Capítulo 5

Análise de Discussões em Fóruns Educacionais Usando Mineração de Texto e Análise de Grafos

Vitor Rolim, Rafael Ferreira Mello e Rafael Dueire Lins

Abstract

The online learning expansion and the gradual implementation of blended learning in face-to-face courses are responsible for a democratization revolution in education. Those learning modalities require the provision of techniques, tools and theoretical frameworks that make the educational experience of students as similar as possible to presential learning. Thus, this chapter will address the Model of the Community of Inquiry, a framework widely used to analyze interactions in virtual learning environments, and how such a model is related to the educational online discussion analysis. This chapter presents the technique called Epistemic Network analysis (ENA), that may be used to analyze textual data, which is processed using text mining techniques. Such concepts are essential to assist the instructor in the challenge of providing a relevant educational experience to the student and to optimize the construction of knowledge.

Resumo

A expansão do ensino à distância e implementação gradual do ensino híbrido (blended learning) nos cursos presenciais, são responsáveis cada vez mais pela democratização do ensino superior, por ter como princípio fundamental que o ensino possa ocorrer "em qualquer lugar e a qualquer momento". Nessas modalidades de ensino existe a necessidade da provisão de técnicas, ferramentas e instrumentos teóricos que atuem de forma a possibilitar que o estudante tenha um experiência educacional similar a do ensino presencial. Assim, este capítulo abordará o conceito do modelo de Comunidade de Investigação, largamente utilizado para análise de interações em ambientes virtuais de aprendizagem, e como esse conceito pode estar relacionado à análise de discussões em fóruns educacionais. Será apresentada a técnica chamada ENA (Epistemic Network Analysis), que pode ser utilizada para análise de dados textuais, e o processamento prévio desses dados pode ser realizado pelas técnicas de mineração de texto. Esses conceitos serão apresentados ao longo do capítulo com o intuito de abranger o que envolve o desafio de prover

uma experiência educacional satisfatória ao aluno de forma a otimizar a construção do conhecimento.

5.1. Fóruns Educacionais

Recentemente o mundo tem experienciado uma grande difusão de cursos a distância focados principalmente em educação superior e MOOCs (do inglês *Massive Open Online Course*). Apesar de existir inicialmente uma grande adesão de alunos a esses cursos, também existe grande evasão ao longo do tempo (Rivard, 2013). Um dos problemas que explicam esse comportamento é a falta de relacionamento direto entre os alunos, fato que no ensino presencial é constante (Wise et al., 2014). Diante desse contexto, interações em fóruns educacionais são fundamentais para facilitar a interação social em cursos totalmente online (T. Anderson & Dron, 2010). O fórum de discussão é uma ferramenta de comunicação assíncrona online, tendo como função principal promover a interação entre os participantes acerca de diferentes temas. Essa ferramenta desempenha um papel essencial na experiência educacional dos alunos, incentivando-os a aumentar sua participação no curso, respondendo a perguntas, compartilhando recursos e resolvendo problemas (Hew & Cheung, 2008; Ferreira-Mello et al., 2019). Quando utilizado no meio educacional, fornece aos alunos um canal de comunicação, onde alunos e professores interagem para expressar suas dúvidas, opiniões e respostas aos questionamentos existentes sobre algum assunto. Comumente, o fórum educacional possui professores ou tutores como mediadores das interações para que se possa auxiliar os alunos no processo de aprendizagem (Batista & Gobara, 2007; M. A. D. Ferreira et al., 2018; Rolim et al., 2017). A referência (Freitas & Auxiliadora, 2009) apresenta vários contextos onde o fóruns podem ser utilizados, dentre eles:

- Incentivar a criação de laços entres os alunos a partir da discussão de temas específicos da disciplina;
- Desenvolver a capacidade de debate crítico acerca de algum tema ou assunto;
- Dar uma resposta a dúvidas e comentários;
- Guiar os estudos dos alunos baseados nas suas postagens;
- Avaliar o aluno.

Existem vários trabalhos na literatura que demonstram o ganho pedagógico advindo da utilização dos fóruns educacionais, mesmo que os alunos envolvidos no processo empreguem pouco tempo para essa atividade, por proporcionar desenvolvimento do pensamento crítico, criatividade e argumentação e promoção da (co-)construção de conhecimento em um grupo de alunos (Dawson et al., 2011; Cheng et al., 2011; Barbosa et al., 2020). Apesar de todos os benefícios listados, o aumento da interação entre os usuários do fórum faz crescer a quantidade de dados gerados pelos fóruns, o que dificulta o acompanhamento. Por isso, ao longo dos anos vários trabalhos que utilizam técnicas de mineração de texto para extrair informações específicas das discussões em fóruns educacionais foram propostos.

Alguns exemplos de aplicações de mineração de texto à fóruns educacionais são: identificação de dúvidas (Rolim et al., 2016a,b), verificação de plágio (Cavalcanti & Ferreira, 2018), monitoramento de colaboração (Dionísio et al., 2017; M. Ferreira et al., 2020), pontuação de atividade automaticamente (Wanas et al., 2008; Rolim et al., 2017) e análise de sentimentos (Azevedo et al., 2017; Hew et al., 2020). Contudo, nem sempre esses métodos estão diretamente alinhados a teorias educacionais existentes. Essas teorias indicam que os aspectos sociais e cognitivos das mensagens dos alunos são os mais importantes de serem analisados em uma discussão online.

Quando se analisa mensagem num ambiente de fórum, o mais comum é olhar para o lado social. Esse aspecto é extremamente importante, pois sem a participação dos alunos no fórum não se consegue avaliar nenhum outro aspecto. Dentro desse contexto, os modelos educacionais mais utilizados são:

- O Modelo de Murphy (Murphy, 2004) apresenta indicadores para reconhecimento de colaboração em discussões assíncronas. As principais características que esse modelo tenta identificar na participação dos alunos são: (1) Reconhecimento de presença social, (2) Articulação de perspectivas individuais, (3) Acolhimento ou reflexão das perspectivas dos outros, (4) Co-construção de perspectivas e significados partilhados, (5) Construção de objetivos e propósitos partilhados e (6) Produção de artefatos partilhados. Todas essas características estão interligadas para direcionar a avaliação final do modelo.
- O Modelo de Comunidade de Investigação (Garrison et al., 1999) que utiliza a Presença Social. Este modelo categoriza as interações em discussões online em três categorias (*Afetiva*, *Interativa* e *Coesiva*), que por sua vez possuem vários indicadores. Mais detalhes sobre esse modelo serão apresentados na próxima seção.

Por outro lado, o aspecto cognitivo é extremamente importante no contexto educacional. Por isso, existem várias teorias educacionais que exploram a cognição dos alunos a partir das interações em discussões online. Como o objetivo deste capítulo não é aprofundar essas teorias, vamos listar aqui as mais utilizadas na literatura:

- A Taxonomia de Bloom (L. W. Anderson & Sosniak, 1994) avalia o nível de processamento cognitivo neste contexto. As categorias da Taxonomia de Bloom são: conhecimento, compreensão, aplicação, análise, síntese e avaliação. Os níveis dessa taxonomia foram projetados para serem uma representação contínua e não categórica. Contudo, existem trabalhos que agrupam indicadores para conseguir discretizar esses valores.
- A Taxonomia SOLO (Biggs & Collis, 2014) avalia a complexidade estrutural refletida na escrita e diferencia o conteúdo da discussão entre o processamento profundo e de superficial. As categorias que compõem a taxonomia, que são acompanhadas por definições e indicadores, são: pré-estrutural, não-estrutural, multi-estrutural, relacional e abstrato estendido. Os usuários da taxonomia tomaram os níveis relacionais e abstratos estendidos para refletir em profundidade, em oposição ao processamento superficial.

- O Modelo de Comunidade de Investigação (Garrison et al., 1999) utiliza a Presença Cognitiva para avaliar o nível de cognição do aluno. Essa presença é dividida entre: evento desencadeador, exploração, integração e resolução. Mais detalhes sobre esse modelo serão apresentados na próxima seção.

Este trabalho foca no modelo de Comunidade de Investigação que integra não só os aspectos sociais e cognitivos, mas também uma dimensão relacionada à metodologia utilizada pelo professor durante a discussão. A seguir são apresentados os principais conceitos desse modelo.

5.1.1. Comunidade de Investigação (COI)

Esta seção detalha o modelo de Comunidade de Investigação (*Community of Inquiry*) que dentre os diversos arcabouços teóricos existentes para modelagem educacional é um dos que mais se destaca atualmente (Garrison et al., 1999).

