Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática
Programa de Pós-Graduação em Informática Aplicada

Máverick André Dionísio Ferreira

UMA ABORDAGEM PARA EXTRAÇÃO AUTOMÁTICA DE *LEARNING ANALYTICS* RELACIONADAS À COLABORAÇÃO EM FÓRUNS EDUCACIONAIS

Dissertação de Mestrado

Recife
Fevereiro de 2018

Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática
Pós-graduação em Informática Aplicada

Máverick André Dionísio Ferreira

UMA ABORDAGEM PARA EXTRAÇÃO AUTOMÁTICA DE *LEARNING ANALYTICS* RELACIONADAS À COLABORAÇÃO EM FÓRUNS EDUCACIONAIS

Trabalho apresentado ao Programa de Pós-graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco como requisito parcial para obtenção do grau de Mestre em Informática Aplicada.

Orientador: *Prof. Dr. Rafael Ferreira Leite de Mello*

Recife
Fevereiro de 2018

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

F383a Ferreira, Máverick André Dionísio
Uma abordagem para extração automática de learning analytics relacionadas à colaboração em fóruns educacionais / Máverick André Dionísio Ferreira. – 2018.
79 f. : il.

Orientador: Rafael Ferreira Leite de Mello.
Dissertação (Mestrado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, BR-PE, 2018.
Inclui referências e apêndice(s).

1 Fóruns de discussão 2. Learning analytics 3. Colaboração 4. Mineração de texto I. Mello, Rafael Ferreira Leite de, orient. II. Título

CDD 004

Dissertação de Mestrado apresentada por **Máverick André Dionísio Ferreira** ao programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, sob o título **Uma abordagem para extração automática de *Learning Analytics* relacionadas à colaboração em fóruns educacionais**, orientada pelo **Prof. Dr. Rafael Ferreira Leite de Mello** e aprovada pela banca examinadora formada pelos professores:

Prof. Dr. Rafael Ferreira Leite de Mello
Departamento de Estatística e Informática/UFRPE

Prof. Dr. Rafael Dueire Lins
Departamento de Estatística e Informática/UFRPE

Prof. Dr. Rodrigo Lins Rodrigues
Departamento de Educação/UFRPE

Recife
Fevereiro de 2018

Dedico esta dissertação aos meus familiares, professores e amigos que contribuíram de alguma forma para o encerramento de mais este ciclo em minha formação.

Agradecimentos

A Deus, pelo dom da vida e por ser meu refúgio diante das dificuldades com as quais me deparei até aqui. Aos meus familiares pelo apoio e por compreenderem a minha ausência, muitas vezes mental, motivada pelo sonho de aprender a ser um pesquisador. À Sebastião Júnior, que nos deixou durante esta jornada, pelo apoio e a torcida cedida ainda em vida. Ao meu orientador, professor Rafael Ferreira, por toda paciência, confiança e positividade demonstrada mesmo com todas as minhas limitações. Espero que seja apenas o início de uma parceria bastante frutífera. Aos colegas e parceiros de pesquisa Alisson, Anderson, Augusto, Débora, Elaine, Quercia, Vitor e tantos outros pela oportunidade de crescimento a mim concedida ao longo dos últimos anos.

—NÃO QUERO QUE TE APEGUES DE TAL MODO ÀS COISAS QUE DISSEMOS, QUE AS SUSTENTES TEIMOSAMENTE, SE ALGUÉM CONSEGUISSE DESTRUÍ-LAS COM ARGUMENTOS MAIS FORTES E ESTABELECER COISAS CONTRÁRIAS. SE ISSO ACONTECER, PELO MENOS NÃO NEGARÁS QUE ESTAS AFIRMAÇÕES NOS SERVIRAM DE EXERCÍCIO PARA AS DISCUSSÕES. (Santo Anselmo de Cantuária)

