DOI: 10.19180/978-85-99968-49-9.2

Mineração de dados textuais educacionais: experiências e perspectivas para a análise de postagens em fóruns de discussão

Breno Fabrício Terra Azevedo
Instituto Federal Fluminense [bterra@iff.edu.br]
Doutor em Informática na Educação/ UFRGS
Heluia Pereira Pinto Bastos
Instituto Federal Fluminense [hbastos@iff.edu.br]
Doutora em Informática na Educação/UFRGS

> *"A verdadeira substância da língua não é constituída por um sistema abstrato de formas linguísticas nem pela enunciação monológica isolada, nem pelo ato psicofisiológico de sua produção, mas pelo fenômeno social da interação verbal, realizada através da enunciação ou das enunciações. A interação verbal constitui assim a realidade fundamental da língua".*
>
> Mikhail Bakhtin, 2006, p. 117

Em face do número crescente de cursos *on-line*[1] e da expressiva quantidade de dados gerados por eles, a mineração de dados constitui uma ferramenta importante para instituições de ensino e, especialmente, para os envolvidos com a oferta de Educação a Distância (EaD) porquanto pode fornecer, de forma rápida e objetiva, informações relevantes dos alunos, entre elas: comportamento, características, padrões de aprendizagem e de relacionamento, interações discursivas, fornecimento de *feedback* (GARCIA et al., 2011).

Essa consolidação de atividades pedagógicas em suportes computacionais tradicionais ou portáteis tem motivado pesquisas sobre os múltiplos aspectos envolvidos nessas metodologias, fornecendo resultados importantes para a melhoria constante e personalização das práticas didáticas de ensino e da aprendizagem (COSTA et al., 2012). Do *design* instrucional aos processos de avaliação, entre outros, destacam-se, nesse contexto, as interações entre os usuários em ambientes virtuais de aprendizagem (AVA)

[1] No Brasil, por exemplo, o Censo realizado em 2012 pela Associação Brasileira de Educação a Distância (ABED, 2013), aponta a existência de 1.856 cursos autorizados/reconhecidos e 7.520 cursos livres (com base em 284 instituições respondentes).

– campo de estudo que se insere na área de Comunicação Mediada por Computador (CMC).

A CMC se refere a todo tipo de interação, síncrona ou assíncrona, via tecnologias da informação e da comunicação, seja por videoconferência ou mensagens de texto. Entretanto, em que pese a popularização crescente da comunicação por vídeo em telefones portáteis, as interações baseadas em texto continuam a prevalecer.

Antes da popularização da Internet, considerava-se que as interações por escrito eram "inferiores" às realizadas face a face (FaF) devido à falta de pistas não-verbais (expressões faciais, gestos, articulações sonoras, distância entre os falantes). Esse alto grau de "pistas filtradas" (*filtered cues*) poderia inviabilizar a formação de relacionamentos mais "verdadeiros" e "estáveis" (LOWENTHAL, 2009). Entretanto, a Teoria do Processamento Social da Informação de Walther (1996), formulada no início da comunicação via Internet, defende que ambas as formas de interlocução (CMC e FaF) são propícias ao desenvolvimento de relacionamentos, uma vez que os interlocutores, em ambiente digital, podem formar impressões sobre os outros com base apenas no conteúdo textual.

Dessa forma, a natureza dialógica da CMC é um aspecto importante em AVA por propiciar suporte acadêmico, intelectual e interpessoal, a troca de diferentes pontos de vista, de novos significados e, particularmente, o sentimento de pertencimento no grupo (BASTOS et al., 2013; BASTOS, 2012). Conforme Moller (1998), o sentimento de pertencimento se desenvolve com base na "territorialidade, permanência, forma de comunicação", entre outros, e contribui para a diminuição do sentimento de isolamento e eventual desistência do aluno.

De fato, as mensagens trocadas em ferramentas de CMC têm natureza híbrida, isto é, apresenta características do discurso oral e do escrito. Por serem usualmente redigidas com maior informalidade do que em outros trabalhos acadêmicos, as interações em fóruns e *chats* possibilitam, segundo Bastos (2012), que os interlocutores se tornem mais "próximos" e mais "presentes" no evento de comunicação.

Nesse contexto, os fóruns de discussão e as salas de bate-papo constituem espaços privilegiados para o ensino e aprendizagem em ambientes digitais por permitirem aos alunos trabalhar de forma colaborativa, compartilhar informações e saberes, desenvolver a produção de textos e fortalecer vínculos afetivos. Essas ferramentas de CMC são igualmente relevantes e adequadas na avaliação do desempenho discente, particularmente a do tipo formativa.

Fornecendo dados quantitativos e qualitativos das participações, o

intercâmbio discursivo em fóruns pode mostrar o percurso de aprendizagem (BASSANI, 2006; SÁNCHEZ, 2005). Segundo Bassani e Behar (2006), o mapeamento das postagens em fóruns de discussão pode auxiliar na avaliação discente porque (i) possibilitam que o aluno regule "seus processos de pensamento e aprendizagem"; (ii) permitem que o professor analise o processo de "construção dos alunos" pelo monitoramento das produções individuais e forneça "subsídios para possíveis/ necessários ajustes no processo ensino-aprendizagem"; (iii) evidenciam "processos coletivos de construção de conhecimento" uma vez que as trocas discursivas são "facilitadoras da aprendizagem".

Dessa forma, considera-se que o envolvimento em fóruns de discussão é uma atividade discente importante. Ao realizar a análise das postagens dos alunos, o professor pode diagnosticar informações sobre os mesmos. No caso de uma grande quantidade de alunos, o docente precisa despender muito tempo na análise das mensagens (AZEVEDO, 2012). Visando facilitar essa tarefa e identificar diferentes aspectos das construções textuais, técnicas de mineração de texto têm sido usadas, de forma crescente, como recurso auxiliar do corpo docente em AVA. A utilização de *softwares* desenvolvidos para minerar textos produzidos por alunos constitui um campo recente de investigação denominado "Mineração de Dados Textuais Educacionais" – MDTE – tema central deste estudo.

Este capítulo está estruturado em três grandes seções. A primeira traz os referenciais teóricos sobre Mineração de Dados Textuais Educacionais e a caracterização do que seja Mineração de Texto, além de sua íntima relação com outros campos de estudo. A seção 2 apresenta programas e experimentos realizados com postagens de alunos com diferentes objetivos. As considerações finais dos autores encontram-se na seção 3, incluindo as perspectivas da MDTE diante da constante evolução dos recursos tecnológicos e sua utilização no cenário educacional.

1. FUNDAMENTOS TEÓRICOS

Nesta seção de embasamento teórico, são apresentados os conceitos, as definições e a inter-relação entre as diferentes áreas de pesquisa voltadas para o processamento computacional de textos.

A evolução da Ciência da Computação e o grande volume de dados gerados em instituições escolares, especialmente em ambientes de ensino e aprendizagem *on-line*, têm resultado no desenvolvimento de *softwares* capazes de identificar e analisar diferentes aspectos em

documentos produzidos nesses contextos. Esses programas visam, sobretudo, auxiliar o trabalho docente no manuseio e na avaliação de grandes quantidades de textos redigidos por alunos nas ferramentas de comunicação fórum e *chat*.