Lipman descreve a comunidade de investigação como algo necessário para uma boa experiência educacional e para que o aluno possa produzir resultados decorrentes de um aprendizado profundo (Lipman, 1991). Ele descreve as características de uma comunidade de investigação: questionar, raciocinar, conectar, deliberar, desafiar, e desenvolver técnicas de solução de problemas. Ramsden argumenta que a oportunidade de negociar significados, diagnosticar equívocos e desafiar crenças aceitas, são elementos essenciais para um aprendizado profundo e uma boa experiência educacional (Ramsden, 1988). Sob essa base teórica, o conceito atual do modelo CoI foi desenvolvido. Esse modelo é composto por três presenças (ou dimensões) e o modelo CoI assume que o aprendizado ocorre através da interação dessas três dimensões. A presença cognitiva (*Cognitive Presence*) (Garrison et al., 1999) é parte central do pensamento crítico e descreve o processo pelo qual o estudante pode produzir resultados de forma que possa atingir as metas estabelecidas. A presença social (*Social Presence*), conforme Rourke et al. (1999), descreve a importância de humanizar as relações entre os participantes do curso, de forma que possam se apresentar como “pessoas reais". A presença de ensino (*Teaching Presence*), segundo T. Anderson et al. (2001), descreve o papel dos instrutores antes e durante o curso. A composição do modelo CoI pode ser vista na Figura 5.1.

Cada presença descrita no CoI possui indicadores, esses indicadores representam características presentes no discurso do estudante (Ex.: palavras-chave, frases), e para uma melhor aplicação esses indicadores são agrupados em categorias. Os conceitos dessas presenças, juntamente com suas respectivas categorias são explorados nas subseções desta seção.

Esse dispositivo teórico fornece uma modelagem precisa de como a construção do conhecimento é desenvolvida por um determinado grupo de indivíduos, esse tipo de informação, sobretudo, é de grande importância para a educação a distância, principalmente pela limitação do acompanhamento dos alunos que existe nesse modelo educacional. A apresentação desse dispositivo à comunidade acadêmica da área de informática na educação é fundamental, para que mais trabalhos usando esse dispositivo possam ser desenvolvidos.

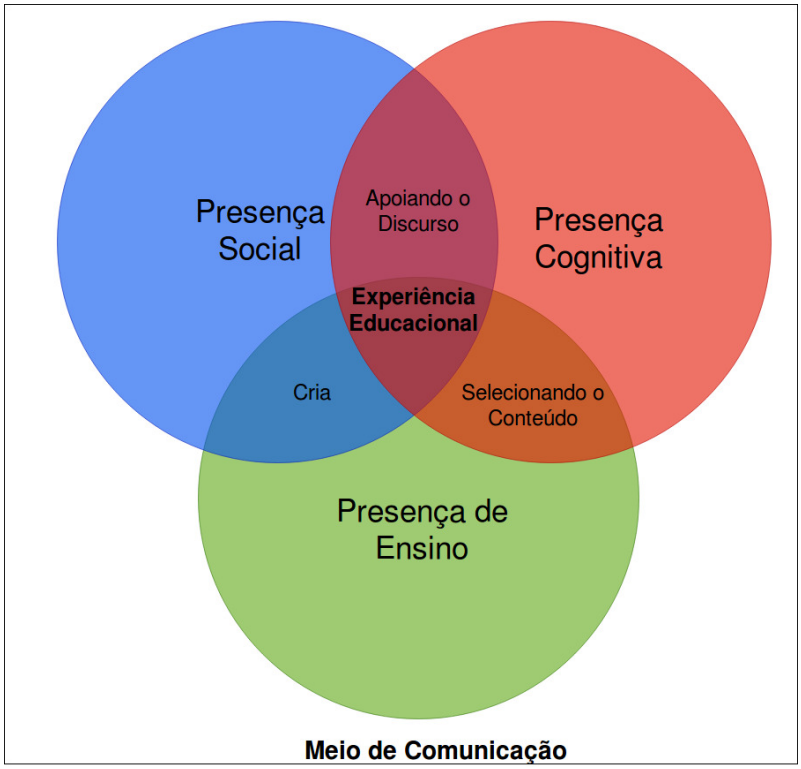


Figura 5.1: **Composição do modelo CoI**

5.1.2. Presença Cognitiva

A presença cognitiva é definida como a exploração, construção, resolução e confirmação da compreensão através da colaboração e reflexão em uma comunidade de investigação. A presença cognitiva é operacionalizada através da prática de investigação (*Practical Inquiry* - PI) (Garrison et al., 2010), que por sua vez é fundamentada no modelo delineado por Dewey (1897), que afirma que uma boa experiência educacional deve ser baseada em um processo de investigação reflexiva.

O processo do PI se desenvolve a partir de um evento inicial que é seguido pelas etapas de percepção, deliberação, concepção e ação. O sucesso desse modelo não está restrito ao processo de reflexão, envolve também o relacionamento interpessoal e o compartilhamento do conhecimento adquirido. Além disso, deve haver sinergia entre os participantes, e pensamentos e ações propositivas se tornam parte essencial desse processo (Garrison et al., 1999).

O modelo PI é divido em quatro categorias, que seguem uma sequência lógica, são elas:

- **Evento desencadeador** (*Triggering event*) - esta categoria é autoexplicativa, pois se trata de um evento inicial, que se dá pela identificação de um problema para em seguida ser realizada a investigação;
- **Exploração** (*Exploration*) - nesta categoria os envolvidos são motivados a buscar por informações, explorar o problema de forma que as alternativas possam ser dis-

cutidas de forma reflexiva e se obtenha entendimento do problema inicial;

- **Integração** (*Integration*) - os estudantes constroem um significado do conhecimento obtido na fase de exploração, integrando o conhecimento à concepção de ideias coerente;
- **Resolução** (*Resolution*) - também autoexplicativa, esta categoria aborda a resolução do problema inicial, aplicando as hipótese e ideias desenvolvidas, e o sucesso dessa aplicação irá definir a continuidade desse processo.

5.1.3. Presença Social

Garrison & Arbaugh (2007) definem a presença social como uma capacidade de autoprojetar e estabelecer relacionamentos pessoais e propositivos. Em contraste com a interação face-a-face, em uma discussão *on-line*, é essencial expressar certas habilidades para estabelecer uma comunicação sócio-emocional de forma textual.

Como mencionado anteriormente, a presença social é responsável por humanizar os relacionamentos nas discussões *on-line* trazendo-as para o âmbito pessoal, tornando-se assim parte fundamental do modelo CoI. A presença social está além de apenas criar relacionamentos pessoais, ela deve promover a coesão do grupo através do estabelecimento de uma comunicação aberta e propositiva conforme Rourke et al. (1999).

A presença social possui alguns indicadores que são agrupados em três categorias: **Afetiva**, **Interativa** e **Coesiva**. A categoria **Afetiva** está associada a emoções, sentimentos e expressões de humor. Esta categoria visa examinar a tradução de emoções reais em texto. A categoria **Interativa** está focada na troca de mensagens, se propõe a implementar uma comunicação aberta entre os participantes. Alguns fatores tem um forte valor nessa categoria, como: a interação social, elogios, expressões de apreciação e consciência mútua. A categoria **Coesiva** tenta descobrir o sentido de união e compromisso do grupo e está associada ao aspecto cognitivo da experiência educacional. As mensagens normalmente citam uma terceira pessoa.

As três categorias da presença social podem ser definidas em termos de os participantes se identificarem com a comunidade, comunicar-se propositivamente em um ambiente de confiança e desenvolver relacionamentos interpessoais. De fato, a interação social deve ser encorajada nas discussões *on-line*.

5.1.4. Presença de Ensino

A presença de ensino é um elemento de ligação na criação de uma comunidade de investigação. O desenvolvimento apropriado das presenças sociais e cognitivas e o estabelecimento de um pensamento crítico em uma comunidade de investigação se deve a presença de um professor, e o sucesso e a falha da experiência estão diretamente ligadas a capacidade de gestão, mediação e liderança desse professor (Garrison et al., 1999).

Diferentemente das outras duas presenças, a presença de ensino começa antes do início do curso, o que envolve todos os preparativos para que o curso seja conduzido de forma a garantir uma boa experiência educacional, e continua durante o curso, com o professor assumindo um papel de facilitador e fornecendo direcionamentos aos alunos (T. Anderson et al., 2001).

