto, foi detectado o efeito de homofilia entre os participantes especialistas. Assim, a hipótese **H1.2 foi rejeitada**.

H1.3 O coeficiente do parâmetro é negativo, portanto, não foi detectado o efeito de homofilia entre os participantes regulares e especialistas. Assim, a hipótese **H1.3 foi aceita**.

Conclusão sobre a QP1: o teste da hipótese H1.1 confirma que os participantes regulares interagem majoritariamente com os participantes especialistas. Apesar de H1.2 confirmar que os participantes especialistas interagem majoritariamente entre si, a hipótese H1.3 confirma que predomina a interação entre especialistas e participantes regulares. Portanto, os testes de hipóteses confirmam que as OLC promovem a troca de conhecimento entre participantes com diferentes níveis de conhecimento.

QP2 [H2.1 a H2.4]: o valor das métricas de centralidade determina a quantidade de participação dos membros da comunidade?

As métricas de centralidade atribuídas aos nós da rede de interações possuem valores contínuos. Logo, para reforçar os efeitos discriminantes de tais métricas, elas passaram por um processo de discretização, com base nos seguintes critérios: os valores acima do terceiro quartil (25% com valores mais altos) foram rotulados com 1, e os demais foram rotulados com 0. Assim, foram criadas duas categorias para cada métrica de centralidade. Para aceitar ou rejeitar as hipóteses de H2.1 a H2.4 é necessário comparar os coeficientes de cada categoria das métricas. No modelo ERG, valores de coeficientes mais altos indicam maior probabilidade da quantidade de participação (grau de saída). O resultado do teste das hipóteses da QP2 é mostrado na Tabela 18 e foi similar para ambas as OLC, o que permitiu inferir as mesmas conclusões para as comunidades *learnprogramming* e *MachineLearning*. Todos os parâmetros possuem significância estatística e podem ser interpretados como segue.

H2.1 O valor do coeficiente da categoria $grau = 1$ é maior que o da categoria $grau = 0$. Assim, a hipótese **H2.1 foi aceita**.

H2.2 O valor do coeficiente da categoria $intermediacao = 1$ é maior que o da categoria $intermediacao = 0$. Assim, a hipótese **H2.2 foi aceita**.

H2.3 O valor do coeficiente da categoria $autovetor = 1$ é menor que o da categoria $autovetor = 0$. Assim, a hipótese **H2.3 foi rejeitada**.

H2.4 O valor do coeficiente da categoria $pagerank = 1$ é maior que o da categoria $pagerank = 0$. Assim, a hipótese **H2.4 foi aceita**.

Conclusão sobre a QP2: os testes das hipóteses H2.1, H2.2 e H2.4 confirmaram que as centralidades de grau, intermediação e *pagerank* são determinantes para identificar a quantidade de participação dos membros do comunidade. Por outro lado, valores mais altos para a centralidade de autovetor não determinaram a quantidade de participação, uma vez que a hipótese H2.3 foi rejeitada.

Tabela 18 – Modelo ERG que investiga a influência das métricas de centralidade na quantidade de participação (**QP2**)

comunidade *learnprogramming*				
Parâmetro ERGM		**Coeficiente**	**Erro padrão**	***z value***
Baseline		-10.76908***	0,03253	-331,00
H2.1 Grau	grau = 0	-0,69374***	0,03425	-20,26
	grau = 1	0,48056***	0,03441	13,96
H2.2 Intermediação	intermediacao = 0	-1,24493***	0,02503	-49,74
	intermediacao = 1	1,13866***	0,02503	45,50
H2.3 Autovetor	autovetor = 0	0,75105***	0,02929	25,64
	autovetor = 1	-0,72106***	0,02907	-24,80
H2.4 *Pagerank*	*pagerank* = 0	-0,38835***	0,03324	-11,68
	pagerank = 1	0,31664***	0,03322	9,53
comunidade *MachineLearning*				
Parâmetro ERGM		**Coeficiente**	**Erro padrão**	***z value***
Baseline		-9,47386***	0,03349	-282,897
H2.1 Grau	grau = 0	-0,90902***	0,03069	-29,624
	grau = 1	0,74214***	0,03123	23,77
H2.2 Intermediação	intermediacao = 0	-1,38711***	0,02202	-62,983
	intermediacao = 1	1,31250***	0,02224	59,01
H2.3 Autovetor	autovetor = 0	0,47893***	0,03404	14,068
	autovetor = 1	-0,47494***	0,03395	-13,99
H3.4 *Pagerank*	*pagerank* = 0	-0,20835***	0,03082	-6,759
	pagerank = 1	0,19374***	0,03071	6,31

Fonte: Elaborada pelo autor.

Nota – *** significa *p-value* $<0,001$.

QP3 [H3.1 e H3.2]: Os participantes especialistas possuem níveis de atividade e popularidade mais altos que os participantes regulares?

O resultado do teste de hipóteses da QP3 foi similar para ambas as OLC, o que permitiu inferir as mesmas conclusões para as comunidades *learnprogramming* e *MachineLearning*. Para aceitar ou rejeitar as hipóteses H3.1 e H3.2, é necessário comparar os coeficientes relacionados aos efeitos de propagação da atividade e popularidades dos participantes regulares e especialistas. Valores de coeficientes mais altos indicam maior probabilidade da ocorrência de tais efeitos. O resultado é apresentado na Tabela 19. Todos os parâmetros possuem significância estatística e podem ser interpretados como segue.

H3.1 Nas comunidades *learnprogramming* e *MachineLearning*, o coeficiente do parâmetro de propagação da atividade dos participantes regulares é negativo, enquanto o coeficiente dos participantes especialistas é positivo. Assim, a hipótese **H3.1 foi aceita**.

H3.2 Na comunidade *learnprogramming*, o coeficiente do parâmetro de propagação da popularidade dos participantes regulares é negativo, enquanto o coeficiente dos participantes

Tabela 19 – Modelo ERG que investiga o nível de atividade e popularidade entre participantes especialistas e regulares (**QP3**)

comunidade *learnprogramming*				
Parâmetro ERGM		**Coeficiente**	**Erro padrão**	***z value***
Baseline		-11.01744***	0,01532	-718,945
H3.1 Atividade	Regulares (*karma*=0)	-0,92965***	0,01905	-48,793
	Especialistas (*karma*=1)	0,19973***	0,02996	6,666
H3.2 Popularidade	Regulares (*karma*=0)	-1,28335***	0,01982	-64,754
	Especialistas (*karma*=1)	0,15016***	0,03849	3,901
comunidade *MachineLearning*				
Parâmetro ERGM		**Coeficiente**	**Erro padrão**	***z value***
Baseline		-9,90128***	0,01803	-549,024
H3.1 Atividade	Regulares (*karma*=0)	-0,84984***	0,02656	-31,992
	Especialistas (*karma*=1)	0,16082***	0,04456	3,609
H3.2 Popularidade	Regulares (*karma*=0)	-1,63011***	0,02408	-67,693
	Especialistas (karma=1)	-0,17984***	0,04307	-4,175

Fonte: Elaborada pelo autor.

Nota – *** significa *p-value* $<0,001$.

especialistas é positivo. Na comunidade *MachineLearning*, os coeficientes do parâmetro de propagação da popularidade tanto dos participantes regulares quanto dos especialistas são negativos, contudo, o dos especialistas possui valor mais alto. Neste contexto, a hipótese **H3.2 foi aceita**.

Conclusão sobre a QP3: os testes das hipóteses H3.1 e H3.2 confirmaram a importância da quantidade de participação para que as contribuições dos membros das comunidades sejam reconhecidas pelos seus pares, pois os participantes especialistas apresentaram níveis de atividade e popularidade mais altos que os participantes regulares.

QP4 [H4.1 a H4.6]: Períodos com prevalência de emoção negativa produzem menos frequentemente efeitos estruturais de rede associados à aprendizagem social *online*?

Na rede utilizada para treinamento do modelo ERG referente às hipóteses de H4.1 a H4.6 foram consideradas apenas as interações ocorridas no período de janeiro a dezembro de 2020, pois de acordo com as evidências apresentadas na discussão da Figura 30, tal período mostra uma prevalência da emoção negativa na comunidade *MachineLearning*, enquanto a comunidade *learnprogramming* apresenta alternância entre as emoções negativa e positiva, com prevalência da emoção positiva. Assim, para aceitar ou rejeitar as hipóteses da QP4 é necessário comparar os coeficientes dos efeitos estruturais das duas comunidades. O resultado é apresentado na Tabela 20. Todos os parâmetros possuem significância estatística e podem ser interpretados como segue.