Quando se trata de produções textuais, pode-se usar o termo Linguística Computacional (LC) para se referir, de forma genérica, aos campos de investigação que utilizam a linguagem humana em sistemas computacionais, seja como objeto ou como meio de investigação. Othero e Menuzzi (2005, p. 22) definem a LC como a "área responsável pela investigação do tratamento computacional da linguagem e das línguas naturais". Conforme os autores, o Processamento da Linguagem Natural (PLN)[2] e a Linguística de Corpus (LCp)[3] são subáreas da LC, podendo utilizar técnicas de Mineração de Texto (MT) como mecanismo auxiliar na pesquisa e no desenvolvimento de programas computacionais direcionados ao estudo da linguagem, pela coleta e análise de dados gerados em AVA.

A subseção seguinte caracteriza esses campos de trabalho.

1.1 Mineração de Dados Educacionais

Costa et al. (2012, p. 4) explicam que a mineração de dados em ambientes virtuais com contextos educacionais é uma área de investigação "emergente" que visa ao desenvolvimento e adaptação de

> [...] métodos e algoritmos de mineração existentes, de tal modo que se prestem a compreender melhor os dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem, tais como AVAs, Sistemas Tutores Inteligentes (STIs), entre outros.

Entre as muitas possibilidades de aplicação da Mineração de Dados Educacionais apontadas por Baker e Isotani (2011), destacam-se:

---

[2] Segundo a Comissão Especial de Processamento de Linguagem Natural da Sociedade Brasileira de Computação (http://www.nilc.icmc.usp.br/cepln/), a PLN é a área que "lida com problemas relacionados à automação da interpretação e da geração da língua humana em aplicações como Tradução Automática, Sumarização Automática de Textos, Ferramentas de Auxílio à Escrita, Perguntas e Respostas, Categorização Textual, Recuperação e Extração de Informação [...]". A PLN colabora, ainda, para o desenvolvimento e disponibilização de dicionários e corpus eletrônicos, além de contribuir com a Inteligência Artificial, notadamente em programas de interação humano-computador, *software* para reconhecimento e síntese de fala (*speech recognition, text to speech*), *chatterbots* (programas capazes de "conversar" com humanos), e *parsers* (analisadores sintáticos).

[3] Área que "se ocupa da coleta e exploração de *corpora*, ou conjunto de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística" (BERBER-SARDINHA, 2004, p. 3).

- obtenção de conhecimento científico relacionado aos estados emocionais do estudante (motivado, frustrado, confuso, entre outros);
- identificação da relação entre estudos emocionais e o comportamento apresentado pelo discente;
- utilização de *softwares* "inteligentes" para fornecimento de suporte e *feedback* apropriados para melhorar a qualidade da aprendizagem do estudante;
- verificação das colaborações dos estudantes nos tópicos dos fóruns de discussão;
- identificação de quem interagiu com quem.

Os autores enfatizam que tais aspectos podem ajudar a compreender quais processos de interação auxiliam a aprendizagem e quais deles a dificultam.

Com propósitos específicos de recuperar e fornecer informações relevantes em contexto educacional, a Mineração de Dados Educacionais recorre, frequentemente, não só às técnicas de Mineração de Texto (MT), mas também à Análise de Conteúdo (AC) para, entre outros: (i) extrair e identificar opiniões; (ii) facilitar o processo de automático de codificação de postagens; (iii) avaliar a sequência dos comentários em fóruns de discussão (*thread discussions)*; (iv) identificar padrões de interação em chats; (v) verificar a qualidade da participação dos alunos (ROMERO; VENTURA, 2010).

A MDTE é uma aplicação da "Descoberta de Conhecimento em Textos" (*Knowledge Discovery in Text – KDT)*[4], área abrangente que trata dos "problemas relacionados ao entendimento, resumo e tratamento de informações (transformando-as em conhecimento útil e aplicável)" (WIVES; LOH, 1999, p.1). A KDT engloba, ainda, a subárea de Mineração de Texto (MT).

As relações de proximidade e interdependência entre as áreas de processamento de dados textuais estão resumidas na figura 1.

O detalhamento e a caracterização da Mineração de Texto e da Análise de Conteúdo encontram-se nas subseções que seguem.

[4] Os termos "mineração de texto" e "descoberta de conhecimento em textos" são considerados sinônimos para Barion e Lago (2008).

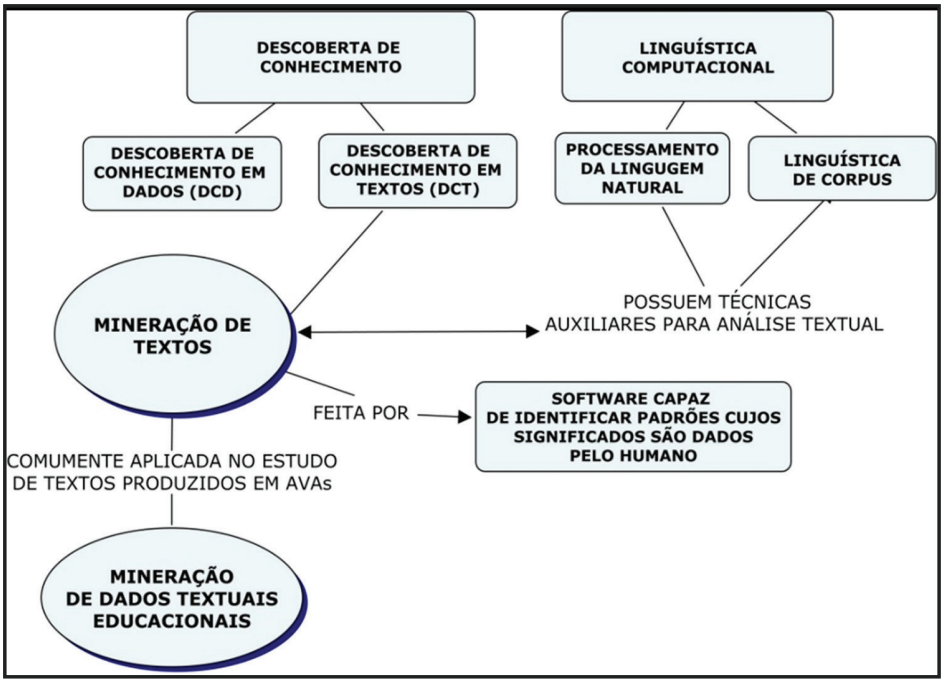

Figura 1 - Relações de proximidade e interdependência entre as áreas de processamento de dados textuais
Fonte: Elaboração própria (2014).