A presença de ensino é responsável por balancear as questões sociais e cognitivas de modo que atendam os resultados esperados. Os indicadores desta presença são agrupados em três categorias: **Projeto e Organização**, ***Facilitando o Discurso***, e **Instrução Direta**.

O **Projeto e Organização** está diretamente relacionado ao desenho e planejamento do curso, em um processo análogo ao do curso presencial, também envolve a administração do grupo e das atividades individuais durante a execução do curso. A segunda categoria (**Facilitando o Discurso**, está concentrada na produtividade e na aquisição válida de conhecimento, está preocupada com a integridade acadêmica da comunidade colaborativa de alunos. Esta categoria se sobrepõe com a presença social, onde o professor assume o papel de criar e manter a presença social. Na terceira categoria o professor avalia o discurso do aluno e a eficácia do processo educacional.

5.2. Modelagem de tópicos

Na era virtual em que vivemos, uma grande quantidade de dados é gerada a cada segundo[1]. Esses dados têm um potencial valor associado, a depender da "riqueza" de informações neles escondidas. No entanto, é muito difícil, ou impossível, para o ser humano extrair essas informações para si mesmo sem recursos computacionais, dada a diversidade de fontes, tipos e complexidade dos dados coletados, além do grande volume. Reduzindo a escala do problema ao nível institucional, a demanda por indexação e recuperação de informação, bem como pela interpretabilidade dos dados é crescente devido a grande quantidade de documentos produzidos diariamente.

O entendimento dos dados é um desafio atribuído ao campo da ciência de dados, e ainda tem muitas tarefas em aberto, embora muitas outras tenham sido resolvidas com excelentes resultados. Podemos atribuir uma parte dessas tarefas relacionadas aos dados não estruturados, dados textuais especificamente.

Observando a demanda pela interpretabilidade dos dados, uma subárea da mineração de dados denominada mineração textual cresceu bastante nas últimas duas décadas. Nesse contexto, são utilizadas técnicas adequadas para o tratamento de dados não estruturados (i.e., textos), normalmente baseadas em processamento de linguagem natural (PLN), como por exemplo, a modelagem de tópicos (*Topic modeling*).

A modelagem de tópicos (MT) faz uso de técnicas baseadas em distribuições probabilísticas para descobrir os tópicos abordados em uma grande coleção de documentos (D. M. Blei, 2012). Diferentemente dos mecanismos de buscas baseados em palavras-chave, a MT permite que documentos sejam agrupados de acordo com o tema, e as relações entre diversos temas de um mesmo documento. Portanto, a MT permite organizar e sumarizar documentos em quantidades que seriam humanamente impossível. Para entender de forma prática o funcionamento dessa técnica, observe-se o exemplo abaixo[2]:

> **Exemplo 1:** *"Após ser ingerida, a vitamina C participa de diversas ações bioquímicas vitais para o organismo. Ela melhora o sistema imunológico, a pele, o humor e evita problemas oftalmológicos e derrames. O nutriente tam-*

[1]**Fonte:** https://www.domo.com/learn/data-never-sleeps-5
[2]**Fonte:** https://www.minhavida.com.br/alimentacao/tudo-sobre/17559-vitamina-c

bém conta com forte ação antioxidante, combatendo os radicais livres. Este nutriente pode ser obtido especialmente em algumas frutas, como a laranja, goji berry, acerola, kiwi e goiaba, e verduras, como a couve e o brócolis."

O Exemplo 1 apresenta um texto que aborda os benefícios da vitamina C. Ao utilizarmos a MT nesse caso, poderíamos identificar que esse documento pertence tanto ao tópico relacionado a saúde, por possuir palavras como "*bioquímicas*", "*organismo*", "*imunológico*", quanto ao tópico relacionado a alimentação, por possuir palavras como "*nutriente*", "*frutas*", "*laranja*". O modelo indicaria a composição das distribuições probabilísticas de cada tópico contido no documento levando em consideração a distribuição de cada palavra.

A modelagem de tópicos é uma técnica de mineração de texto bastante consolidada e utilizada do não só no meio acadêmico como também para resolver demandas do mercado das mais diversas áreas. Nesse contexto, a importância de abordar esse tema no curso se dá pela atualidade da técnica e pela utilização crescente em trabalhos recentes. Ademais para a área educacional podemos utilizar essa técnica para auxiliar a entender o que é discutido nos fóruns de discussão.

5.2.1. Evolução da Técnica

Atualmente existem vários algoritmos que podem ser usados para aplicações de MT. Dentre eles, o mais conhecido é o LDA (*Latent Dirichlet Allocation*) (D. M. Blei et al., 2003a). Contudo outros algoritmos o precederam servindo como base para o seu desenvolvimento e outros algoritmos surgiram posteriormente a sua criação motivados por suas limitações, como a ausência de correlações entre os tópicos, a informação prévia necessária sobre quantos tópicos existem na coleção de documentos e a falta de rótulos nos tópicos gerados. Por isso, nesta seção serão apresentados os principais algoritmos de MT junto com uma comparação entre eles. Como a MT não é o objeto de análise principal deste capítulo, mas apenas um meio, não entraremos em detalhes dos algoritmos. Os principais algoritmos de TM são:

Latent Semantic Analysis (LSA) Também chamado de *Latent Semantic Indexing* (LSI), Deerwester et al. (1990) identifica os termos para inferir o tópico latente a partir desses termos. Esta técnica foi pioneira e foi inicialmente idealizada para recuperação de informações. Este método faz uso de uma representação de *bag-of-words* para criar uma matriz contendo documentos e termos como linhas e colunas, respectivamente. Na etapa seguinte, a matriz é decomposta usando SVD (Decomposição em Valores Singulares) nas outras 3 matrizes para cada tópico: (T) contendo os termos; (S) com os valores singulares da matriz diagonal; (D) contendo os documentos.

Latent Dirichlet Allocation (LDA) D. M. Blei et al. (2003b) define LDA como um modelo de corpus probabilístico generativo. Esta técnica teve como base duas outras técnicas: LSA (Deerwester et al., 1990) e *Probabilistic Latent Semantic Analysis* (pLSA) (Hofmann, 1999). Diferentemente da LSA, que usa decomposição em valor singular na matriz termo-documento, o LDA assume uma distribuição de Dirichlet sobre os tópicos

latentes. É importante mencionar que um tópico (no contexto de LDA) é uma distribuição de palavras sobre um vocabulário fixo; cada tópico possui o mesmo vocabulário, porém com distribuição diferente para cada palavra. O algoritmo LDA tem duas etapas principais: **i) Etapa de inicialização**: o algoritmo atribui cada palavra a um tópico temporário usando a distribuição Dirichlet; **ii) Etapa iterativa**: o algoritmo atualiza as atribuições de tópicos (para cada palavra em cada documento) com base em quão comum é aquela palavra específica entre tópicos e quão comuns são tópicos no documento atual. Após a conclusão da etapa iterativa, o algoritmo retorna um modelo treinado que pode ser usado para extrair tópicos de um novo documento.

Hierarchical Dirichlet Process (HDP) é um modelo bayesiano não paramétrico para problemas de agrupamento envolvendo vários grupos de dados (Teh et al., 2005). Esta técnica aborda o problema da necessidade de que o número de tópicos (valor K) seja informado *a priori* no momento de geração dos modelos LDA, usando o processo de Dirichlet. Além disso, os grupos de dados podem compartilhar os tópicos, causando um efeito de dependência.

Correlated Topic Models (CTM) O CTM foi proposto para atender a uma limitação do modelo LDA, e fornece uma correlação entre os tópicos. CTM é um modelo hierárquico em que as proporções dos tópicos exibem correlação através da distribuição normal logística (D. Blei & Lafferty, 2006) em vez da distribuição de Dirichlet e incorpora uma estrutura de covariância entre os tópicos, no entanto, ele usa a abordagem metodológica do LDA (Shanmugam, 2019).

Supervised LDA (SLDA) É um modelo estatístico para documentos rotulados, ou seja, atua como uma extensão do LDA com aprendizagem supervisionada. Assim, o número de tópicos (valor K) é conhecido *a priori*. Um rótulo é anexado a cada documento, para melhor prever os novos documentos não rotulados, usando os tópicos latentes gerados pelo modelo (Mcauliffe & Blei, 2008).

Labeled LDA (LLDA) É considera que esta técnica é uma extensão natural de LDA e Multinomial Naive Bayes, eles também definem este modelo probabilístico como um modelo de tópico que restringe LDA ao definir uma correspondência um-para-um entre os tópicos latentes de LDA e os rótulos (Ramage et al., 2009). Além disso, como no SLDA, o valor K é conhecido *a priori*.