Tabela 20 – Modelo ERG que investiga a menor frequência de efeitos estruturais de redes associados à aprendizagem social *online* na prevalência de emoção negativa (**QP4**)

Parâmetro ERGM		**Coeficiente**	**Erro padrão**	***z value***
H4.1 Reciprocidade	*learnprogramming*	8,095062***	0,334221	24,22
	MachineLearning	6,275010***	0,080510	77,94
H4.2 Conectividade Simples	*learnprogramming*	0,275605***	0,007749	35,57
	MachineLearning	0,036740***	0,000010	3674
H4.3 Atividade	*learnprogramming*	1,613400**	0,610400	2,64
	MachineLearning	1,511380***	0,060460	25,00
H4.4 Popularidade	*learnprogramming*	1,150820***	0,075160	15,31
	MachineLearning	0,860600**	0,333700	2,579
H4.5 Transitividade	*learnprogramming*	14,876856***	0,877302	16,958
	MachineLearning	7,311422***	0,202530	36,100
H4.6 Múltipla conectividade	*learnprogramming*	-1,684824***	0,040402	-41,702
	MachineLearning	-2,270117***	0,012326	-184,177

Fonte: Elaborada pelo autor.

Notas – *** significa *p-value* $<0,001$; ** significa *p-value* $<0,01$.
Baseline para o modelo da comunidade *learnprogramming*: coeficiente: -12,110940***; erro padrão: 0,009202; *z value*: -1316,12.
Baseline para o modelo da comunidade *MachineLearning*: coeficiente: -11,147695***; erro padrão: 0,008741; *z value*: -1275,33.

H4.1 O coeficiente de reciprocidade da comunidade *MachineLearning* é menor que o da comunidade *learnprogramming*. Portanto, **H4.1 foi aceita**.

H4.2 O coeficiente de conectividade simples da comunidade *MachineLearning* é menor que o da comunidade *learnprogramming*. Portanto, **H4.2 foi aceita**.

H4.3 O coeficiente de propagação da atividade da comunidade *MachineLearning* é menor que o da comunidade *learnprogramming*. Portanto, **H4.3 foi aceita**.

H4.4 O coeficiente de propagação da popularidade da comunidade *MachineLearning* é menor que o da comunidade *learnprogramming*. Portanto, **H4.4 foi aceita**.

H4.5 O coeficiente de transitividade da comunidade *MachineLearning* é menor que o da comunidade *learnprogramming*. Portanto, **H4.5 foi aceita**.

H4.6 Os coeficientes de múltipla conectividade das duas comunidades são negativos, contudo, o de *MachineLearning* é menor que o da comunidade *learnprogramming*. Portanto, **H4.6 foi aceita**.

Conclusão sobre a QP4: os testes das hipóteses de H4.1 a H4.6 mostram que todos os coeficientes da comunidade *MachineLearning* são menores que o de *learnprogramming* no período investigado. Nesse contexto, é possível evidenciar que a prevalência da emoção negativa está

associada à produção menos frequente de efeitos estruturais de rede relacionados à aprendizagem social *online*. Este resultado está em conformidade com as evidências apresentadas em estudos que descrevem o papel das emoções positivas e negativas em outros contextos de aprendizagem (AZEVEDO *et al.*, 2017; XIONG; FENG; TANG, 2020; ZOU *et al.*, 2021).

6.4 Considerações finais

Este capítulo apresentou os resultados de um estudo de caso para o processo SLIM. O objetivo de tal estudo foi investigar a efetividade da aplicação do processo SLIM na geração de conhecimento sobre a aprendizagem social em ambientes informais, por meio da avaliação de duas comunidades *online* da rede social Reddit e realização de testes de hipóteses. Com base na questão de pesquisa que motivou e guiou este trabalho de doutorado e na avaliação dos modelos criados pelo processo SLIM, foram formuladas 15 hipóteses com o objetivo de identificar padrões comportamentais associados à aprendizagem social *online*. Após o teste de tais hipóteses, 13 delas foram aceitas e suas conclusões permitiram gerar conhecimento sobre o comportamento dos participantes em ambientes informais de aprendizagem, conforme descrito no próximo capítulo.

CAPÍTULO

7

CONCLUSÕES

7.1 Considerações iniciais

Este trabalho de doutorado apresentou o SLIM, um processo formado por um conjunto de atividades e diretrizes para guiar a análise de dados educacionais produzidos pelos participantes de ambientes informais de aprendizagem, como as redes sociais e as comunidades *online*. Seu objetivo é evidenciar como a aprendizagem social ocorre nos contextos informais, por meio da avaliação de comunidades *online*. Neste cenário, a aplicação do SLIM responde a principal questão de pesquisa formulada no trabalho de doutorado: "como a análise dos dados gerados pelos participantes de comunidades *online* pode contribuir para ampliar a compreensão e gerar conhecimento sobre a aprendizagem social em ambientes informais de aprendizagem?"

Os resultados do estudo de caso apresentado no capítulo 6 possibilitaram identificar evidências e inferir conclusões que ampliaram o entendimento sobre como a aprendizagem social *online* acontece nos ambientes informais. Tais evidências e conclusões foram inferidas por meio dos seguintes pontos: ***(i)*** a definição de um conjunto de métricas que identifica condições associadas à aprendizagem social; ***(ii)*** a criação de um conjunto de modelos que reconhece padrões comportamentais associados à aprendizagem social; ***(iii)*** a especificação de atividades e diretrizes que orientam a análise dos dados extraídos das comunidades *online*; e ***(iv)*** a formulação e teste de hipóteses, baseadas na análise dos dados educacionais, que geram conhecimento sobre a aprendizagem social. Esses pontos atendem a necessidade de pesquisadores e profissionais na compreensão das atividades de aprendizagem em ambientes informais, uma área ainda sub-representada na literatura sobre tecnologias educacionais.

O restante deste capítulo está estruturado como descrito a seguir. A seção 7.2 apresenta as principais contribuições do trabalho de doutorado. A seção 7.3 descreve as limitações da pesquisa desenvolvida. A seção 7.4 sugere possíveis trabalhos futuros. Por fim, a seção 7.5 apresenta as publicações resultantes do trabalho.

7.2 Contribuições da pesquisa

As principais contribuições da pesquisa realizada neste trabalho de doutorado são sumarizadas como descrito a seguir.

- **Caracterização do estado da arte da avaliação de comunidades *online*** - um MS foi conduzido para caracterizar o cenário atual da avaliação de OLC em ambientes informais de aprendizagem (SILVA; GIMENES; MALDONADO, 2020c). Além disso, foi realizada uma revisão de literatura com o objetivo de complementar as lacunas de pesquisa evidenciadas pelo MS.

- **Definição de um conjunto de métricas associadas à aprendizagem social *online*** - um conjunto de métricas quantitativas foi definido com o objetivo de identificar, a partir dos dados coletados, condições associadas à aprendizagem social *online*. Tais métricas foram sugeridas com base nos *frameworks* teóricos de criação de valor e presença social, descritos na fundamentação teórica desta tese de doutorado. Assim, elas podem ser utilizadas para avaliar as interações sociais (SILVA; GIMENES; MALDONADO, 2020b) e o discurso (SILVA; GIMENES; MALDONADO, 2020a) produzido pelos participantes dos ambientes informais de aprendizagem.

- **Definição de um conjunto de modelos associados à aprendizagem social *online*** - um conjunto de modelos de ML e do tipo ERG foi criado, treinado, avaliado e disponibilizado com o objetivo de identificar nos dados coletados padrões comportamentais e efeitos estruturais de rede associados à aprendizagem social *online*.

- **Definição de um processo para guiar a análise de dados em ambientes informais** - com base nas lacunas de pesquisa descritas no capítulo 3 desta tese de doutorado, foi definido o processo SLIM, com o objetivo de sistematizar a avaliação da aprendizagem social *online* em ambientes informais (SILVA; GIMENES; MALDONADO, 2020a). SLIM é formado por um conjunto de atividades e diretrizes que guiam o analista de dados educacionais na avaliação das OLC estabelecidas em contextos informais.

- **Automatização do processo SLIM** - com o objetivo de automatizar as principais atividades do processo SLIM, foram implementados e disponibilizados os seguintes elementos: ***(i)*** um *script* para automatizar a extração, limpeza, transformação, cálculo das métricas e armazenamento dos dados de comunidades de aprendizagem *online*; ***(ii)*** um *pipeline* para automatizar as atividades de criação, treinamento e avaliação dos modelos de ML que apoiam o processo SLIM; e ***(iii)*** um LAD para ampliar a cognição e o entendimento sobre a dinâmica temporal das OLC investigadas (SILVA; GIMENES; MALDONADO, 2021). Os elementos implementados podem ser utilizados para apoiar, de forma automática, a avaliação de comunidades *online* de outros contextos informais de aprendizagem, como as

redes sociais *Twitter* e *Facebook*, ou comunidades do tipo Q&A como *Stack Overflow*[1], desde que os dados sejam armazenados no formato especificado pelo processo SLIM.