1.2 Mineração de Texto

A Mineração de Texto (MT) é o processo no qual um usuário interage com uma coleção de documentos utilizando um conjunto de ferramentas de análise (FELDMAN; SANGER, 2007). Conhecida, também, como Mineração de Dados Textuais (*Text Data Mining*) e Descoberta de Conhecimento a partir de Bancos de Dados Textuais (*Knowledge Discovery from Textual Databases*), a MT é uma área da Ciência da Computação que visa à descoberta de informações novas, ou desconhecidas, por meio da extração automática dos dados em documentos escritos (GUPTA; LEHAL, 2009).

A MT é análoga à Mineração de Dados, distinguindo-se desta por focar no processamento de dados não-estruturados[5] ou semi-estruturados (*e-mails*, arquivos em diferentes formatos, páginas em HTML, entre outros), para fornecer uma visualização final sistematizada do documento (FELDMAN; SANGER, 2007; FAN et al., 2006).

[5] Devido à sua estrutura semântica ou sintática, Feldman e Sanger (2007) relatam que todo documento pode ser um "objeto estruturado". Elementos tipográficos e o *layout* do texto são exemplos de estruturação em documentos.

O processamento computacional de documentos escritos costuma se apoiar nas seguintes abordagens (LOPES et al., 2009; MORAIS, AMBRÓSIO, 2007):

- Estatística: em que se mede a frequência de ocorrência dos termos;
- Linguística: em que os termos são anotados segundo sua classificação morfológica, sintática e semântica, constituindo as chamadas *tags*. Essa abordagem deve considerar, também, o "discurso" como um todo, particularmente os aspectos pragmáticos (o uso da língua em diferentes eventos comunicativos que afetam a produção textual);
- Híbrida: na qual são usadas as duas técnicas anteriores de forma conjunta.

Segundo Fan et al. (2006), a MT costuma empregar as seguintes técnicas: recuperação de informação, classificação ou categorização, extração de informação, sumarização e agrupamento (*clustering*).

Lidando com caracteres, termos (palavras ou sintagmas) e conceitos encontrados em *corpora*, a MT envolve as seguintes etapas:

- *Coleta ou Recuperação da informação:* localização e recuperação de documentos considerados relevantes para o estudo (o *corpus* de análise);
- *Pré-processamento*: em que se faz a "limpeza" do documento pela remoção dos elementos não necessários à compreensão textual. Essa limpeza envolve passos (não obrigatórios conforme o objetivo da pesquisa):
    - Correção ortográfica: quando os termos do documento são comparados aos verbetes de um dicionário (por exemplo, o programa "br.inspell para o português do Brasil, distribuído sob a licença GNU *General Public License*);
    - Remoção de *stopwords*: filtragem de palavras sem significado semântico relevante (artigos, preposições, conectivos, pronomes relativos, entre outros) visando diminuir o *corpus* de análise;
    - Etapa de *stemming:* nesta fase as palavras são reduzidas à sua unidade mínima de significação (*radical / stem*), como marcas de plural, de conjugação verbal, de gênero, entre outros. A etapa de *stemming* converte todas as palavras com um mesmo radical à sua unidade básica;
    - Etiquetamento morfossintático (*Part-of-Speech Tagging*): as palavras são etiquetadas com base no contexto textual, isto é, com base na função que exercem nas frases;

- Processamento: separação das partes constitutivas do texto (capítulos, seções, parágrafos, sentenças, palavras, sílabas e fonemas). O sistema mais usado é a separação de frases e palavras – *tokens* (KAPLAN, 2005; FELDMAN; SANGER, 2007). O processo de tokenização (*tokenization*) implica inúmeras dificuldades, entre elas, a distinção entre palavras e abreviações, palavras compostas, e ambiguidade[6] de significados (KAPLAN, 2005). Essa etapa envolve, ainda, o desenvolvimento dos algoritmos a serem utilizados na mineração textual;
- Pós-processamento: etapa em que se faz a avaliação e a validação dos resultados visando obter melhor conhecimento do algoritmo usado na mineração.

O analista pode voltar e rever cada passo do processo de modo a refinar os dados obtidos. As etapas do processo de mineração de textos encontram-se ilustradas na figura 2.

Garcia et al. (2011) alertam para o fato de que a maioria das ferramentas de MT são muito complexas, podendo ser de difícil utilização por professores e tutores. Os autores consideram que os programas de MT devem ter interfaces mais intuitivas e com recursos de colaboração que tornem os resultados disponíveis para aprimoramento e utilização em outros cursos.

Figura 2 - Etapas do processo de mineração de textos
Fonte: Elaboração própria (2014).

Entre as possibilidades de aplicação da MT, Pang e Lee (2008) destacam estas abordagens usadas na inferência de afetividade em textos: Mineração de Opinião (*Opinion Mining*), Análise de Sentimento (*Sentiment Analysis*), Análise da Subjetividade *(Subjectivity Analysis*), também conhecida como Análise de Julgamento (*Appraisal Analysis*). Na visão dos autores, a Análise de Sentimento visa determinar a atitude dos falantes em relação a algum assunto ou ao texto como um todo.

É importante destacar a utilização frequente da Análise de Conteúdo em conjunto com a MT na descoberta de conhecimento em textos.

[6] A resolução de ambiguidades costuma ser feita por "analisadores sintáticos" (*parsers*). Esses programas realizam a "interpretação automática (ou semiautomática) de sentenças" pela classificação morfossintática de palavras e expressões nas frases (OTHERO; MENUZZI, 2005).

1.3 Análise de Conteúdo

A Análise de Conteúdo (AC) é definida por Bardin (2010, p. 44) como "um conjunto de técnicas de análise das comunicações visando obter, por procedimentos, sistemáticos e objetivos de descrição de conteúdo das mensagens, indicadores (quantitativos ou não) que permitam a inferência de conhecimentos relativos às condições de produção / recepção (variáveis inferidas) destas mensagens". Esses processos possibilitam determinar a presença de palavras e conceitos (*unidades de registro* ou *unidades de análise*) em determinado texto ou conjuntos de textos (*corpora*), permitindo a análise dos dados de forma qualitativa (busca de dados não-explícitos) ou quantitativa (número de ocorrências do termo).

A AC pode se apoiar em duas técnicas: a *análise lexical* e a *análise categorial* (BARDIN, 2010). A primeira busca a taxa de ocorrência das unidades lexicais consideradas relevantes na pesquisa (o "repertório léxico" dos sujeitos da investigação). Por sua vez, o levantamento categorial envolve o "desmembramento do texto" em categorias determinadas segundo os objetivos da pesquisa.

Essas técnicas de análise costumam seguir os passos mostrados na figura 3.

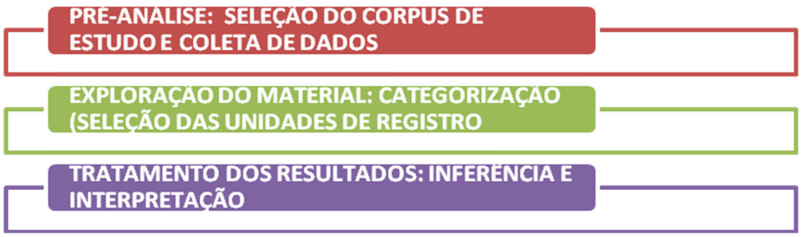

Figura 3 - Etapas da Análise de Conteúdo
Fonte: Elaboração própria baseada em Bardin, 2010, p.128.