Como mencionado anteriormente, o algoritmo mais comumente utilizado é o LDA, principalmente pelo fato de estar disponível em diversas bibliotecas de processamento de linguagem natural e por usar pouco recurso computacional para realizar a geração dos modelos. Mais vantagens e desvantagens dos algoritmos mencionados podem ser observadas na Tabela 5.1.

Os algoritmos apresentados são a evolução natural do LDA (exceto LSA), eliminando algumas de suas limitações amplamente conhecidas. Embora alguns dos algoritmos ofereçam uma solução para as limitações, a escolha de qual algoritmo utilizar deve

Tabela 5.1: **Vantagens e desvantagens de cada algoritmo de modelagem de tópico**

Algoritmo	Vantagens	Desvantagens
LDA	Gera modelos usando pouco recurso computacional; Disponivel em diversas bibliotecas em diversas linguagens de programação.	Precisa informar o número de tópicos antes de gerar o modelo; Não possui relacionamento hierárquico entre os tópicos; Não é possível gerar o modelo com documentos categorizados; Os tópicos gerados não são rotulados.
LSA	Procura gerar modelos comparáveis aos que um ser humano faria.	Consome mais recurso computacional na geração do modelo; Existem poucas bibliotecas com esse algoritmo implementado; Possui as mesmas limitações do LDA.
HDP	Possui relacionamento hierárquico entre os tópicos; Não é necessário informar o número de tópicos.	A quantidade de tópicos gerados pode não ser o ideal para o problema abordado; Os tópicos gerados não são rotulados.
CTM	Possui relacionamento hierárquico entre os tópicos.	Precisa informar o número de tópicos antes de gerar o modelo; Existem poucas bibliotecas com esse algoritmo implementado.
SLDA	Possibilita gerar o modelo com documentos categorizados; A quantidade de tópicos é conhecida na geração do modelo.	Necessidade de documentos rotulados previamente; Existem poucas bibliotecas com esse algoritmo implementado.
LLDA	Gera rótulos para cada tópico gerado; A quantidade de tópicos é conhecida na geração do modelo.	Necessidade de documentos rotulados previamente; Existem poucas bibliotecas com esse algoritmo implementado.

levar em conta o problema que etá sendo estudado e como os dados a serem utilizados estão estruturados.

5.2.2. Aplicações de Modelagem de Tópico

A técnica de modelagem de tópicos é comumente utilizada na área de processamento de linguagem natural. É possível encontrar facilmente diversos artigos aplicando esta técnica para os mais diferentes tipos de problemas em diferentes áreas do conhecimento, como área de educação (Rolim, Ferreira, et al., 2019), saúde (Paul & Dredze, 2014), negócios (Maskeri et al., 2008), economia (D'Amato et al., 2017) e muitos outros. Essa técnica fornece percepções compreensíveis, mostrando como os termos e até mesmo os tópicos estão relacionados, em uma coleção de documentos, e isso explica a ampla adoção da técnica.

Na área educacional, a modelagem de tópicos pode ser utilizada para visualizar os tópicos que são abordados no discurso do estudante no ambiente virtual, assim como relacionar essa informação com outras que também podem ser extraídas do texto; um exemplo disso é apresentado no trabalho de Rolim e colegas (Rolim, de Mello, et al., 2019) relacionam o tópico da postagem do aluno com a sua categoria (dúvida, neutra, resposta) para identificar pontos fracos e fortes do aluno na disciplina cursada. Outro exemplo da usabilidade da modelagem de tópicos na área de informática na educação é a possibilidade de através dos tópicos das postagens dos alunos verificar a evolução das fases da presença cognitiva do modelo COI (R. Ferreira et al., 2018).

5.2.3. Modelagem de tópicos na prática

Esta seção apresenta exemplos práticos de como criar um modelo LDA para a realização da modelagem de tópicos numa coleção de documentos utilizando a ferramenta gratuita *Topic Modeling Tool*[3] que possui uma interface gráfica para as etapas de treinamento e visualização dos tópicos gerados pelo algoritmo LDA. Essa ferramenta permite, com poucos passos criar uma aplicação de TM. A Figura 5.2 apresenta a tela de configuração do algoritmo de modelagem de tópicos com todos os principais parâmetros necessários.

Os arquivos que serão utilizados para gerar o modelo de tópicos devem ter o diretório indicado no campo "*Input Dir*", cada documento deve ser um arquivo no formato

[3]https://senderle.github.io/topic-modeling-tool/documentation/2017/01/06/quickstart.html

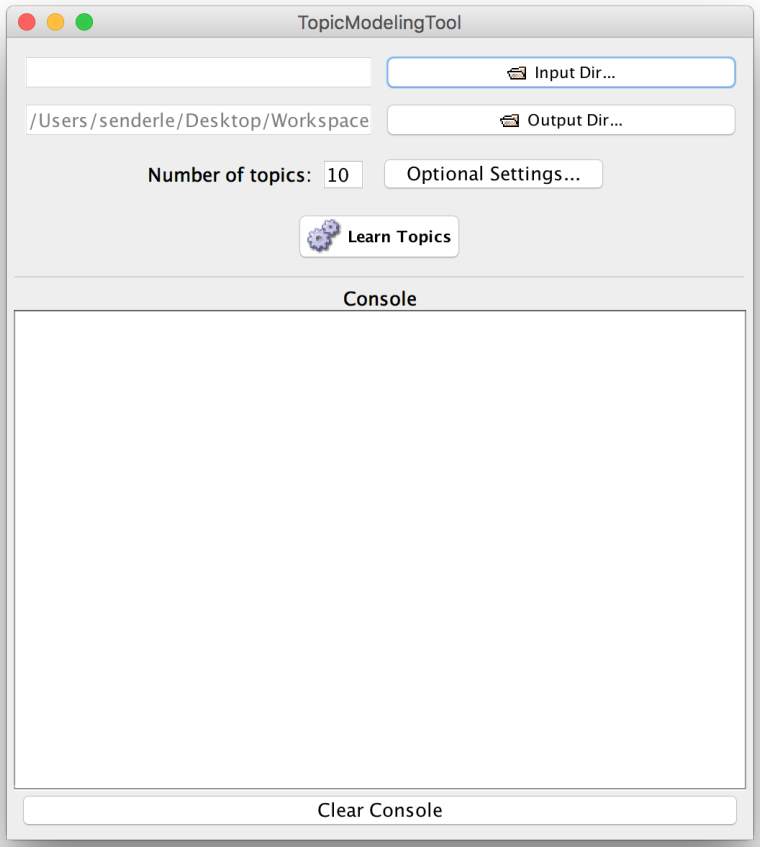


Figura 5.2: **Tela inicial do *Topic Modeling Tool***

“TXT"e o conteúdo deve ser o corpo do texto. No campo de “*Output Dir*"deve ser informado o diretório que os resultados serão salvos. Nas configurações do “*Optional Settings*"os valores padrão já definidos não devem ser alterados. Em seguida na tela inicial (Figura 5.2) o número de tópicos dever ser inserido no campo de “*Number of topics*"e logo após o modelo pode ser gerando clicando em “*Learn Topics*". Os resultados são gerados tanto em arquivos do tipo “CSV"quanto em páginas “HTML"para uma melhor visualização.

5.3. Análise de redes epistêmicas

A análise de redes epistêmicas (ENA, do inglês *Epistemic Network Analysis*) é uma técnica comumente utilizada em combinação com TM para análise de interações em fóruns educacionais. Para entender o funcionamento do ENA, é necessário entender o processo de concepção desta técnica, que se inicia no desenvolvimento do conceito de *epistemic frames* (quadros epistêmicos) como um mecanismo pelo qual os alunos podem usar experiências em jogos eletrônicos e outros ambientes de aprendizagem interativos para ajudá-los a lidar de maneira eficaz com situações fora do contexto original de aprendizagem

(D. W. Shaffer, 2006). De acordo com a hipótese do quadro epistêmico (D. W. Shaffer et al., 2009):

a. um quadro epistêmico conecta as habilidades, conhecimentos, valores, identidade e epistemologia, que um indivíduo assume esse quadro por fazer parte de uma comunidade;

b. esse quadro é internalizado através dos processos de treinamento e indução pelos quais levam o indivíduo a se tornar membro dessa comunidade;

c. uma vez internalizado, o quadro epistêmico de uma comunidade é usado quando um indivíduo encara uma situação sob o ponto de vista de um membro da comunidade.