- **Caracterização de comunidades *online* da rede social Reddit** - no capítulo 5 desta tese de doutorado é apresentada a caracterização de duas comunidades de aprendizagem *online* da rede social Reddit, com a descrição dos seus participantes, moderadores, nível de atividade e principais temas discutidos. Tal caracterização é importante, pois apresenta inúmeros gráficos e estatísticas que mostram como reconhecer efetivamente um grupo de usuários como uma comunidade de aprendizagem *online*.

- **Geração de conhecimento sobre a aprendizagem social *online* em ambientes informais** - no estudo de caso sobre a aplicação do processo SLIM, foram elaboradas quatro QP específicas e formuladas quinze hipóteses baseadas nos resultados da avaliação das comunidades *online*. As conclusões inferidas a partir do teste destas hipóteses permitiram criar um entendimento mais amplo sobre como a aprendizagem social ocorre em ambientes informais, como sumarizado a seguir:

  - a participação, o nível de atividade e popularidade são conceitos-chaves na aprendizagem social; os testes de hipóteses confirmaram que tais conceitos são determinantes para que os membros da comunidade reconheçam entre os seus pares os especialistas que detêm um maior nível de conhecimento;

  - as métricas de centralidade de grau, intermediação e *pagerank* foram reconhecidas como determinantes para identificar a quantidade de participação dos membros da comunidade; assim, os participantes que ocupam uma posição centralizada na rede são aqueles que interagem mais frequentemente com os seus pares;

  - as comunidades de aprendizagem *online* promovem a interação entre membros especialistas e regulares (não especialistas), o que evidencia a troca de informações entre participantes com diferentes níveis de conhecimento; tal conclusão foi inferida pois os testes de hipóteses identificaram a inexistência do efeito de homofilia entre participantes especialistas e regulares.

  - os testes de hipóteses permitiram evidenciar a influência das emoções positiva e negativa na produção de efeitos estruturais de rede associados à aprendizagem social; tais evidências foram detectadas pois períodos com a prevalência da emoção negativa sobre a positiva produziram, por exemplo, menor nível de interação, menor quantidade de respostas às discussões publicadas, menor conectividade local e menor propensão à participação.

[1] http://stackoverflow.com

7.3 Limitações da pesquisa

Entre as limitações da pesquisa, pode-se destacar as descritas a seguir.

- **Uso de métodos de observação indireta**: o processo SLIM utiliza o método indireto de coleta de dados. A literatura sobre tecnologias educacionais apresenta fortes evidências que a análise dos dados produzidos pelos participantes das OLC, como a pontuação referente a reputação dos participantes e os votos positivos associados a um tópico de discussão e às suas respostas, podem ser usados para avaliar a frequência e a qualidade das contribuições dos membros das comunidades *online*. A análise em larga escala envolvendo uma grande quantidade de participantes e tópicos de discussão, por um longo período de tempo (acima de um ano), tende a reduzir possíveis ruídos nesses dados e aumentar a probabilidade que eles reflitam a *expertise* dos participantes e a qualidade das interações. Contudo, métodos de observação direta, como a aplicação de questionários que investiguem a efetividade da aprendizagem social, podem ser usados para complementar os resultados apresentados neste trabalho de doutorado.

- **Análise do discurso com número limitado de atributos**: o processo SLIM utiliza na análise do discurso um subconjunto dos atributos do *framework* LIWC. Existem diversos estudos que evidenciam a relevância de tais atributos na identificação da presença social em discussões assíncronas *online*. No entanto, pode-se investigar a influência que outros atributos do *framework* LIWC possuem no discurso gerado pelos participantes das OLC investigadas.

7.4 Trabalhos futuros

Com base nas limitações discutidas previamente e em oportunidades de pesquisa descritas na literatura sobre tecnologias educacionais, identificou-se algumas possibilidades para a condução de trabalhos futuros no contexto desta tese de doutorado, como descrito a seguir.

- **Inclusão de métodos de observação direta**: a área da pesquisa educacional possui alguns instrumentos, no formato de questionários, que podem ser usados para investigar a efetividade da aprendizagem social em discussões assíncronas. Tais instrumentos podem ser aplicados entre participantes das OLC analisadas, com o objetivo de ampliar as possibilidades de generalização das evidências apresentadas neste trabalho de doutorado.

- **Aprimoramento da análise do discurso**: investigar a influência de outros atributos do *framework* LIWC no discurso gerado pelos participantes de grandes comunidades de aprendizagem *online*. Adicionalmente, podem ser utilizados métodos específicos para a investigação da emoção positiva e negativa, como a análise de sentimentos.

- **Criação de novos modelos**: incluir no processo SLIM novos modelos de ML que reconheçam comportamentos associados à aprendizagem social não investigados atualmente, como o conceito da Legitimação da Participação Periférica (LPP). Considerando a evolução da rede de interações ao longo do tempo, a LPP identifica indivíduos que deixam uma posição periférica rumo a uma posição central da comunidade, onde encontram-se os participantes especialistas. Diversos estudos descrevem a necessidade de investigar o tempo necessário para que os membros de uma comunidade sejam reconhecidos como especialistas pelos seus pares.

- **Formulação e teste de novas hipóteses**: com base nos elementos que automatizam a aplicação do processo SLIM, outras comunidades de aprendizagem podem ser caracterizadas e avaliadas. A avaliação de tais comunidades pode permitir a elaboração de novas questões de pequisa, a formulação e o teste de hipóteses que podem gerar novos conhecimentos e ampliar o entendimento sobre a aprendizagem social *online* em ambientes informais.

- **Criação de diretrizes para incrementar o nível de atividade e participação**: de forma geral, as abordagens que visam a avaliar comunidades de aprendizagem oferecem resultados que mensuram um ou mais dos seus aspectos, como o nível de atividade e participação. No entanto, comumente tais abordagens não oferecem diretrizes que podem guiar ações com base no resultado dos aspectos mensurados. Neste contexto, pode-se adicionar ao processo SLIM atividades que orientem os responsáveis pela comunidade a como regular os níveis de atividade e participação.

- **Investigação do impacto da integração entre ambientes formais e informais de aprendizagem**: a área de pesquisa em tecnologias educacionais descreve amplamente as vantagens do uso dos ambientes informais de aprendizagem. Alguns destes benefícios, como a interação entre participantes com diferentes níveis de conhecimento, foram evidenciados por meio dos testes de hipóteses apresentados no capítulo 6. Neste contexto, podem ser propostos experimentos que investiguem o impacto do uso dos contextos informais no desempenho dos estudantes em ambientes formais de aprendizagem.

7.5 Publicações resultantes

As publicações resultantes das atividades conduzidas neste trabalho de doutorado são descritas como segue.

- SILVA, R. F.; GIMENES, I. M. S.; MALDONADO, J. C. The challenge of evaluating virtual communities of practice: a systematic mapping study. In: **Interdisciplinary Journal of Information, Knowledge, and Management**. 2020, v. 15, n. 1, p. 39–64.

- SILVA, R. F.; GIMENES, I. M. S.; MALDONADO, J. C. Assessing virtual communities of practice in informal learning environments. In: **2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)**. 2020, p. 117–121.
- SILVA, R. F.; GIMENES, I. M. S.; MALDONADO, J. C. An approach for assessing online communities in informal learning environments. In: **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. 2020, p. 642–651.
- SILVA, R. F.; GIMENES, I. M. S.; MALDONADO, J. C. Analyzing learners' behavior and discourse within large online communities: a Social Learning Analytics Dashboard. In: **XXXII Simpósio Brasileiro de Informática na Educação**. 2021.

Encontram-se em avaliação as publicações descritas a seguir.

- SILVA, R. F.; GIMENES, I. M. S.; MALDONADO, J. C. SLIM: a process for assessing social learning within large online communities in informal environments. In: **Journal of Universal Computer Science**, 2021.

REFERÊNCIAS

AGUINIS, H.; GOTTFREDSON, R. K.; JOO, H. Best-practice recommendations for defining, identifying, and handling outliers. **Organizational Research Methods**, Sage Publications Sage CA: Los Angeles, CA, v. 16, n. 2, p. 270–301, 2013. Citado na página 78.