2. MINERAÇÃO DE TEXTO NA ANÁLISE DE POSTAGENS EM FÓRUNS

Os fóruns são ferramentas de interação em AVA que permitem o trabalho colaborativo, o compartilhamento de informação, de opiniões e sentimentos. Constituem, portanto, espaços privilegiados para a socialização dos sujeitos em cursos a distância. Os fóruns são também um recurso chave na avaliação formativa uma vez que possibilitam aos professores e/ou tutores verificar o progresso dos alunos. Como costumam apresentar uma quantidade expressiva de dados textuais, o monitoramento

e a análise desses dados constituem uma tarefa complexa e extenuante (MACEDO et al., 2011).

Segundo Bastos (2012), a observação cuidadosa das postagens envolve "ler-ouvir" e "escrever-dizer", em outras palavras, manter diálogo constante com e entre os alunos, fato que justifica o desenvolvimento de programas de MT que podem, entre outros, fornecer ao professor / tutor informações que justificam intervenções pontuais.

Os *softwares* de MT precisam ser integrados ao ambiente virtual e apresentar, em uma só interface, todos os recursos de mineração (pré-processamento, mineração e pós-processamento). Dessa forma, os programas ficam melhor disponibilizados para uso por parte do professor / tutor, fornecendo *feedback* rápido para a tomada de decisões (ROMERO; VENTURA, 2010).

2.1 Experimentos de mineração de interações textuais com geração de grafos

Com o objetivo de apoiar o acompanhamento docente em face do grande volume de produções textuais no ETC – Editor de Texto Coletivo (ZANK, 2010), Macedo (2010) utilizou técnicas da mineração de textos para desenvolver a ferramenta "Rede de Conceitos". O *software* fornece dados quantitativos e qualitativos em forma de grafos, permitindo a identificação de autores (alunos) que demandam atenção. Os resultados do processamento textual podem indicar, entre outros, a necessidade de aprofundamento teórico.

Os grafos também são usados para visualização de dados minerados no software SOBEK (LORENZATTI, 2007), programa aplicado no estudo de Corrêa, Reategui e Biazus (2012), para mostrar se as postagens estão coerentes com a tarefa proposta. Outro estudo utilizando o SOBEK foi realizado por Klemann, Lorenzatti e Reategui (2009) visando verificar como esse programa pode auxiliar a produção textual. Nesse experimento, o processamento textual se deu pela "extração de termos frequentes; criação de uma base de conceitos e relacionamento a partir dos termos extraídos automaticamente; geração de um grafo correspondente aos termos e relacionamentos estabelecidos; escrita do texto com base no grafo gerado" (KLEMMAN et al., 2009, p. 9). Os autores concluíram que o processo de elaboração textual com ajuda do software promoveu melhor compreensão e tratamento do tema proposto. A figura 4 apresenta o grafo gerado no experimento sobre o tema "Escrita".

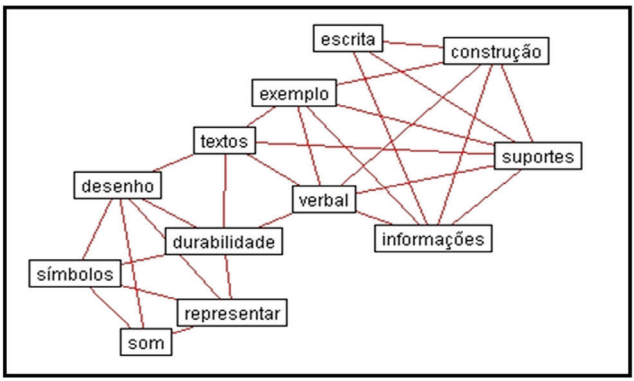

Figura 4 - Exemplo de grafo gerado pelo *software* SOBEK
Fonte: Klemman et al., 2009, p.6.

Azevedo (2011) desenvolveu o MineraFórum – *software* que realiza a análise qualitativa das mensagens em fóruns de discussão. O programa calcula a relevância da postagem em relação ao tópico proposto para debate, e a análise das contribuições textuais dos alunos é feita por meio de mineração de texto e subsequente geração de grafos. A relevância temática da(s) postagem(ns) é definida pela "relevância da mensagem (RT)", pela "relevância de citações da mensagem (RM)", e pela "similaridade da mensagem (SM)".

A análise das postagens fornece maiores subsídios para avaliação do desempenho do aluno e a elaboração de estratégias de fomento às discussões na ferramenta fórum (AZEVEDO, BEHAR, REATEGUI, 2011). A Figura 5 apresenta a interface do MineraFórum após a seleção da opção "Minerar Fórum" com as mensagens agrupadas por aluno.

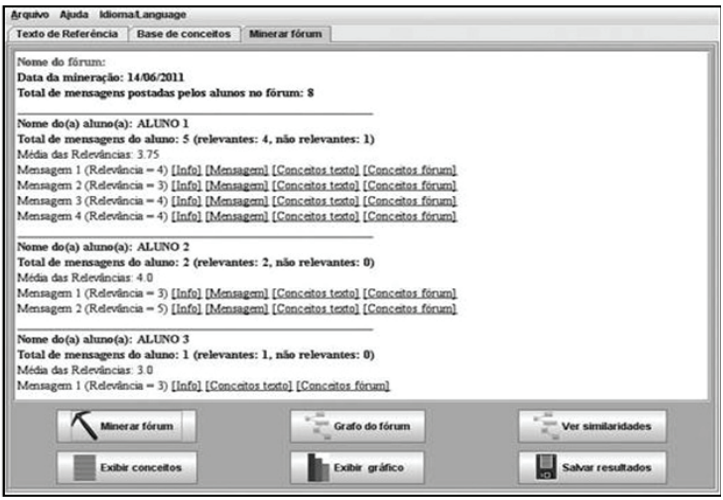

Figura 5 - Interface do MineraFórum após a seleção da opção "Minerar Fórum"
Fonte: Azevedo, 2011b, p.24.

2.2 Identificação de estados afetivos do aluno pela análise de interações discursivas

A relevância da afetividade no desenvolvimento sociocognitivo é apontada por Piaget (2007), Vigotski (1998), Freire (2007), que a consideram indissociável dos aspectos cognitivos. Entre as várias acepções para "afetividade", este trabalho se baseia em Bercht (2001, p.59), para quem o termo se refere ao conjunto de fenômenos de ordem física e psíquica, incluindo "o domínio das emoções propriamente ditas, dos sentimentos, das emoções, das experiências sensíveis e, principalmente, da capacidade em se poder entrar em contato com sensações". Considerando o distanciamento físico inerente à EaD, a comunicação por escrito é a forma mais recorrente para se construir e manter vínculos afetivos entre os participantes de cursos a distância (BASTOS et al., 2013).

São apresentados, a seguir, estudos de mineração textual realizados com o objetivo de verificar sentimentos e grau de relacionamento entre os sujeitos em AVA, assim como entre estes e o próprio ambiente de aprendizagem.