Tendo compreendido esse conceito fundamental, pode-se afirmar que o ENA é uma forma de análise de rede para avaliar os quadros epistêmicos (D. W. Shaffer et al., 2009). Esta técnica realiza uma análise baseada em grafos para examinar relacionamentos entre um conjunto de conceitos. Segundo D. W. Shaffer et al. (2016), o ENA possui um conjunto de passos que identifica e mede conexões entre elementos cognitivos em dados codificados e os representa em modelos de redes dinâmicas. As conexões são essenciais ao analisar as redes epistêmicas, porque fornecem enfoques úteis e mostram os aspectos salientes do objeto de estudo. Ao contrário de outras ferramentas de análise de rede, o ENA foi projetado principalmente para problemas com um conjunto relativamente pequeno de conceitos caracterizados por interações altamente dinâmicas e densas.

Dentro do ENA, as ligações entre os diferentes conceitos (*códigos*) são derivadas para cada *unidade de análise* (por exemplo, aluno) com base no conceito de co-ocorrências em subconjuntos de dados chamados *conversação* (por exemplo, frase, parágrafo, documento). A partir de co-ocorrências de código, o ENA cria primeiro uma representação de alta dimensão, denominada espaço analítico, de todas as unidades de análise. As unidades de análise são então projetadas em um espaço de representação inferior, chamado espaço de projeção, que é derivado do espaço analítico por meio do SVD.

Em geral, as aplicações de ENA são anotadas de acordo com a presença ou ausência de um determinado código específico. Contudo, o ENA também pode ser usado para códigos que representam a força ou a probabilidade de um dado código. Por exemplo quanto utilizado com a saída da etapa de mineração de tópicos. Nesse caso, o espaço analítico não é construído a partir da co-ocorrência do código, mas do produto ponderado dos valores dos códigos; a ponderação pode ser feita como: 1) produto direto, 2) raiz quadrada do produto direto, ou 3) logaritmo natural do produto direto. No final, a saída do ENA é uma série de modelos gráficos que capturam as relações entre diferentes categorias de codificação (D. W. Shaffer et al., 2016).

Para entender de forma prática e visual como o ENA opera, a Figura 5.3 apresenta o gráfico de projeção gerado. Note-se que as porcentagens no topo do eixo Y e ao final do eixo X representam a variança nos dados. Os pontos plotados no gráfico representam as *unidades de análise*, e os quadrados representam os centroides com seus respectivos intervalos de confiança ao redor, e apenas observando os centroides já é possível inferir onde

os pontos serão posicionados, nesse caso em lados opostos ao longo do eixo X. Através desse gráfico podemos determinar se a diferença dos grupos é estatisticamente significante (D. Shaffer, 2017), aplicando por exemplo o *t-test* ou o teste de *Mann-Whitney*.

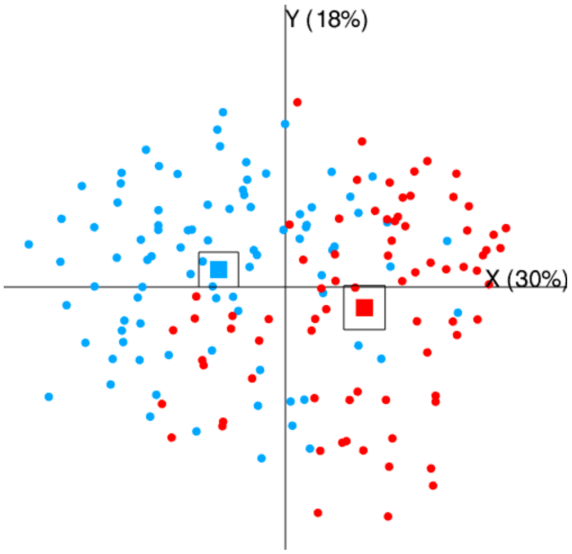


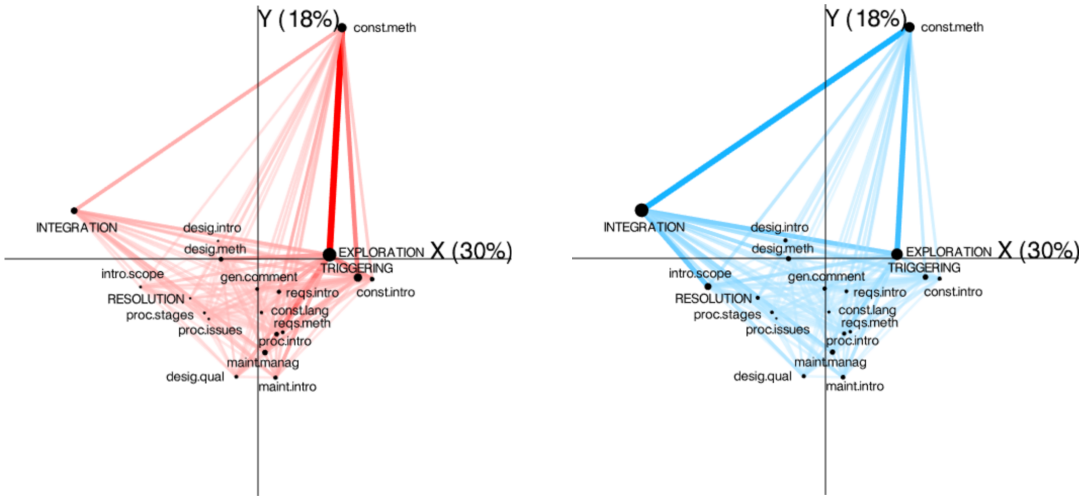
Figura 5.3: **Exemplo do gráfico de projeção do ENA**

Continuando a análise visual, as Figuras 5.4a e 5.4b apresentam análises dos relacionamentos de diferentes grupos. Neste caso ela apresenta os relacionamentos das médias de cada grupo apresentado na figura 5.3. Cada nó (ou vértice) representa um conceito (*código*) analisado, enquanto que as arestas representam a *conversação*. A partir desse grafo é possível analisar relacionamentos existentes entre os códigos selecionados e a intensidade desses relacionamentos (nesse caso, os relacionamentos são entre as presenças cognitivas e os tópicos de curso, de dois grupos diferentes), quanto mais escura é a cor da aresta, mais forte é a conexão. As figuras 5.4a e 5.4b deixam claro a interação entre esses dois conceitos. Por causa das posições fixas dos nós, o ENA constrói também uma rede de subtração, que habilita a identificar diferenças salientes na comparação de duas redes (veja-se a Figura 5.5). Para fazer isso, o ENA subtrai o peso de cada conexão de uma rede pelo peso da conexão diretamente correspondente na outra rede, após esse processo é possível observar as diferenças.

Adicionalmente aos pontos abordados acima, deve-se destacar na visualização das redes epistêmicas:

- o tamanho dos nós (códigos) representa a importância desse nó para a rede;
- a força da conexão entre dois códigos representa quão frequente é a co-ocorrência desses códigos;
- as posições dos nós dentro da rede indicam o quão semelhantes eles são uns aos outros, em outras palavras, a similaridade é proporcional a distância entre os nós.

Outra análise bastante utilizada quando se aplica ENA é o grafo de subtração. A Figura 5.5 mostra a rede de subtração entre os grupos 1 e 2 apresentados acima. Nessa



(a) **Grupo 1** (b) **Grupo 2**

Figura 5.4: **Exemplos das redes epistêmicas**

rede, os vértices permanecem no mesmo ponto das redes originais, mas as arestas agora tem cores diferentes. Nesse caso, o grupo que tiver um relacionamento mais intenso entre dois conceitos vai ter sua cor ressaltada. Quanto mais o relacionamento tende a ser igual para os dois grupos a cor da aresta tende a ser mais transparente. Enquanto no caso contrário, o grupo dominante tem uma cor mais intensa.