AHN, J. What Can We Learn from Facebook Activity ? Using Social Learning Analytics to Observe New Media Literacy Skills. In: **Proceedings of the third international conference on learning analytics and knowledge**. [S.l.: s.n.], 2013. p. 135–144. Citado nas páginas 23, 24, 25, 31, 38, 56 e 62.

ANAYA, A. R.; GONZÁLEZ-BOTICARIO, J.; LETÓN, E.; OLMO, F. Hernández-del. An approach of collaboration analytics in moocs using social network analysis and influence diagram. In: **EDM**. [S.l.: s.n.], 2015. p. 492–495. Citado na página 41.

ANTONE, B.; GUPTA, A.; BELL, S.; DECHURCH, L.; CONTRACTOR, N. Testing influence of network structure on team performance using stergm-based controls. In: SPRINGER. **International Conference on Complex Networks and Their Applications**. [S.l.], 2019. p. 1018–1030. Citado na página 66.

ARTASANCHEZ, A.; JOSHI, P. **Artificial Intelligence with Python: Your complete guide to building intelligent apps using Python 3. x**. [S.l.]: Packt Publishing Ltd, 2020. Citado na página 75.

ARTINO, A. R.; JONES, K. D. Exploring the complex relations between achievement emotions and self-regulated learning behaviors in online learning. **The Internet and Higher Education**, Elsevier, v. 15, n. 3, p. 170–175, 2012. Citado na página 125.

AZEVEDO, R. Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. **Educational psychologist**, Taylor & Francis, v. 50, n. 1, p. 84–94, 2015. Citado na página 46.

AZEVEDO, R.; TAUB, M.; MUDRICK, N. V.; MILLAR, G. C.; BRADBURY, A. E.; PRICE, M. J. Using data visualizations to foster emotion regulation during self-regulated learning with advanced learning technologies. In: **Informational environments**. [S.l.]: Springer, 2017. p. 225–247. Citado nas páginas 125 e 134.

BARBOSA, A.; FERREIRA, M.; MELLO, R. F.; LINS, R. D.; GASEVIC, D. The impact of automatic text translation on classification of online discussions for social and cognitive presences. In: **LAK21: 11th International Learning Analytics and Knowledge Conference**. [S.l.: s.n.], 2021. p. 77–87. Citado na página 71.

BHOLOWALIA, P.; KUMAR, A. Ebk-means: A clustering technique based on elbow method and k-means in wsn. **International Journal of Computer Applications**, Citeseer, v. 105, n. 9, 2014. Citado na página 82.

BILBRO, R.; OJEDA, T.; BENGFORT, B. **Applied text analysis with Python**. [S.l.]: O'reilly, 2019. Citado na página 61.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **The Journal of machine Learning research**, JMLR. org, v. 3, p. 993–1022, 2003. Citado na página 100.

BOEKAERTS, M.; MINNAERT, A. Self-regulation with respect to informal learning. **International journal of educational research**, Elsevier, v. 31, n. 6, p. 533–544, 1999. Citado na página 56.

BRINT, S. Gemeinschaft revisited: A critique and reconstruction of the community concept. **Sociological theory**, Wiley Online Library, v. 19, n. 1, p. 1–23, 2001. Citado na página 33.

CAI, Z.; EAGAN, B.; DOWELL, N. M.; PENNEBAKER, J. W.; SHAFFER, D. W.; GRAESSER, A. C. Epistemic network analysis and topic modeling for chat data from collaborative learning environment. In: **Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017**. [S.l.: s.n.], 2017. p. 104–111. Citado na página 63.

CAPLAN, S. E.; TURNER, J. S. Bringing theory to research on computer-mediated comforting communication. **Computers in human behavior**, Elsevier, v. 23, n. 2, p. 985–998, 2007. Citado na página 90.

CARRINGTON, P. J.; SCOTT, J.; WASSERMAN, S. **Models and methods in social network analysis**. [S.l.]: Cambridge university press, 2005. v. 28. Citado na página 45.

CELA, K. L.; SICILIA, M. Á.; SÁNCHEZ, S. Social network analysis in e-learning environments: A preliminary systematic review. **Educational Psychology Review**, Springer, v. 27, n. 1, p. 219–246, 2015. Citado nas páginas 40, 65 e 66.

CHARNES, A.; COOPER, W. W.; RHODES, E. Measuring the efficiency of decision making units. **European journal of operational research**, Elsevier, v. 2, n. 6, p. 429–444, 1978. Citado na página 59.

CHATTI, M. A.; DYCKHOFF, A. L.; SCHROEDER, U.; THÜS, H. A reference model for learning analytics. **International Journal of Technology Enhanced Learning**, Inderscience Publishers, v. 4, n. 5-6, p. 318–331, 2012. Citado nas páginas 35, 36, 37 e 67.

CHATTI, M. A.; MUSLIM, A.; SCHROEDER, U. Toward an open learning analytics ecosystem. In: DANIEL, B. K. (Ed.). **Big data and learning analytics in higher education**. [S.l.]: Springer International Publishing, 2017. p. 195–219. Citado nas páginas 24 e 35.

CHEN, B.; CHANG, Y.-H.; OUYANG, F.; ZHOU, W. Fostering student engagement in online discussion through social learning analytics. **The Internet and Higher Education**, Elsevier, v. 37, p. 21–30, 2018. Citado nas páginas 38, 63 e 74.

CHEN, B.; HUANG, T. It is about timing: Network prestige in asynchronous online discussions. **Journal of Computer Assisted Learning**, Wiley Online Library, v. 35, n. 4, p. 503–515, 2019. Citado na página 30.

CHEN, X.; VORVOREANU, M.; MADHAVAN, K. Mining social media data for understanding students' learning experiences. **IEEE Transactions on learning technologies**, IEEE, v. 7, n. 3, p. 246–259, 2014. Citado nas páginas 56 e 57.

CHUANG, J.; MANNING, C. D.; HEER, J. Termite: Visualization techniques for assessing textual topic models. In: **Proceedings of the international working conference on advanced visual interfaces**. [S.l.: s.n.], 2012. p. 74–77. Citado na página 101.

CORBI, A.; BURGOS, D. How to integrate formal and informal settings in massive open online courses through a transgenic learning approach. In: **Radical Solutions and Learning Analytics**. [S.l.]: Springer, 2020. p. 173–191. Citado na página 31.

CZERKAWSKI, B.; HERNÁNDEZ, J. N. Formal, non-formal, and informal e-learning experiences with emerging technologies: a case study of a graduate educational technology program. In: **Cases on formal and informal e-learning environments: Opportunities and practices**. [S.l.]: IGI Global, 2013. p. 337–355. Citado na página 31.

CZERKAWSKI, B. C. Blending Formal and Informal Learning Networks for Online Learning. **The International Review of Research in Open and Distributed Learning**, v. 17, n. 3, 2016. Citado nas páginas 30 e 31.

DADO, M.; BODEMER, D. A review of methodological applications of social network analysis in computer-supported collaborative learning. **Educational Research Review**, Elsevier, v. 22, p. 159–180, 2017. Citado nas páginas 40, 65 e 66.

DASCALU, M.; MCNAMARA, D. S.; TRAUSAN-MATU, S.; ALLEN, L. K. Cohesion network analysis of CSCL participation. **Behavior Research Methods**, v. 50, n. 2, p. 604–619, 2018. Citado nas páginas 46, 62, 63 e 65.

DAWSON, S. Analytics to literacies: Emergent learning analytics to evaluate new literacies. In: **Workshop on New Media, New Literacies, and New Forms of Learning**. [S.l.: s.n.], 2011. Citado na página 38.

DE-MARCOS, L.; GARCÍA-LÓPEZ, E.; GARCÍA-CABOT, A.; MEDINA-MERODIO, J.-A.; DOMÍNGUEZ, A.; MARTÍNEZ-HERRÁIZ, J.-J.; DIEZ-FOLLEDO, T. Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance. **Computers in Human Behavior**, Elsevier, v. 60, p. 312–321, 2016. Citado nas páginas 40, 41 e 43.

De Wever, B.; SCHELLENS, T.; VALCKE, M.; Van Keer, H. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. **Computers and Education**, v. 46, n. 1, p. 6–28, 2006. Citado nas páginas 30 e 46.

DEDE, C. Comparing frameworks for 21st century skills. In: **21st century skills: Rethinking how students learn**. 20. ed. Bloomington: Solution Tree Press, 2010. p. 51–76. Citado na página 23.