Com o objetivo de analisar o "conteúdo emocional" de textos redigidos por alunos, Longhi et al. (2010) usaram o *framework* AWM (*Affect Word Mining*)[7] para identificar e classificar termos com conotação afetiva em postagens de fórum no ambiente ROODA[8]. Os autores consideraram palavras de conotação afetiva aquelas que exprimem sentimentos, desejos e julgamento, além de adjetivos que indicam valor positivo ou negativo.

O processo de mineração no AWM identifica lexemas afetivos que são, em seguida, classificados conforme os "estados de ânimo" estabelecidos na "Roda dos Estados Afetivos – REA" (LONGHI; BEHAR; BERCHT, 2009). Os quadrantes e subquadrantes da REA apresentam os seguintes "estados de ânimo": *satisfeito, insatisfeito, desanimado, animado.* Após a identificação das palavras de significado afetivo na etapa de *tokenização*, a validação de seu "caráter afetivo" é checada no banco WordAffectBR(adapt)[9].

A arquitetura do *Affect Word Mining* encontra-se na figura 6.

---

[7] *Software* desenvolvido no Núcleo de Tecnologia Digital Aplicada à Educação (NUTED) da Universidade Federal do Rio Grande do Sul.

[8] Rede Cooperativa de Aprendizagem. Disponível em: <http://www.ead.ufrgs.br/rooda>.

[9] Extensão do WordAffectBR, banco lexical de termos de natureza afetiva na língua portuguesa idealizado por Pasqualotti (2008).

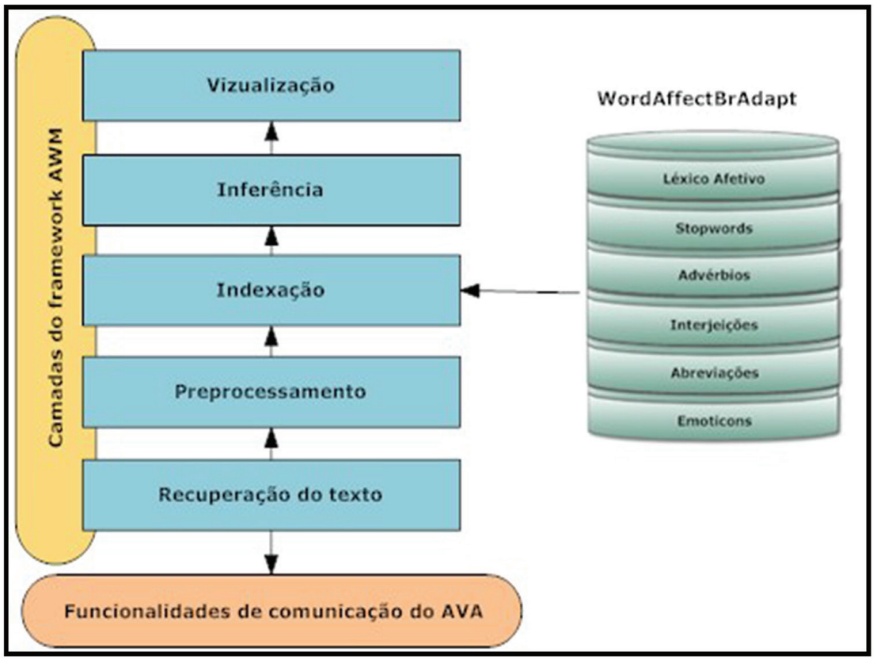

Figura 6 - Arquitetura do *framework* AWM
Fonte: Longhi et al., 2010, p.6.

A satisfação do aluno e sua permanência em cursos a distância podem ser identificadas por seu grau de "presença social" (PS). Entre as várias definições para PS, toma-se a de Bastos (2012, p. 23) em que PS é a "manifestação verbal e percepção da afetividade e interatividade dos sujeitos em relação ao ambiente virtual de ensino e aprendizagem". Para verificar o grau de PS em fóruns e *chats* ofertados em AVA, Kambara-Silva (2011) desenvolveu o *software Presente!* – um programa de mineração que fornece esse índice conforme as categorias de análise dadas pelo Modelo Presença Plus – PPLUS (BASTOS, 2012; BASTOS; BERCHT; WIVES, 2011). O Modelo PPLUS apresenta quatro grandes classes de indicadores textuais: *afetividade, interatividade, coesão e força.* Essas classes contêm várias subcategorias e unidades de análise de modo a contemplar as diversas estratégias discursivas usadas pelos participantes nos eventos de comunicação.

Não podendo utilizar um analisador sintático, devido ao alto custo, para realizar a análise textual proposta por Bastos (2012), Kambara-Silva (2011) desenvolveu um programa que faz a análise lexicométrica[10] do *corpus* linguístico selecionado. Entre outras características, o *Presente!* (i) permite

[10] A lexicometria faz tratamento estatístico de dados qualitativos gerando uma caracterização topológica e combinatória das pistas discursivas no *corpus* de estudo (DAMASCENO, 2008).

ao usuário acrescentar novas pistas textuais; (ii) gera resultados por classe e subclasse do PPlus, (iii) fornece relatórios dos graus de PS por aluno, por turma, por curso, por fórum, entre outros. A imagem mostra a ferramenta "construtor de categorias" com suas três áreas: (i) a da esquerda, com as classes e pistas textuais correspondentes, (ii) a do meio, que permite a alteração e inclusão de novas pistas, (iii) a da direita, que permite a seleção do tipo análise que se deseja realizar.

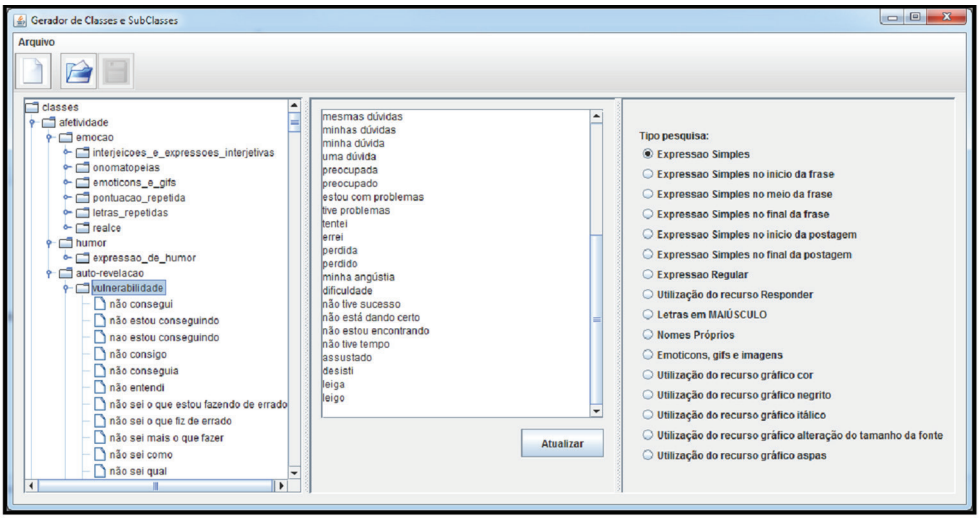

Figura 7 - Interface do módulo "Construtor de Categorias" do *software Presente!*
Fonte: Kambara-Silva, 2011, p. 21.