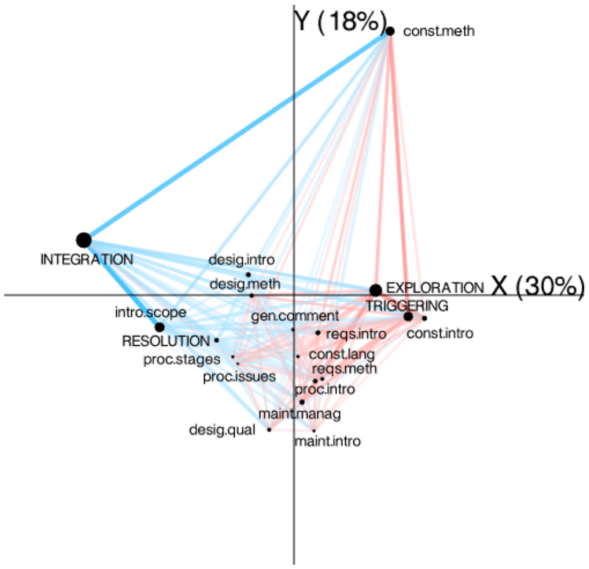


Figura 5.5: **Exemplo da rede de subtração do ENA**

Além dos gráficos mencionados, o ENA permite a representação das mudanças das redes ao longo do tempo. Esse gráfico é chamado de análise de trajetórias. No exemplo a seguir apresentaremos um gráfico de trajetória.

O ENA é uma técnica recente que fornece diversas informações sobre os relacio-

namentos de um determinado grupo, e essas informações tornam possível uma avaliação qualitativa desses relacionamentos. Apesar de ganhar cada vez mais espaço no ambiente acadêmico internacional, essa técnica ainda é pouco conhecida no cenário nacional, por isso a importância da apresentação da mesma. Para essa parte utilizaremos a aplicação em https://app.epistemicnetwork.org, que é disponível gratuitamente.

5.3.1. Aplicação de redes epistêmicas para análise de COI

O ENA é tipicamente utilizado na área educacional, o que inclui examinar as relações entre diferentes elementos em um conjunto de dados codificados, como transcrições do discurso do estudante codificados. Para exemplificar, utilizando técnicas de modelagem de tópico para inferir os tópicos presentes numa postagem de um estudante, esse texto pode ser representado pelos valores inferidos.

Compreender o desenvolvimento dos estudantes nas dimensões do COI não é uma tarefa simples, por isso um número crescentes de estudos tem abordado esse tema levando em conta adicionalmente o relacionamento epistêmico dessas dimensões (Rolim, De Mello, et al., 2019; Rolim, Ferreira, et al., 2019; Mello & Gašević, 2019; R. Ferreira et al., 2018). Nesse contexto, a ENA pode fornecer novas percepções qualitativas e quantitativas sobre o desenvolvimento das habilidades de pensamento social e crítico dos alunos em comunidades de investigação (Rolim, Ferreira, et al., 2019). Mais especificamente, a ENA pode ser utilizada para cumprir três objetivos:

1. descobrir ligações entre as presenças sociais e cognitivas das comunidades de investigação;
2. avaliar a eficácia de intervenções instrucionais na experiência do aluno, medida por conexões entre presenças cognitivas e sociais; e
3. explorar como a relação entre as presenças sociais e cognitivas muda ao longo do tempo durante um curso.

5.3.2. Exemplo da aplicação de ENA

O principal objetivo do exemplo apresentado a seguir é analisar os relacionamentos entre as fases da presença cognitiva e os indicadores da presença social (Rolim, Ferreira, et al., 2019). Os dados utilizados no presente estudo consistiram em seis ofertas (inverno 2008, outono 2008, verão 2009, outono 2009, inverno 2010, inverno 2011) de um curso pós-graduação nível de mestrado em engenharia de software oferecido inteiramente online, por meio do Moodle, em uma universidade pública canadense entre 2008 e 2011. Nessas seis ofertas, um total de 81 alunos postaram 1.747 mensagens. O curso abrangeu seis módulos que abrangeram 14 tópicos diferentes relacionados à engenharia de software. Os alunos foram avaliados pelos instrutores do curso em quatro tarefas (TMA1–4):

- **TMA1**: 15 % - apresentação de artigo publicado com revisão por pares sobre um dos tópicos do curso, através de publicações em um fórum educacional;
- **TMA2**: 25 % - redação de um artigo de revisão de literatura sobre um tópico selecionado em engenharia de software;

- **TMA3**: 15 % - respondendo a seis questões (uma para cada módulo) para demonstrar o pensamento crítico e habilidades de síntese;
- **TMA4**: 30 % - projeto final.

Como parte da avaliação TMA1, os alunos foram solicitados a selecionar um artigo de pesquisa sobre um tópico em engenharia de software, gravar uma apresentação de vídeo e postar um URL para uma nova discussão online do curso, na qual os outros alunos se envolveriam no debate em torno sua apresentação. Nesse caso, os alunos que postaram o vídeo foram considerados os *experts* das discussões, enquanto o resto da classe foram os *practicing*. A participação em tal discussão online respondeu pelos 15% restantes da nota (Gašević et al., 2015).

Durante as duas primeiras ofertas do curso, a participação dos alunos foi impulsionada principalmente por fatores motivacionais extrínsecos (ou seja, nota do curso). Neste estudo, os alunos das duas primeiras ofertas são referidos como grupo de controle (*control group*), que consistia em 37 alunos que produziram 845 mensagens. Após as duas primeiras ofertas de cursos, foi realizado uma intervenção pedagógica para incentivar a participação na discussão por meio de atribuições de funções e instruções claras. No total, 44 alunos, referidos como grupo de tratamento (*treatment group*), foram expostos a essa intervenção e produziram um total de 902 mensagens. Mais detalhes sobre a intervenção são apresentados em (Gašević et al., 2015).

Neste contexto, foi aplicada a análise de redes epistêmicas para identificar qual o relacionamento entre as presenças social e cognitiva, como os alunos de diferentes grupos interagiram no fórum e como essas interações mudaram ao longo do tempo. A Figura 5.6 apresenta a rede ENA contendo o relacionamento entre as duas presenças, gerada a partir da interação de todos os estudantes. Observamos que os indicadores da categoria **Interativa** foram localizados mais à direita do gráfico, especialmente o indicador "*Asking_Question*". Os indicadores da categoria *Coesiva* ficaram no meio e os da categoria **Afetiva** no lado esquerdo. Isso pode ser explicado pelo fato de que a categoria **Interativa** de presença social está geralmente associada à fase de **Evento Desencadeador** da presença cognitiva, enquanto a categoria **Afetiva** de presença social está relacionada a níveis mais elevados de fase cognitiva (Morueta et al., 2016).

Na Figura 5.7 é apresentado o gráfico de projeção das unidades analisadas pelo ENA, nesse caso os estudantes estão divididos em 4 categorias de acordo com o seu papel no fórum e com o tipo do fórum. Nesse gráfico é possível observar através dos agrupamentos apresentados como a categoria do estudante tem uma relação direta com o seu comportamento. Também para analisar esses grupos, a Figura 5.8 mostra o gráfico de subtração (ou seja, a diferença entre as duas redes) entre duas das categoria apresentadas na Figura 5.7: os especialistas (*experts*) e os pesquisadores praticantes (*practicing*). A figura revela que o grupo de especialistas tiveram mais conexões com os códigos dentro dos indicadores **Afetivos** e **Coesivos** da presença social do que os grupos de pesquisadores praticantes. O grupo de pesquisadores praticantes teve um maior número de conexões com os indicadores **Interativos** da presença social, especialmente para fazer perguntas (*Asking_Question*). Os especialistas tiveram mais conexões com as fases de **Integração** e **Resolução** da presença cognitiva. Os pesquisadores praticantes tendem a postar mais

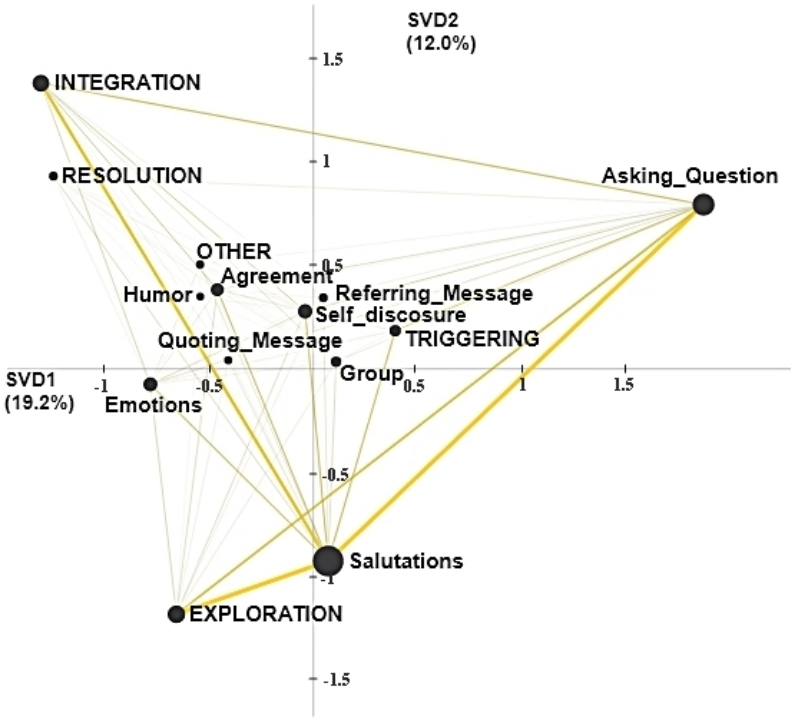


Figura 5.6: **Rede ENA dos relacionamentos entre as presenças social e cognitiva**

mensagens com conexões com as categorias de **Exploração** e **Evento Desencadeador** da presença cognitiva.