DUBOIS, C.; BUTTS, C.; SMYTH, P. Stochastic blockmodeling of relational event dynamics. In: PMLR. **Artificial intelligence and statistics**. [S.l.], 2013. p. 238–246. Citado na página 45.

ESHACH, H. Bridging in-school and out-of-school learning: Formal, non-formal, and informal education. **Journal of science education and technology**, Springer, v. 16, n. 2, p. 171–190, 2007. Citado na página 31.

FARROW, E.; MOORE, J.; GAŠEVIĆ, D. Dialogue attributes that inform depth and quality of participation in course discussion forums. In: **Proceedings of the Tenth International Conference on Learning Analytics & Knowledge**. [S.l.: s.n.], 2020. p. 129–134. Citado nas páginas 42 e 44.

FERREIRA, M.; ROLIM, V.; MELLO, R. F.; LINS, R. D.; CHEN, G.; GAŠEVIĆ, D. Towards automatic content analysis of social presence in transcripts of online discussions. In: **Proceedings of the Tenth International Conference on Learning Analytics & Knowledge**. [S.l.: s.n.], 2020. p. 141–150. Citado nas páginas 30, 47, 49, 51, 62, 63, 65 e 71.

FINCHAM, E.; GAŠEVIĆ, D.; PARDO, A. From Social Ties to Network Processes: Do Tie Definitions Matter? **Journal of Learning Analytics**, v. 5, n. 2, p. 9–28, 2018. Citado nas páginas 45, 62 e 63.

GALANIS, N.; MAYOL, E.; ALIER, M.; GARCÍA-PEÑALVO, F. J. Supporting, evaluating and validating informal learning. a social approach. **Computers in Human Behavior**, Elsevier, v. 55, p. 596–603, 2016. Citado nas páginas 30, 31, 33, 55, 56 e 73.

GARCIA-PENALVO, F. J. Learning Analytics as a Breakthrough in Educational Improvement. In: **Radical Solutions and Learning Analytics**. 1. ed. [S.l.]: Springer, 2020. p. 1–15. Citado nas páginas 24 e 36.

GARCÍA-PEÑALVO, F. J.; COLOMO-PALACIOS, R.; LYTRAS, M. D. Informal learning in work environments: training with the social web in the workplace. **Behaviour & Information Technology**, Taylor & Francis, v. 31, n. 8, p. 753–755, 2012. Citado nas páginas 55 e 73.

GARRISON, D. R. Online collaboration principles. **Journal of Asynchronous Learning Networks**, v. 10, n. 1, p. 25–34, 2006. Citado na página 30.

____. Online community of inquiry review: Social, cognitive, and teaching presence issues. **Journal of Asynchronous Learning Networks**, ERIC, v. 11, n. 1, p. 61–72, 2007. Citado na página 33.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. Critical inquiry in a text-based environment: Computer conferencing in higher education. **The internet and higher education**, Elsevier, v. 2, n. 2-3, p. 87–105, 1999. Citado nas páginas 48, 51, 52 e 68.

____. The first decade of the community of inquiry framework: A retrospective. **The internet and higher education**, Elsevier, v. 13, n. 1-2, p. 5–9, 2010. Citado nas páginas 49 e 51.

GAŠEVIĆ, D.; DAWSON, S.; ROGERS, T.; GASEVIC, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. **The Internet and Higher Education**, Elsevier, v. 28, p. 68–84, 2016. Citado na página 46.

GAŠEVIĆ, D.; JOKSIMOVIĆ, S.; EAGAN, B. R.; SHAFFER, D. W. Sens: Network analytics to combine social and cognitive perspectives of collaborative learning. **Computers in Human Behavior**, Elsevier, v. 92, p. 562–577, 2019. Citado nas páginas 39, 41, 43, 46 e 63.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. [S.l.]: O'Reilly Media, 2019. Citado na página 112.

GILBERT, S. Portraits of participation: Exploring the relationship between social motivators and facets of participation in a twitter-based community. In: **Proceedings of the 50th Hawaii International Conference on System Sciences**. [S.l.: s.n.], 2017. Citado na página 34.

GÓMEZ-AGUILAR, D. A.; HERNÁNDEZ-GARCÍA, Á.; GARCÍA-PEÑALVO, F. J.; THERÓN, R. Tap into visual analysis of customization of grouping of activities in elearning. **Computers in Human Behavior**, Elsevier, v. 47, p. 60–67, 2015. Citado na página 36.

GOODREAU, S. M.; KITTS, J. A.; MORRIS, M. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. **Demography**, Springer, v. 46, n. 1, p. 103–125, 2009. Citado na página 45.

GREENHOW, C.; GIBBINS, T.; MENZER, M. M. Re-thinking scientific literacy out-of-school: Arguing science issues in a niche Facebook application. **Computers in Human Behavior**, Elsevier Ltd, v. 53, p. 593–604, 2015. Citado nas páginas 25, 55, 57, 62 e 73.

GRUZD, A.; KUMAR, P.; ABUL-FOTTOUH, D.; HAYTHORNTHWAITE, C. Coding and classifying knowledge exchange on social media: A comparative analysis of the# twitterstorians and askhistorians communities. **Computer Supported Cooperative Work (CSCW)**, Springer, v. 29, n. 6, p. 629–656, 2020. Citado nas páginas 25, 31 e 47.

GRUZD, A.; PAULIN, D.; HAYTHORNTHWAITE, C. Analyzing Social Media and Learning Through Content and Social Network Analysis: A Faceted Methodological Approach. **Journal of Learning Analytics**, v. 3, n. 3, p. 46–71, 2016. Citado nas páginas 25, 31, 33, 41, 43, 63, 73 e 103.

HAFEEZ, K.; ALGHATAS, F. M.; FOROUDI, P.; NGUYEN, B.; GUPTA, S. Knowledge sharing by entrepreneurs in a virtual community of practice (VCoP). **Information Technology & People**, v. 32, n. 2, p. 405–429, 2019. Citado nas páginas 25, 42 e 44.

HALBERSTAM, Y.; KNIGHT, B. Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. **Journal of public economics**, Elsevier, v. 143, p. 73–88, 2016. Citado na página 126.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. **Journal of the royal statistical society. series c (applied statistics)**, JSTOR, v. 28, n. 1, p. 100–108, 1979. Citado na página 82.

HAYTHORNTHWAITE, C. Learning , connectivity and networks. **Information and Learning Science**, 2018. Citado nas páginas 29, 30, 33, 34, 39, 56 e 96.

HAYTHORNTHWAITE, C.; KUMAR, P.; GRUZD, A.; GILBERT, S.; VALLE, M. Esteve del; PAULIN, D. Learning in the wild: coding for learning and practice on reddit. **Learning, media and technology**, Taylor & Francis, v. 43, n. 3, p. 219–235, 2018. Citado nas páginas 30, 34 e 98.

HAYTHORNTHWAITE, C.; LAAT, M. D. Social networks and learning networks: Using social network perspectives to understand social learning. In: LANCASTER UNIVERSITY AALBORG, DENMARK. **Proceedings of the 7th international conference on networked learning**. [S.l.], 2010. p. 183–190. Citado na página 115.

HAYTHORNTHWAITE, C.; LAAT, M. de; SCHREURS, B. A Social Network Analytic Perspective on E-Learning. In: **The SAGE Handbook of E-learning Research**. London: SAGE Publications, 2016. p. 251–269. Citado nas páginas 24 e 43.

HELMS, R.; IGNACIO, R.; BRINKKEMPER, S.; ZONNEVELD, A. Limitations of network analysis for studying efficiency and effectiveness of knowledge sharing. **Electronic Journal of Knowledge Management**, Academic Conferences International Limited, v. 8, n. 1, p. 53, 2010. Citado na página 43.

HENRI, F.; PUDELKO, B. Understanding and analysing activity and learning in virtual communities. **Journal of Computer Assisted Learning**, Wiley Online Library, v. 19, n. 4, p. 474–487, 2003. Citado na página 33.

HERNÁNDEZ-GARCÍA, Á.; GONZÁLEZ-GONZÁLEZ, I.; JIMÉNEZ-ZARCO, A. I.; CHAPARRO-PELÁEZ, J. Applying social learning analytics to message boards in online distance learning: A case study. **Computers in Human Behavior**, Elsevier, v. 47, p. 68–80, 2015. Citado nas páginas 38, 41, 43 e 62.

HOD, Y.; BIELACZYC, K.; BEN-ZVI, D. Revisiting learning communities: Innovations in theory and practice. **Instructional Science**, Springer, v. 46, n. 4, p. 489–506, 2018. Citado na página 33.

HOLLWECK, T. Case study research design and methods, robert k. yin. **The Canadian Journal of Program Evaluation**, v. 30, n. 1, p. 108, 2015. Citado na página 109.