Apesar de alguns impasses ainda por serem superados, os experimentos feitos com o *software Presente!* mostraram que o programa é capaz de identificar as marcas discursivas cadastradas no programa e apontar o grau de PS dos sujeitos envolvidos nos testes. Uma necessidade verificada é integrar o *software* ao um banco de dados lexicais visando a constantes atualizações das pistas lexicais oferecidas no *software*.

Com o objetivo de facilitar o trabalho docente pela identificação de colaborações que demandam maior atenção, Oliveira Jr. e Esmin (2012) conceberam a ferramenta "Classificador de Fóruns" que classifica mensagens em positivas ou negativas. São exemplos de mensagens consideradas positivas no experimento: "*Estou muito satisfeito; Ótimo curso; Achei muito legal".* Esses trechos foram classificados como tendo natureza negativa: "*Falta de suporte de professores; Preciso de orientação; Aguardo contato urgente".* Para realização do experimento, os autores integraram seu classificador ao *software* GiAva – "ambiente de gestão e acompanhamento de qualidade

em AVA", desenvolvido por Esmin et al. (2010). A figura 8 mostra como as mensagens são apresentadas.

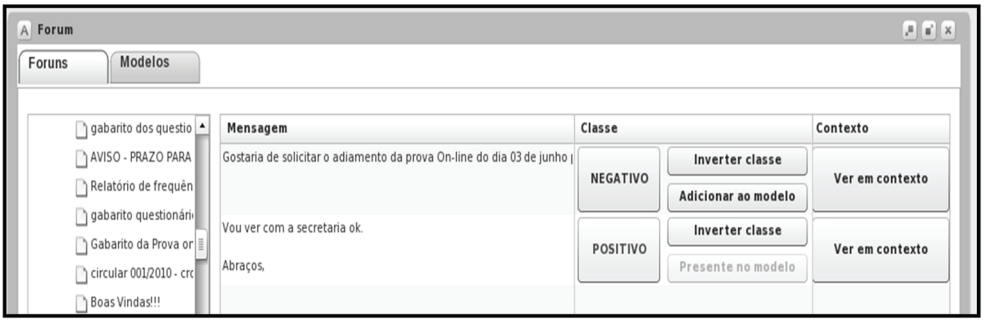

Figura 8 - Interface de apresentação das mensagens
Fonte: Oliveira Jr.; Esmin, 2012, p.6.

O experimento de D'Mello et al. (2008) usou a ferramenta AutoTutor – um sistema de tutor inteligente que monitora as emoções e a aprendizagem dos alunos por meio de interações em linguagem natural. No estudo citado, os autores objetivaram detectar o estado afetivo dos participantes durante a realização de uma atividade de aula em ambiente virtual. Os dados foram avaliados por alunos e dois julgadores treinados para a observação e comparados com os fornecidos pelo programa, gerando resultados semelhantes. Os estados afetivos colhidos no experimento foram, em sua maioria, *tédio, confusão, frustração e neutralidade*. Os autores consideram que versões mais aperfeiçoadas do *software* AutoTutor poderão fornecer *feedback* e correções de forma estabelecer maior empatia com o aluno, aliviar os sentimentos negativos e aumentar seu interesse no curso.

2.3 Estudos de classificação e descoberta de padrões em postagens de alunos

Percebendo a necessidade de se fazer o monitoramento automático de fóruns de discussão, Oliveira Jr. e Esmin (2012) desenvolveram uma ferramenta com uso de algoritmo semissupervisionado – SVM-KNN (LIN; HSIEH; CHUANG, 2009). Segundo os autores, um algoritmo desse tipo "aprende a partir de um pequeno número de dados rotulados juntamente com informações e estruturas internas contidas em um grande número de dados não rotulados". Para o estudo, as mensagens foram classificadas em "negativas" (quando continham dúvidas, insatisfação ou conteúdo indevido) ou "positivas" (sem padrões para "negativas"). A opção pelo algoritmo SVM-KNN mostrou-se adequado por sua alta taxa de acerto.

O fluxo das postagens em fóruns foi objeto da investigação de Chen e Chiu (2008). Para analisar como as primeiras mensagens afetam as posteriores, os autores consideraram cinco dimensões: (i) julgamento (concordância, discordância e mensagens não respondidas); (ii) conhecimento (contribuição, repetição e falta de conteúdo pertinente); (iii) marcas sociais (positivas, negativas ou nenhuma); (iv) informações pessoais (número de visitas); (v) solicitação (obtenção de respostas ou nenhuma). A pesquisa mostrou que discordância, contribuição, marcas sociais e visitas a mensagens anteriores podem afetar as postagens subsequentes, fato que pode auxiliar o professor a gerenciar o nível das discussões e facilitar o debate de questões polêmicas.

Lin, Hsieh e Chuang (2009) propõem um sistema de classificação das discussões em cascata feita em fóruns por gêneros textuais (anúncio, explicação, pergunta, interpretação, afirmação, conflito, entre outros). Essa classificação – *Genre Classification System* – visa à facilitação da codificação de conteúdos em fóruns. Os resultados obtidos no estudo são inconclusivos quanto à precisão dos resultados obtidos pela mineração de texto quando comparados ao do analista humano.

Baseados na Teoria dos Atos de Fala[11] de Austin (1962) e Searle (1981), Ravi e Kim (2007) apresentam uma abordagem para realizar a identificação automática dos tipos de interações em fóruns, particularmente questionamentos sem respostas que demandam atenção do docente. Para o estudo, foi definido um conjunto de "atos de fala" para relacionar as mensagens dentro de uma sequência, ou seja, em relação às postagens anteriores – "perguntas", "respostas", "elaboração" e/ ou "correção". A definição dessas categorias / classificadores foi obtida pela "sequência de palavras" e algoritmos SVM (*Support Vector Machine*). Os autores alertam para a dificuldade de se definir unidades de análise em fóruns porque as discussões, nessa ferramenta, costumam ocorrer de forma não-estruturada e sem coerência.

Considerando que o desempenho do aluno em ambiente virtual não deve se limitar aos aspectos meramente quantitativos (registros de entrada, número de postagens e outros), Bassani e Behar (2006), baseadas no ideário construtivista-interacionista, desenvolveram a ferramenta interROODA. As autoras apresentam um modelo de mapeamento das interações no módulo "Trocas Interindividuais" do ambiente ROODA[12], categorizando as mensagens em "enunciados" (mensagens que iniciam o diálogo) e "citações" (mensagens de resposta aos enunciados). Conforme Bassani e Behar (2006,

[11] A noção fundadora da Teoria dos Atos de Fala (*Speech Act Theory*) é que "todo dizer é um fazer". Isso significa que o principal objetivo da linguagem não é informar, mas realizar algum tipo de ação.