Observamos na Figura 5.9 que os gráficos de rede e de trajetória dos relacionamentos desenvolvidos entre as duas dimensões do modelo COI no decorrer de quatro semanas (duração do curso). Ambos os gráficos apresentados são plotados no mesmo espaço dimensional, o que possibilita realizar análises levando conta as correlações existentes.

A visualização da trajetória (Figura 5.9b) mostra a localização da atividade principal realizada por cada grupo de alunos em cada semana da discussão conectada por uma linha que representa como os grupos evoluíram de uma semana para a seguinte. Essa análise revelou que os alunos da maioria dos grupos representados demonstraram um progresso constante fazendo menos ligações com as fases dos níveis mais baixos da presença cognitiva e aumentando as ligações com os níveis mais altos da presença cognitiva com o passar das semanas do curso.

Este exemplo demostra como o ENA pode auxiliar no entendimento do relacionamento entre grupos e até mesmo na evolução de aspectos específicos, como a presença social e cognitiva. Examinando essas duas presenças no nível do aluno em vez de no nível da mensagem, um entendimento muito mais rico do desenvolvimento dos alunos foi adquirido, indo além da simples contagem de mensagens e correlações estatísticas. Além disso, a análise do desenvolvimento das presenças cognitivas e sociais do aluno ao longo

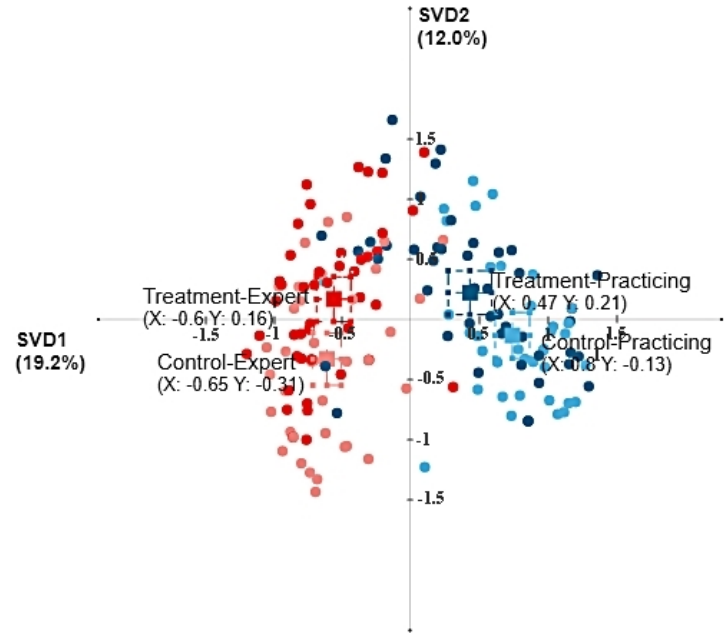


Figura 5.7: **Gráfico de projeção**

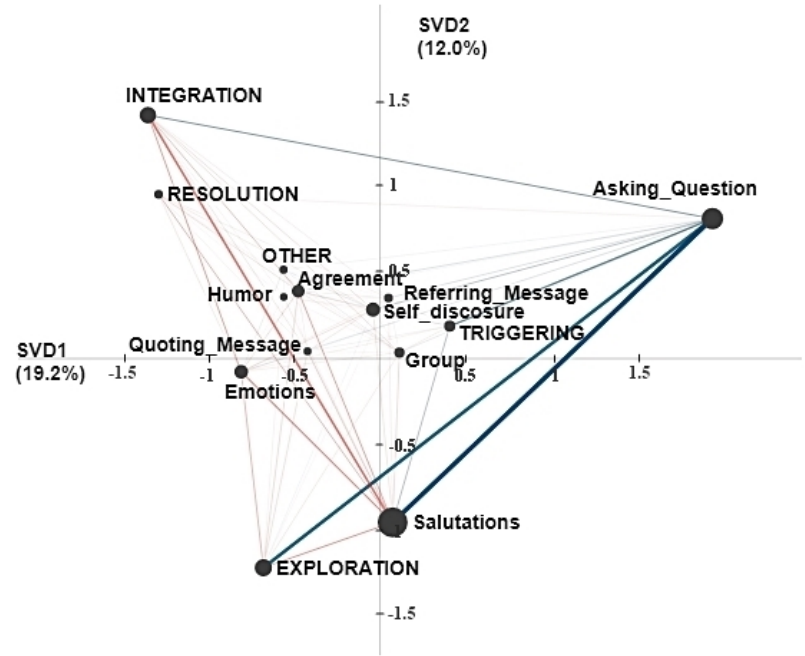
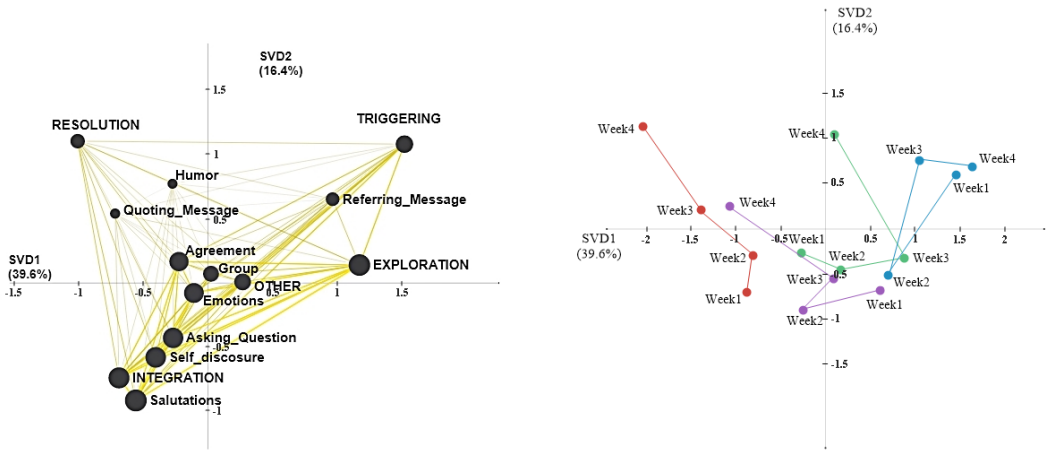


Figura 5.8: **Rede de subtração**



(a) **Rede geral**

(b) **Gráfico de trajetória. Cada cor representa uma categoria de estudante**

Figura 5.9: **Análise ao longo de 4 semanas da evolução dos alunos**

do tempo forneceu *insights* sobre como seu comportamento mudou em cada semana do curso.

A abordagem apresentada nesse exemplo com ENA, pode ser usada, por exemplo, como base para o desenvolvimento de uma plataforma que pode ajudar instrutores e alunos a alcançar uma experiência educacional aprimorada em discussões online assíncronas seguindo o modelo COI.

5.4. Considerações Finais

A educação contemporânea atravessa uma nova era nas modalidades de ensino superior a distância, e vem experimentando como a tecnologia pode beneficiar e otimizar as modalidades de ensino. Contudo, devido ao crescimento dessas modalidades de ensino, surgem diversos desafios a serem abordados pela comunidade científica, dada a importância de garantir uma boa experiência educacional aos estudantes que fazem uso dessas modalidades. O modelo CoI tem sido bastante utilizado no desafio da modelagem educacional, permitindo uma análise da aquisição do conhecimento e do desenvolvimento cognitivo do estudante. Por isso este trabalho se concentrou em analisar as presenças social e cognitiva, sendo pioneiro a realizar essa análise no contexto de fóruns utilizando ENA.