HOSKINS, A. The end of collective memory. **Digital memory studies: Media pasts in transition**, Routledge, 2017. Citado na página 33.

HUDGINS, W.; LYNCH, M.; SCHMAL, A.; SIKKA, H.; SWENSON, M.; JOYNER, D. A. Informal Learning Communities: The Other Massive Open Online 'C'. **L@S 2020 - Proceedings of the 7th ACM Conference on Learning @ Scale**, p. 91–101, 2020. Citado nas páginas 24, 25, 30, 33, 56, 57, 73, 90 e 92.

JAN, S. K. Investigating virtual communities of practice with social network analysis: guidelines from a systematic review of research. **International Journal of Web Based Communities**, Inderscience Publishers (IEL), v. 15, n. 1, p. 25–43, 2019. Citado nas páginas 41, 43, 65, 66 e 116.

JAN, S. K.; VLACHOPOULOS, P. Influence of learning design of the formation of online communities of learning. **International Review of Research in Open and Distributed Learning**, Athabasca University, v. 19, n. 4, 2018. Citado na página 43.

____. Social network analysis: A framework for identifying communities in higher education online learning. **Technology, Knowledge and Learning**, Springer, v. 24, n. 4, p. 621–639, 2019. Citado na página 41.

JAN, S. K.; VLACHOPOULOS, P.; PARSELL, M. Social network analysis and learning communities in higher education online learning: A systematic literature review. **Online Learning**, v. 23, n. 1, 2019. Citado nas páginas 40, 62 e 65.

JENKINS, H. **Confronting the challenges of participatory culture: Media education for the 21st century**. [S.l.]: The MIT Press, 2009. Citado nas páginas 33 e 96.

JOKSIMOVIC, S.; GASEVIC, D.; KOVANOVIC, V.; ADESOPE, O.; HATALA, M. Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions. **The internet and higher education**, Elsevier, v. 22, p. 1–10, 2014. Citado na página 47.

JOKSIMOVIĆ, S.; GAŠEVIĆ, D.; KOVANOVIĆ, V.; RIECKE, B. E.; HATALA, M. Social presence in online discussions as a process predictor of academic performance. **Journal of Computer Assisted Learning**, Wiley Online Library, v. 31, n. 6, p. 638–654, 2015. Citado nas páginas 49, 51, 52 e 74.

JOKSIMOVIC, S.; JOVANOVIC, J.; KOVANOVIC, V.; GASEVIC, D.; MILIKIC, N.; ZOUAQ, A.; STAALDUINEN, J. P. van. Comprehensive Analysis of Discussion Forum Participation: From Speech Acts to Discussion Dynamics and Course Outcomes. **IEEE Transactions on Learning Technologies**, v. 13, n. 1, p. 38–51, 2019. ISSN 1939-1382. Citado nas páginas 25, 39, 46 e 63.

JOKSIMOVIC, S.; KOVANOVIC, V.; DAWSON, S. The Journey of Learning Analytics. **Review of Higher Education**, v. 6, n. 4, p. 37–63, 2019. Citado na página 38.

____. The Journey of Learning Analytics. **Review of Higher Education**, v. 6, n. 4, p. 37–63, 2019. Citado na página 48.

JOKSIMOVIC, S.; KOVANOVIC, V.; DOWELL, N.; GASEVIC, D.; GRAESSER, A. C.; SKRYPNYK, O.; DAWSON, S. How do you connect? Analysis of social capital accumulation in connectivist MOOCs. In: **ACM International Conference Proceeding Series**. [S.l.: s.n.], 2015. v. 16-20-Marc, p. 64–68. Citado nas páginas 41, 46 e 47.

JOKSIMOVIĆ, S.; MANATAKI, A.; GAŠEVIĆ, D.; DAWSON, S.; KOVANOVIĆ, V.; KEREKI, I. F. D. Translating network position into performance: importance of centrality in different network configurations. In: **Proceedings of the sixth international conference on learning analytics & knowledge**. [S.l.: s.n.], 2016. p. 314–323. Citado na página 45.

KALIISA, R.; MØRCH, A. I.; KLUGE, A. Exploring social learning analytics to support teaching and learning decisions in online learning environments. In: SPRINGER. **European conference on technology enhanced learning**. [S.l.], 2019. p. 187–198. Citado na página 47.

KAROLY, L. A.; PANIS, C. **The 21st century at work: Forces shaping the future workforce and workplace in the United States**. 1. ed. Pittsburgh: RAND Corporation, 2004. Citado na página 23.

KASSAMBARA, A. **Practical guide to cluster analysis in R: Unsupervised machine learning**. [S.l.]: Sthda, 2017. v. 1. Citado nas páginas 81 e 82.

KEIM, D.; ANDRIENKO, G.; FEKETE, J.-D.; GÖRG, C.; KOHLHAMMER, J.; MELANÇON, G. Visual analytics: Definition, process, and challenges. In: **Information visualization**. [S.l.]: Springer, 2008. p. 154–175. Citado nas páginas 36, 37 e 67.

KEIM, D.; KOHLHAMMER, J.; ELLIS, G.; MANSMANN, F. Mastering the information age: solving problems with visual analytics. Goslar: Eurographics Association, 2010. Citado na página 36.

KENT, C.; LASLO, E.; RAFAELI, S. Interactivity in online discussions and learning outcomes. **Computers and Education**, Elsevier Ltd, v. 97, p. 116–128, 2016. Citado nas páginas 30 e 63.

KENT, C.; RECHAVI, A. Deconstructing online social learning: network analysis of the creation, consumption and organization types of interactions. **International Journal of Research & Method in Education**, Taylor & Francis, v. 43, n. 1, p. 16–37, 2020. Citado nas páginas 29, 30, 33, 38 e 43.

KENT, C.; RECHAVI, A.; RAFAELI, S. Networked learning analytics: A theoretically informed methodology for analytics of collaborative learning. In: **Learning In a Networked Society**. [S.l.]: Springer, 2019. p. 145–175. Citado nas páginas 38, 41 e 43.

KIM, D.; YOON, M.; JO, I. H.; BRANCH, R. M. Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea. **Computers and Education**, v. 127, n. September 2016, p. 233–251, 2018. Citado na página 63.

KITCHENHAM, B. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004. Citado na página 57.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Citeseer, 2007. Citado na página 57.

KLAMMA, R. Community Learning Analytics – Challenges and Opportunities. In: **International Conference on Web-Based Learning**. [S.l.: s.n.], 2013. p. 284–293. Citado nas páginas 25, 31, 33 e 56.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. **Journal of the ACM**, v. 46, n. 5, p. 604–632, 1999. Citado na página 40.

KOVANOVIĆ, V.; JOKSIMOVIĆ, S.; POQUET, O.; HENNIS, T.; ČUKIĆ, I.; VRIES, P. D.; HATALA, M.; DAWSON, S.; SIEMENS, G.; GAŠEVIĆ, D. Exploring communities of inquiry in massive open online courses. **Computers & Education**, Elsevier, v. 119, p. 44–58, 2018. Citado nas páginas 49 e 51.

KOVANOVIĆ, V.; JOKSIMOVIĆ, S.; WATERS, Z.; GAŠEVIĆ, D.; KITTO, K.; HATALA, M.; SIEMENS, G. Towards automated content analysis of discussion transcripts: A cognitive presence case. In: **Proceedings of the sixth international conference on learning analytics & knowledge**. [S.l.: s.n.], 2016. p. 15–24. Citado nas páginas 42, 44 e 63.

LAAT, M. D.; PRINSEN, F. Social Learning Analytics: Navigating the Changing Settings of Higher Education. **Research & Practice in Assessment**, v. 9, n. 1, p. 51–60, 2014. Citado nas páginas 25, 26, 29, 30, 38, 39, 48 e 57.

LANGER, T.; MEISEN, T. System design to utilize domain expertise for visual exploratory data analysis. **Information**, Multidisciplinary Digital Publishing Institute, v. 12, n. 4, p. 140, 2021. Citado na página 74.

LANTZ, B. **Machine learning with R: expert techniques for predictive modeling**. [S.l.]: Packt publishing ltd, 2019. Citado nas páginas 80 e 83.

LEE, A. V. Y.; TAN, S. C. Temporal analytics with discourse analysis: Tracing ideas and impact on communal discourse. In: **Proceedings of the Seventh International Learning Analytics & Knowledge Conference**. [S.l.: s.n.], 2017. p. 120–127. Citado na página 47.