[12] A Rede Cooperativa de Aprendizagem – ROODA – é um *software* livre para apoiar o ensino e aprendizagem em ambiente virtual desenvolvido pelo Núcleo de Tecnologia Aplicada à Educação – NUTED / UFRGS.

p.7), o "mapeamento das trocas interindividuais pretende refletir a dinâmica das interações que se constituem entre os sujeitos participantes de um AVA". Da mesma forma, o "percurso da aprendizagem" pode ser avaliado de forma contextualizada.

Com a finalidade de obter um retrato mais abrangente das interações em fóruns, Li e Huang (2008) apresentam um modelo de análise multidimensional em que são aplicadas as seguintes abordagens: (i) Análise de Conteúdo para investigar como se dão as interações e descobrir padrões discursivos nas postagens; (ii) Mineração de texto para identificar os tópicos usados para debate. Além dessas abordagens, o estudo contou com o "componente exportador de dados" do VINCA[13] para analisar o *corpus* de estudo e identificar o padrão de interações entre os pares, tópicos de discussão mais usados no experimento e a rede de relacionamentos desenvolvida pelos participantes.

3. CONSIDERAÇÕES FINAIS

Este capítulo discorreu sobre o campo de pesquisa "Mineração de Dados Textuais Educacionais", mostrando, particularmente, a relevância de se fazer a mineração de interações discursivas realizadas na ferramenta fórum de discussão. O monitoramento e a análise de postagens feitas por alunos constituem um instrumento auxiliar da avaliação de seu desempenho, sendo, entretanto, uma tarefa complexa e morosa para professores e /ou tutores. Para facilitar o trabalho, programas de mineração de texto, tais como os apresentados neste trabalho, podem mostrar diferentes aspectos da participação do aluno (relevância dos comentários, envolvimento afetivo, entre outros) e orientar as devidas tomadas de decisão.

Considerando a complexidade e os múltiplos fatores envolvidos na elaboração dos intercâmbios linguísticos, pode-se afirmar que o processamento automático das interações feitas em ambientes virtuais oferece desafios e um largo espectro de possibilidades de estudo.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE EDUCAÇÃO A DISTÂNCIA – ABED. *Relatório Analítico da Aprendizagem a Distância no Brasil 2012*. Curitiba: Ibpex, 2013.

[13] URL do *software*: <http://vincaconnect.com>.

AUSTIN, J. L. *How to do things with words*. The William James Lectures delivered at Harvard University in 1955. Oxford: Clarendon, 1962.

AZEVEDO. B. F. T. *MineraFórum: um recurso de apoio para análise qualitativa em fóruns de discussão*. 2011. Tese (Doutorado em Informática na Educação) - Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, 2011.

AZEVEDO, B. F. T; BEHAR, P. A.; REATEGUI, E. B. Análise das mensagens de fóruns de discussão através de um software para mineração de textos. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO – SBIE, 22, 2011, Aracaju. *Anais...*, Aracaju, SE, 2011, p. 20-29.

AZEVEDO, B. F. T.; BEHAR, P. A.; REATEGUI, E. B. Automatic Analysis of Asynchronous Discussions. In: INTERNATIONAL CONFERENCE ON COMPUTER SUPPORTED EDUCATION, 4, 2012, Porto. *Proceedings...* Porto, 2012. v. 1. p. 5-12.

BAKER, R. J. D.; ISOTANI, S. Mineração de dados educacionais: oportunidades para o Brasil. *Revista Brasileira de Informática na Educação - RBIE*, v. 19, n. 2, 2011.

BAKHTIN, M. (Volochinov). *Marxismo e filosofia da linguagem*. 12. ed. São Paulo: Hucitec, 2006.

BARDIN, L. *Análise de conteúdo.* Lisboa: Edições 70, 2010.

BARION, E. C.; LAGO, D. Mineração de textos. *Revista de Ciências Exatas e Tecnologia*, v. 3, n. 3, 2008.

BASSANI, P. S. *Mapeamento das interações em ambiente virtual de aprendizagem: uma possibilidade para avaliação em educação a distância*. 2006. Tese (Doutorado em Informática na Educação) - Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, 2006.

BASSANI. P. S.; BEHAR, P. A. Análise das interações em ambientes virtuais de aprendizagem: uma possibilidade para avaliação da aprendizagem em EAD. *Revista Novas Tecnologias na Educação – RENOTE*, Porto Alegre, v. 4, n.1, 2006.

BASTOS, H. P. P. *Presença Plus: Modelo de identificação de presença social em Ambientes Virtuais de Ensino e Aprendizagem*. 2012. Tese (Doutorado em Informática na Educação) - Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, 2012.

BASTOS, H. P. P.; BERCHT, M.; WIVES, L. K.; KAMBARA-SILVA, J.; MARTINS, Y. Text mining indicators of affect and interaction: a case study of students' postings in a blended-learning course of English for Specific Purposes. In: Kacprzyk et al. (Ed.). *Advances in Intelligent Systems and Computing*. Berlim: Springer, 2013. v. 206. p.861-872. Disponível em: <http://link.springer.com/book/10.1007/978-3-642-36981-0/page/1>. Acesso em: 11 abr. 2015.

BASTOS, H.; BERCHT, M.; WIVES, L. K. Presença Social e Pertencimento em Fóruns Educacionais: Manifestação e Percepção de Afetividade. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 22, 2011, Aracaju, SE. *Anais*... Aracaju, SE, 2011. p. 1047-1056.

BERBER-SARDINHA, T. *Linguística de Corpus*. São Paulo: Manole, 2004.

BERCHT, M. *Em direção a agentes pedagógicos com dimensões afetivas.* 2001. Tese (Doutorado em Computação) - Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, 2001.

CHEN, G.; CHIU, M. M. Online discussion processes: Effects of earlier messages' evaluations, knowledge content, social cues and personal information on later messages. *Computers & Education*, v. 50, n. 3, p. 678-692, 2008.

CORRÊA, Y.; REATEGUI, E. B.; BIAZUS, M. C. A mineração textual de práticas discursivas em um chat: uma perspectiva pedagógica em contexto de EAD. *Revista Novas Tecnologias na Educação* – RENOTE, Porto Alegre, v. 10, n. 1, 2012.

COSTA, E.; RYAN, S. J. D.; BAKER, L. A.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. In: JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA NA EDUCAÇÃO, 2012. *Anais*... 2012, p.1-29.

DAMASCENO, E. A. A dinâmica da análise lexicométrica e de conteúdo: perspectivas e aplicações ao ensino de língua materna. *Estudos Lingüísticos de São Paulo – GEL*, v.2, p. 42-51, 2007.

D'MELLO, S.; CRAIG, S. D.; WITHERSPOON, A.; McDANIEL, B.; GRAESSER, A. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, v. 18, n. 1-2, p. 45-80, 2008.

ESMIN, A. A. A.; ALONSO, L. S.; FONSECA, E. B.; COELHO, T. A.; OLIVEIRA Jr.,

R.; GIROTO, R. Giava: Ambiente inteligente de acompanhamento e gestão de qualidades em AVA. In: ENCONTRO DE SOFTWARE LIVRE NA EDUCAÇÃO – ESLE, 2010, João Pessoa, PB. *Anais*... João Pessoa, PB, 2010.