A utilização de técnicas como modelagem de tópicos e análise de redes epistêmicas quando juntas, possibilitam tanto uma análise quantitativa quanto qualitativa dos resultados, por disponibilizar dispositivos para o entendimento do discurso do estudante relacionado ao seu conteúdo e como isso está relacionado com o seu desenvolvimento cognitivo na disciplina e social com os seus pares.

Este capítulo apresentou os principais conceitos para criação de aplicação para entender os relacionamentos entre os estudantes desenvolvidos num fórum de discussão. Ressalta-se também a contribuição do ENA na avaliação qualitativa desses relacionamen-

tos. Esta técnica ainda é pouco utilizada no cenário nacional brasileiro.

Referências

Anderson, L. W., & Sosniak, L. A. (1994). *Bloom's taxonomy*. Univ. Chicago Press Chicago, IL.

Anderson, T., & Dron, J. (2010). Three generations of distance education pedagogy. *The International Review of Research in Open and Distance Learning*, *12*(3), 80–97. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/890/

Anderson, T., Rourke, L., Garrison, D. R., & Archer, W. (2001). Assessing Teaching Presence in a Computer Conferencing Context. *Journal of Asynchronous Learning Networks*, *5*, 1–17.

Azevedo, D., Ferreira, R., Mendonca, V., & Miranda, P. (2017). Aplicação de análise de sentimento em fóruns educacionais para prevenir evasão. In *Brazilian symposium on computers in education (simpósio brasileiro de informática na educaçao-sbie)* (Vol. 28, p. 1097).

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanović, V., & Gašević, D. (2020). Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 605–614).

Batista, E. M., & Gobara, S. T. (2007). O fórum on-line e a interação em um curso a distância. *RENOTE-Revista Novas Tecnologias na Educação*, *5*(1).

Biggs, J. B., & Collis, K. F. (2014). *Evaluating the quality of learning: The solo taxonomy (structure of the observed learning outcome)*. Academic Press.

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, *18*, 147.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003a). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944937

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003b). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Cavalcanti, A., & Ferreira, R. (2018). Uma medida de similaridade textual para identificação de plágio em fóruns educacionais. In *Anais dos workshops do congresso brasileiro de informática na educação* (Vol. 7, p. 63).

Cheng, C. K., Paré, D. E., Collimore, L.-M., & Joordens, S. (2011). Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers & Education*, *56*(1), 253–261.

D'Amato, D., Droste, N., Allen, B., Kettunen, M., Lähtinen, K., Korhonen, J., ... Toppinen, A. (2017). Green, circular, bio economy: A comparative analysis of sustainability avenues. *Journal of Cleaner Production*, *168*, 716–734.

Dawson, S., Tan, J. P.-L., & McWilliam, E. (2011). Measuring creative potential: Using social network analysis to monitor a learners' creative capacity. *Australasian Journal of Educational Technology*, *27*(6), 924-942.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Dewey, J. (1897). My Pedagogical Creed. *School Journal*, *54*(3), 77–80.

Dionísio, M., Ferreira, R., Cavalcanti, A., Carvalho, R., & Neto, S. (2017). Mineração de texto aplicada à identificação de colaboração em fóruns educacionais. In *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)* (Vol. 28, p. 1437).

Ferreira, M., Rolim, V., Mello, R. F., Lins, R. D., Chen, G., & Gašević, D. (2020). Towards automatic content analysis of social presence in transcripts of online discussions. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 141–150).

Ferreira, M. A. D., Mello, R. F. L., Garrozi, C., Rolim, V. B., & Cavalcanti, A. P. (2018). Um sistema baseado em pln e ag para apoiar a mediação pedagógica em fóruns de discussão. *Revista Brasileira de Informática na Educação*, *26*(03), 61.

Ferreira, R., Kovanović, V., Gašević, D., & Rolim, V. (2018). Towards combined network and text analytics of student discourse in online discussions. In *International conference on artificial intelligence in education* (pp. 111–126).

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(6), e1332.

Freitas, M., & Auxiliadora, S. (2009). *Avaliação da aprendizagem em ambientes de formação online: aportes para uma abordagem hermenêutica* (Unpublished doctoral dissertation). PhD thesis, UFBA: Faculdade de Educação.[GS Search].

Garrison, D. R., Anderson, T., & Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education*, *2*(2-3), 87–105.

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

Garrison, D. R., & Arbaugh, J. B. (2007). Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and Higher Education*, *10*(3), 157–172.

Gašević, D., Adesope, O., Joksimovic, S., & Kovanovic, V. (2015). Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, *24*, 53–65. doi: 10.1016/j.iheduc.2014.09.006

Hew, K. F., & Cheung, W. S. (2008). Attracting student participation in asynchronous online discussions: A case study of peer facilitation. *Computers & Education*, *51*(3), 1111–1124.

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with moocs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, *145*, 103724.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289–296).

Lipman, M. (1991). *Thinking in education*. Cambridge University Press.

Maskeri, G., Sarkar, S., & Heafield, K. (2008). Mining business topics in source code using latent dirichlet allocation. In *Proceedings of the 1st india software engineering conference* (pp. 113–120).

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121–128).

Mello, R. F., & Gašević, D. (2019). What is the effect of a dominant code in an epistemic network analysis? In *International conference on quantitative ethnography* (pp. 66–76).

Morueta, R. T., López, P. M., Gómez, Á. H., & Harris, V. W. (2016). Exploring social and cognitive presences in communities of inquiry to perform higher cognitive tasks. *The Internet and Higher Education*, *31*, 122–131.

Murphy, E. (2004). Recognising and promoting collaboration in an online asynchronous discussion. *British Journal of Educational Technology*, *35*(4), 421–431.

Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. *PloS one*, *9*(8).

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1-volume 1* (pp. 248–256).

Ramsden, P. (1988). *Improving learning: New perspectives*. Nichols Pub Co.

Rivard, R. (2013). Measuring the mooc dropout rate. *Inside Higher Ed*, *8*, 2013.

Rolim, V., de Mello, R. F. L., Ferreira, M., Cavalcanti, A. P., & Lima, R. (2019). Identifying students' weaknesses and strengths based on online discussion using topic modeling. In *2019 ieee 19th international conference on advanced learning technologies (icalt)* (Vol. 2161, pp. 63–65).

Rolim, V., De Mello, R. F. L., Kovanovic, V., & Gaševic, D. (2019). Analysing social presence in online discussions through network and text analytics. In *2019 ieee 19th international conference on advanced learning technologies (icalt)* (Vol. 2161, pp. 163–167).

Rolim, V., Ferreira, R., & Costa, E. (2016a). Identificação automática de dúvidas em fóruns educacionais. In *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)* (Vol. 27, p. 936).

Rolim, V., Ferreira, R., & Costa, E. (2016b). Método supervisionado para identificação de dúvidas em fóruns educacionais. In *Anais dos workshops do congresso brasileiro de informática na educação* (Vol. 5, p. 102).

Rolim, V., Ferreira, R., & Costa, E. (2017). Utilizaçao de técnicas de aprendizado de máquina para acompanhamento de fóruns educacionais. *Revista Brasileira de Informática na Educação*, *25*(03), 112.

Rolim, V., Ferreira, R., Lins, R. D., & Găsević, D. (2019). A network-based analytic approach to uncovering the relationship between social and cognitive presences in communities of inquiry. *The Internet and Higher Education*, *42*, 53–65.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (1999). Assessing Social Presence In Asynchronous Text-based Computer Conferencing. *The Journal of Distance Education*, *14*(2), 50–71. Retrieved from `http://www.ijede.ca/index.php/jde/article/view/153`

Shaffer, D. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.

Shaffer, D. W. (2006). Epistemic frames for epistemic games. *Computers & Education*, *46*(3), 223–234. doi: 10.1016/j.compedu.2005.11.003

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data. *Journal of Learning Analytics*, *3*(3), 9–45. doi: 10.18608/jla.2016.33.3

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., ... Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media*.

Shanmugam, R. (2019). *Practical text analytics: maximizing the value of text data: by murugan anandarajan, chelsey hill and thomas nolan, switzerland, ag, springer press, springer nature, 2019, 285+ xxviii pp., isbn: 978-3-319-956663-3*. Taylor & Francis.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems* (pp. 1385–1392).

Wanas, N., El-Saban, M., Ashour, H., & Ammar, W. (2008). Automatic scoring of online discussion posts. In *Proceedings of the 2nd acm workshop on information credibility on the web* (pp. 19–26).

Wise, A., Zhao, Y., & Hausknecht, S. (2014). Learning analytics for online discussions: Embedded and extracted approaches. *Journal of Learning Analytics*, *1*(2), 48–71.