LEE, B. Social media as a non-formal learning platform. **Procedia-Social and Behavioral Sciences**, Elsevier, v. 103, p. 837–843, 2013. Citado na página 56.

LIN, Y.; YU, R.; DOWELL, N. Liwcs the same, not the same: Gendered linguistic signals of performance and experience in online stem courses. In: SPRINGER. **International Conference on Artificial Intelligence in Education**. [S.l.], 2020. p. 333–345. Citado nas páginas 47 e 71.

LUKE, D. A. **A user's guide to network analysis in R**. [S.l.]: Springer, 2015. Citado nas páginas 45, 87 e 129.

MA, X.; HUMMER, D.; GOLDEN, J. J.; FOX, P. A.; HAZEN, R. M.; MORRISON, S. M.; DOWNS, R. T.; MADHIKARMI, B. L.; WANG, C.; MEYER, M. B. Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. **ISPRS International Journal of Geo-Information**, Multidisciplinary Digital Publishing Institute, v. 6, n. 11, p. 368, 2017. Citado na página 74.

MACIÀ, M.; GARCÍA, I. Informal online communities and networks as a source of teacher professional development: A review. **Teaching and teacher education**, Elsevier, v. 55, p. 291–307, 2016. Citado na página 33.

MALCOLM, J.; HODKINSON, P.; COLLEY, H. The interrelationships between informal and formal learning. **Journal of workplace learning**, MCB UP Ltd, 2003. Citado na página 31.

MALINEN, S. Understanding user participation in online communities: A systematic literature review of empirical studies. **Computers in human behavior**, Elsevier, v. 46, p. 228–238, 2015. Citado na página 33.

MAMAS, C.; JR, P. B.; DALY, A. J.; MOUKARZEL, S. Friendship and support networks among students with disabilities in middle school. **International Journal of Educational Research**, Elsevier, v. 103, p. 101608, 2020. Citado nas páginas 45, 46, 62, 64 e 66.

MARSICK, V. J.; WATKINS, K. E. Informal and incidental learning. **New directions for adult and continuing education**, v. 2001, n. 89, p. 25, 2001. Citado na página 31.

MEIER, A.; SPADA, H.; RUMMEL, N. A rating scheme for assessing the quality of computer-supported collaboration processes. **International Journal of Computer-Supported Collaborative Learning**, Springer, v. 2, n. 1, p. 63–86, 2007. Citado na página 48.

MENSHIKOVA, A. Evaluation of expertise in a virtual community of practice: The case of stack overflow. In: SPRINGER. **International Conference on Digital Transformation and Global Society**. [S.l.], 2018. p. 483–491. Citado nas páginas 41, 42 e 44.

MIEGHEM, P. V. Human psychology of common appraisal: The reddit score. **IEEE Transactions on Multimedia**, IEEE, v. 13, n. 6, p. 1404–1406, 2011. Citado na página 57.

MISLOVE, A.; MARCON, M.; GUMMADI, K. P.; DRUSCHEL, P.; BHATTACHARJEE, B. Measurement and Analysis of Online Social Networks Systems. In: **Proceedings of the 7th ACM SIGCOMM conference on Internet measurement**. [S.l.: s.n.], 2007. p. 1297–1300. Citado na página 43.

NACU, D.; UPADHYAY, P.; SKOREPA, E.; EVERETTE, T.; FLORES, E.; JACKSON, M.; PINKARD, N. Implementing learning analytics to foster a stem learning ecosystem at the city-level: Emerging research and design challenges. **Proceedings of the 6th 2019 ACM Conference on Learning at Scale, L@S 2019**, 2019. Citado nas páginas 24 e 35.

NISTOR, N.; DASCALU, M.; TARNAI, C.; TRAUSAN-MATU, S. Predicting newcomer integration in online learning communities: Automated dialog assessment in blogger communities. **Computers in Human Behavior**, v. 105, n. September 2019, p. 106202, 2020. Citado nas páginas 25, 31, 33, 34, 41 e 64.

NISTOR, N.; DERNTL, M.; KLAMMA, R. Learning Analytics : Trends and Issues of the Empirical Research of the Years 2011 – 2014. **Design for Teaching and Learning in a Networked World**, v. 4, p. 453–459, 2015. Citado nas páginas 25, 26, 35, 48 e 57.

PAGE, L.; BRIN, S.; MOTWANI, R. *et al.* **The PageRank citation ranking: Bringing order to the web (Technical report). 1999. Standord InfoLab**. 2019. Citado na página 40.

PAL, A.; PRAKASH, P. **Practical time series analysis: master time series data processing, visualization, and modeling using python**. [S.l.]: Packt Publishing Ltd, 2017. Citado nas páginas 75, 81, 84 e 85.

PALAZUELOS, C.; GARCÍA-SAIZ, D.; ZORRILLA, M. Social network analysis and data mining: An application to the e-learning context. In: SPRINGER. **International Conference on Computational Collective Intelligence**. [S.l.], 2013. p. 651–660. Citado na página 39.

PEKRUN, R. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. **Educational psychology review**, Springer, v. 18, n. 4, p. 315–341, 2006. Citado na página 125.

PENNEBAKER, J. W.; BOYD, R. L.; JORDAN, K.; BLACKBURN, K. **The development and psychometric properties of LIWC2015**. [S.l.], 2015. Citado na página 47.

PENTREATH, N. **Machine learning with spark**. [S.l.]: Packt Publishing Ltd, 2015. Citado na página 78.

PENUEL, W. R.; LEE, T.; BEVAN, B. **Designing and building infrastructures to support equitable STEM learning across settings**. San Francisco, 2014. Disponível em: <https://www.edc.org/research-practice-rp-collaboratory>. Citado na página 24.

PESARE, E.; ROSELLI, T.; ROSSANO, V. Visualizing student engagement in e-learning environment. In: **Proceedings of 22th International Conference on Distributed Multimedia Systems**. [S.l.: s.n.], 2016. p. 26–33. Citado nas páginas 25 e 26.

PINKARD, N. Freedom of movement: Defining, researching, and designing the components of a healthy learning ecosystem. **Human Development**, v. 62, n. 1-2, p. 40–65, 2019. Citado na página 24.

POQUET, O.; DAWSON, S.; DOWELL, N. How effective is your facilitation? group-level analytics of mooc forums. In: **Proceedings of the Seventh International Learning Analytics & Knowledge Conference**. [S.l.: s.n.], 2017. p. 208–217. Citado nas páginas 48, 56 e 63.

REDDIT. **semi colon not required to terminate if statement?** 2021. <https://www.reddit.com/r/learnprogramming/comments/oal4il>. Acesso em: 10 Nov. 2021. Citado na página 91.

RISSER, H. S.; BOTTOMS, S. "newbies" and "celebrities": Detecting social roles in an online network of teachers via participation patterns. **International Journal of Computer-Supported Collaborative Learning**, Springer, v. 9, n. 4, p. 433–450, 2014. Citado na página 41.

ROBINS, G.; LUSHER, D. Simplified account of an exponential random graph model as a statistical model. **Exponential random graph models for social networks**, p. 29–36, 2013. Citado na página 45.

ROTHERHAM, A. J.; WILLINGHAM, D. T. "21st century skills": Not new, but a worthy challenge. **American Educator**, v. 34, n. 1, p. 17–20, 2010. Citado nas páginas 23 e 56.

ROURKE, L.; ANDERSON, T.; GARRISON, D. R.; ARCHER, W. Assessing social presence in asynchronous text-based computer conferencing. **The Journal of Distance Education/Revue de l'ducation Distance**, Athabasca University Press, v. 14, n. 2, p. 50–71, 1999. Citado nas páginas 52 e 53.

ROVAI, A. P. Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks. **The Internet and Higher Education**, Elsevier, v. 5, n. 4, p. 319–332, 2002. Citado na página 33.

RUNESON, P.; HÖST, M. Guidelines for conducting and reporting case study research in software engineering. **Empirical software engineering**, Springer, v. 14, n. 2, p. 131–164, 2009. Citado nas páginas 110, 111 e 113.

SAQR, M.; FORS, U.; NOURI, J. Using social network analysis to understand online problem-based learning and predict performance. **PloS one**, Public Library of Science San Francisco, CA USA, v. 13, n. 9, p. e0203590, 2018. Citado na página 41.

SCHELTER, S.; BIESSMANN, F.; JANUSCHOWSKI, T.; SALINAS, D.; SEUFERT, S.; SZARVAS, G. **On challenges in machine learning model management**. [S.l.], 2018. Disponível em: <http://web.kaust.edu.sa/Faculty/MarcoCanini/classes/CS290E/F19/papers/challenges.pdf>. Citado na página 75.