FAN, W.; WALLACE, L.; RICH, S.; ZHANG, Z. Tapping the power of text mining. *Communications of ACM*, Nova York, v. 9, n. 49, p. 76-82, 2006.

FELDMAN, R.; SANGER, J. *The Text Mining Handbook*: *Advanced Approaches in Analyzing Unstructured Data*. Cambridge, MA: Cambridge University Press, 2007.

FREIRE, P. *Pedagogia da autonomia*: *saberes necessários à prática educativa.* 35. ed. Rio de Janeiro: Paz e Terra, 2007.

GARCIA, E.; ROMERO, C.; VENTURA, S.; de CASTRO, C. A collaborative educational association rule mining tool. *Internet and Higher Education*, v. 14, n. 2, p. 77-88, 2011.

GUPTA, V.; LEHAL, G. S. A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, v. 1, n. 1, 2009.

KAMBARA-SILVA, J. K. *Automatização do processo de aquisição de Presença Social em fóruns e chats*. 2011. Trabalho de Conclusão de Curso (Graduação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, 2011.

KAPLAN, R. A method for tokenizing text. In: ARPPE, A. et al. (Ed.) *Inquiries into Words, Constraints and Context*. CSLI Studies in Computational Linguistics. [S.l.: S.n.], 2005. p. 55-64. Disponível em: <http://cslipublications.stanford.edu/site /SCLO.html>. Acesso em: 11 set. 2011.

KLEMANN, M.; LORENZATTI, A.; REATEGUI, E. O Emprego da Ferramenta de Mineração de Textos SOBEK como Apoio à Produção Textual. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2009, Florianópolis. *Anais*... Florianópolis, SC, 2009.

LI, Y; HUANG, R. Analyzing peer interactions in computer-supported collaborative learning: model, method and tool. *Hybrid Learning and Education – Lecture Notes in Computer Science*, v. 5169, p. 125-136, 2008.

LIN, F. R.; HSIEH, L. S.; CHUANG, F. T. Discovering genres of online discussion

thread via text mining. *Computers & Education*, v. 54, n. 2, p. 481-495, 2009.
LONGHI, M. T.; BEHAR, P. A.; BERCHT, M. AnimA-K: recognizing student's mood during the learning process. In: IFIP WORLD CONFERENCE ON COMPUTERS IN EDUCATION – WCCE, 9, 2009, Bento Gonçalves, RS. *Proceedings ...* Bento Gonçalves, RS, 2009.

LONGHI, M. T.; SIMONATO, G.; BEHAR, P. A.; BERCHT, M. Um framework para tratamento do léxico afetivo a partir de textos disponibilizados em um ambiente virtual de aprendizagem. *Revista Novas Tecnologias na Educação* – RENOTE, Porto Alegre, RS, v. 8, n. 2, 2010.

LOPES, L.; VIEIRA, R.; FINATTO, M. J.; MARTINS, D.; ZANETTE, A.; RIBEIRO Jr, L. C. Extração automática de termos compostos para construção de ontologias: um experimento na área de saúde. *Revista Eletrônica de Comunicação Informação & Inovação em Saúde*, Rio de Janeiro, v. 3, n. 1, p. 76-88, 2009.

LORENZATTI, A. *Uma ferramenta de mineração de texto para um editor de texto coletivo*. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade de Caxias do Sul, Caxias do Sul, RS. 2007.

LOWENTHAL, P. The evolution and influence of social presence theory on online learning. In: KIDD, T. T. (Ed.). *Online education and adult learning*: new frontiers for teaching practices. Hershey, PA: IGI GLOBAL, 2009.

MACEDO A. L.; AZEVEDO, B. T.; BEHAR, P. A.; REATEGUI. E. Acompanhamento da interação e produção textual coletiva através de mineração de textos. *Informática na Educação: teoria e prática*, Porto Alegre, RS, v. 14, n. 2. 2011.

MACEDO, A. L. *Rede de Conceitos: uma ferramenta para contribuir com a prática pedagógica no acompanhamento da produção textual coletiva*. Dissertação (Mestrado em Educação) - Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, 2010.

MOLLER, L. Designing communities of learners for asynchronous distance education. *Educational Technology, Research and Development*, v. 46, n. 4, 1998.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. *Mineração de Textos*. Relatório Técnico. Universidade Federal de Goiás. 2007. Disponível em: <http://www.inf.ufg.br/sites/ default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>. Acesso em: 11 abr. 2013.

OLIVEIRA Jr., R.; ESMIN, A. Monitoramento automático de fóruns de discussão usando técnica de classificação de texto semi-supervisionado. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 23., 2012, Rio de Janeiro. *Anais*...Rio de Janeiro, RJ, 2012.

OTHERO, G. A.; MENUZZI, S. M. *Linguística Computacional: teoria e prática*. São Paulo: Parábola, 2005.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, v. 2, n. 1-2, p. 1-135, 2008.

PASQUALOTTI, P. R. *Reconhecimento de expressões de emoções na interação mediada por computador*. Dissertação (Mestrado em Computação Aplicada). Universidade do Vale dos Sinos - UNISINOS. São Leopoldo, RS, 2008.

PIAGET, J. *Epistemologia Genética*. 3. ed. São Paulo: Martins Fontes, 2007.

RAVI, S.; KIM, J. Profiling student interactions in threaded discussions with Speech Act classifiers. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE IN EDUCATION, 2007, Los Angeles, CA. *Proceedings…* Los Angeles, CA, 2007. p. 357-364.

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state-of-the-art. *IEEE Transactions on Systems, Man, and Cybernectics – Part C*: Applications and reviews, v. 40, n. 6, 2010.

ROMERO, C.; VENTURA, S.; HERVÁS, C.; GONZALES, P. Data mining algorithms to classify students. In: INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING – EDM'08, 1, 2008, Montreal. *Proceedings...* Montreal, Canada, 2008. p. 8-17.

SÁNCHEZ, L.P. El foro virtual como espacio educativo: propuestas didácticas para su uso. *Revista Quaderns Digitals*. n. 40, 2005.

SEARLE, J. R. *Os actos de fala*. Coimbra: Almedina, 1981.

VIGOTSKI, L. S. *Linguagem e pensamento*. 2. ed. São Paulo: Martins Fontes, 1998.

WALTHER, J. B. Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. *Communication Research,* v. 23, n.1, p. 3-43, 1996.

WIVES, L. K.; LOH, S. Tecnologias de descoberta de conhecimento em informações textuais (ênfase em agrupamento de informações). In: *OFICINA DE INTELIGÊNCIA ARTIFICIAL* (OIA), 1999, Pelotas, RS. *Anais*... 1999, Pelotas, RS. p. 28-48.

ZANK, C. *Editor de texto coletivo (ETC*): contribuições para o desenvolvimento da competência para o trabalho em equipe. 2010. Dissertação (Mestrado em Educação) - Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, 2010.