SCHREURS, B.; De Laat, M. The Network Awareness Tool: A web 2.0 tool to visualize informal networked learning in organizations. **Computers in Human Behavior**, Elsevier Ltd, v. 37, n. 1, p. 385–394, 2014. Citado nas páginas 24, 31, 57 e 73.

SCHUMACHER, C. Supporting informal workplace learning through analytics. In: **Digital Workplace Learning**. [S.l.]: Springer, 2018. p. 43–61. Citado nas páginas 31 e 32.

SCHWENDIMANN, B. A.; RODRIGUEZ-TRIANA, M. J.; VOZNIUK, A.; PRIETO, L. P.; BOROUJENI, M. S.; HOLZER, A.; GILLET, D.; DILLENBOURG, P. Perceiving learning at a glance: A systematic literature review of learning dashboard research. **IEEE Transactions on Learning Technologies**, IEEE, v. 10, n. 1, p. 30–41, 2016. Citado na página 113.

SHUM, S. B.; FERGUSON, R. Social learning analytics. **Journal of educational technology & society**, JSTOR, v. 15, n. 3, p. 3–26, 2012. Citado nas páginas 30, 31, 37, 38, 39, 47 e 59.

SIEMENS, G.; LONG, P. Penetrating the fog: Analytics in learning and education. **EDUCAUSE review**, ERIC, v. 46, n. 5, p. 30, 2011. Citado na página 35.

SILVA, R. F.; GIMENES, I. M. S.; MALDONADO, J. C. An Approach for Assessing Large Online Communities in Informal Learning Environments. In: **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. [S.l.: s.n.], 2020. p. 642–651. Citado nas páginas 25, 34, 35, 42, 44, 48, 55, 62, 65, 71, 73, 74, 90 e 136.

____. Assessing virtual communities of practice in informal learning environments. In: IEEE. **2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)**. [S.l.], 2020. p. 117–121. Citado nas páginas 29, 34, 42, 43, 44, 55, 57, 62, 70, 73, 116 e 136.

____. The Challenge of Evaluating Virtual Communities of Practice: A Systematic Mapping Study. **Interdisciplinary Journal of Information, Knowledge, and Management**, v. 15, n. 1, p. 39–64, 2020. Citado nas páginas 25, 26, 57, 58, 60 e 136.

____. Analyzing learnersb́ehavior and discourse within large online communities: a Social Learning Analytics Dashboard. In: **XXXII Simpósio Brasileiro de Informática na Educação**. [S.l.: s.n.], 2021. Citado na página 136.

SINGRODIA, V.; MITRA, A.; PAUL, S. A review on web scrapping and its applications. In: IEEE. **2019 International Conference on Computer Communication and Informatics (ICCCI)**. [S.l.], 2019. p. 1–6. Citado na página 77.

SOUSA-VIEIRA, M. E.; LÓPEZ-ARDAO, J. C.; FERNÁNDEZ-VEIGA, M.; RODRÍGUEZ-PÉREZ, M.; LÓPEZ-GARCÍA, C. Mining relationships in learning-oriented social networks. **Computer Applications in Engineering Education**, Wiley Online Library, v. 25, n. 5, p. 769–784, 2017. Citado na página 41.

SPEILY, O. R. B.; REZVANIAN, A.; GHASEMZADEH, A.; SAGHIRI, A. M.; VAHIDIPOUR, S. M. Lurkers versus posters: Investigation of the participation behaviors in online learning communities. In: **Educational Networking**. [S.l.]: Springer, 2020. p. 269–298. Citado nas páginas 31, 33, 55, 57, 64, 73, 123 e 126.

SWAN, K.; ICE, P. **The community of inquiry framework ten years later: Introduction to the special issue**. [S.l.]: Elsevier, 2010. Citado nas páginas 49 e 51.

SWIECKI, Z.; SHAFFER, D. W. ISENS: An integrated approach to combining epistemic and social network analyses. In: **Proceedings of the 10th International Conference on Learning Analytics and Knowledge**. [S.l.: s.n.], 2020. p. 305–313. Citado nas páginas 62 e 65.

TOIKKANEN, T.; LIPPONEN, L. The applicability of social network analysis to the study of networked learning. **Interactive Learning Environments**, v. 19, n. 4, p. 365–379, 2011. Citado nas páginas 43 e 48.

VENKATESH, V.; THONG, J. Y.; XU, X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. **MIS quarterly**, JSTOR, p. 157–178, 2012. Citado na página 59.

VYGOTSKY, L. Mind in society: development of higher psychological processes harward university press. **Cambridge, Mass**, 1978. Citado na página 29.

WANG, G. A.; LIU, X.; WANG, J.; ZHANG, M.; FAN, W. Examining micro-level knowledge sharing discussions in online communities. **Information Systems Frontiers**, Springer, v. 17, n. 6, p. 1227–1238, 2015. Citado nas páginas 42, 44, 57 e 62.

WANG, P.; ROBINS, G.; PATTISON, P.; LAZEGA, E. Exponential random graph models for multilevel networks. **Social Networks**, Elsevier, v. 35, n. 1, p. 96–115, 2013. Citado na página 45.

WENGER, E. **Communities of practice: Learning, meaning, and identity**. [S.l.]: Cambridge university press, 1999. Citado na página 33.

WENGER, E.; MCDERMOTT, R. A.; SNYDER, W. **Cultivating communities of practice: A guide to managing knowledge**. [S.l.]: Harvard business press, 2002. Citado na página 107.

WENGER, E.; TRAYNER, B.; De Laat, M. **Promoting and assessing value creation in communities and networks: a conceptual framework**. [S.l.], 2011. v. 18, n. August, 60 p. Citado nas páginas 30, 33, 44, 48, 49, 52, 59 e 68.

WENINGER, T. An exploration of submissions and discussions in social news: Mining collective intelligence of reddit. **Social Network Analysis and Mining**, Springer, v. 4, n. 1, p. 173, 2014. Citado na página 56.

WISE, A.; ZHAO, Y.; HAUSKNECHT, S. Learning Analytics for Online Discussions: Embedded and Extracted Approaches. **Journal of Learning Analytics**, v. 1, n. 2, p. 48–71, 2014. Citado nas páginas 30 e 62.

WISE, A. F. Learning Analytics: Using Data-Informed Decision-Making to Improve Teaching and Learning. **Contemporary Technologies in Education**, p. 119–143, 2019. Citado na página 24.

WOHLIN, C.; RUNESON, P.; HÖST, M.; OHLSSON, M. C.; REGNELL, B.; WESSLÉN, A. **Experimentation in software engineering**. [S.l.]: Springer Science & Business Media, 2012. Citado nas páginas 109, 110, 112 e 113.

WORSLEY, M.; OCHOA, X. Towards collaboration literacy development through multimodal learning analytics. In: **Proceedings 10th International Conference on Learning Analytics and Knowledge (LAK20)**. [S.l.: s.n.], 2020. v. 2610, p. 53–63. Citado nas páginas 24 e 25.

XIONG, J.; FENG, X.; TANG, Z. Understanding user-to-user interaction on government microblogs: An exponential random graph model with the homophily and emotional effect. **Information Processing & Management**, Elsevier, v. 57, n. 4, p. 102229, 2020. Citado na página 134.

ZHANG, J.; ACKERMAN, M. S.; ADAMIC, L. Expertise Networks in Online Communities: Structure and Algorithms. In: **International World Wide Web Conference**. [S.l.: s.n.], 2007. p. 221–230. Citado na página 41.

ZHU, M.; BERGNER, Y.; ZHANG, Y.; BAKER, R.; WANG, Y.; PAQUETTE, L. Longitudinal engagement, performance, and social connectivity: a mooc case study using exponential random graph models. In: **Proceedings of the Sixth International Conference on Learning Analytics & Knowledge**. [S.l.: s.n.], 2016. p. 223–230. Citado nas páginas 45 e 66.

ZHU, M.; HERRING, S. C.; BONK, C. J. Exploring presence in online learning through three forms of computer-mediated discourse analysis. **Distance Education**, Taylor & Francis, v. 40, n. 2, p. 205–225, 2019. Citado nas páginas 63 e 71.

ZINOVIEV, D. **Complex network analysis in Python: Recognize-construct-visualize-analyze-interpret**. [S.l.]: Pragmatic Bookshelf, 2018. Citado na página 40.

ZOU, W.; HU, X.; PAN, Z.; LI, C.; CAI, Y.; LIU, M. 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**, Elsevier, v. 115, 2021. Citado nas páginas 125 e 134.

ICMC USP
SÃO CARLOS