Resumo

O crescimento da Educação a Distância (EAD), no Brasil, tem contribuído para democratizar o acesso à educação, principalmente, dos níveis técnico e superior. Apesar disso, a distância física pode provocar nos estudantes uma sensação de isolamento que, por sua vez, consiste de uma das variáveis com poder de influenciar o aprendiz a evadir. Visando minorar o sentimento de abandono, os Ambientes Virtuais de Aprendizagem (AVA) contam com os fóruns de discussão. Estes são pontuados na literatura como altamente colaborativos, por possibilitarem a construção de conhecimento por meio de debates. No entanto, o potencial colaborativo dos fóruns pouco tem sido explorado pois a maioria das postagens inseridas nas discussões são direcionadas do estudante para o mediador. Faz-se necessário o provimento de métodos capazes dar suporte aos estudantes no desenvolvimento de habilidades colaborativas. Os moderadores precisam acompanhar todo o andamento da discussão e isso configura um trabalho intensivo a medida em que o número de postagens aumenta no decorrer do curso. É importante ressaltar que o estabelecimento de uma discussão colaborativa nos fóruns além de diminuir o surgimento da impressão de isolamento, por parte do estudante, incentiva o desenvolvimento de habilidades como o pensamento crítico/reflexivo.

Esta dissertação de mestrado apresenta uma abordagem baseada em técnicas de Mineração de Textos, Aprendizagem de Máquina e Computação Evolucionária, para extrair automaticamente *Learning Analytics* (LA) relacionadas à colaboração em postagens de fóruns educacionais conduzidos em português. A solução proposta foi fundamentada em um modelo de identificação de colaboração, proposto por MURPHY (2004), do qual foram consideradas cinco características colaborativas: Solicitar *feedback*; Responder a questões; Elogiar/expressar apreciação pelos outros participantes; Partilhar informações e recursos e; Reconhecer a presença do grupo.

Para avaliar o desempenho do enfoque dado, foram conduzidos experimentos em quatro bases de dados, compostas por postagens oriundas de fóruns educacionais, chegando a atingir *F-measure* de até 0,98. Com o objetivo de mensurar os impactos na mediação pedagógica e na colaboração dos estudantes, também foi realizado um quase-experimento em um ambiente educacional real. Os resultados mostraram indícios de que a abordagem tem potencial para fornecer o cenário colaborativo do fórum para o mediador e para os estudantes.

Palavras-chave: Fóruns de discussão, *Learning Analytics*, Colaboração, Mineração de Texto

Abstract

The growth of distance learning in Brazil has contributed to democratizing the access to education, especially in higher education. Despite of that, physical distance can cause students a feeling of isolation, which in turn consists of one of the variables that can influence the learner to evade.

To avoid such a problem, Virtual Learning Environments often encompass collaborative resources like discussion forums. The literature points that forums are highly collaborative resources because they provide a platform where the participants can debate to enrich their knowledge construction experience. However, the forums collaborative potential has not been fully exploited because most of the messages published in the discussions are from students to the instructor.

Thus, it is necessary to provide methods capable of helping students to develop the ability to collaborate. For this, the instructors need to follow up the whole progress of the discussion, and this could be an enormous work as the number of posts increases. It is important to emphasize that the establishment of collaborative discussion in the forums decrease the students‘ sense of isolation, which promote the development of skills such as critical/reflective thinking.

This dissertation presents an approach based on Text Mining, Machine Learning, and Evolutionary Computing, to automatically extract Learning Analytics related to collaboration in forums messages conducted in Portuguese. The proposed approach was based on a collaborative identification model, proposed by MURPHY (2004), of which five collaborative features were explored: soliciting feedback; Answer questions; Praise/Express appreciation by the other participants; Share information and resources and; Recognize the presence of the group.

To evaluate the performance of the approach were conducted experiments in four databases, composed of messages from educational forums. The proposed method reached F-measure of up to 0.98. In order to measure the impacts of pedagogical mediation and the collaboration of students, a quasi-experiment was carried out in a real educational environment. The results showed that the approach provided the collaborative scenario of the forum for the mediator, enabling a formative evaluation, besides contributing to the increase of the students' collaboration rates.

Keywords: Discussion Forums, Learning Analytics,Collaboration, Text Mining

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

Sumário

1

Introdução

A Educação à Distância (EAD) é uma modalidade de ensino considerada acessível a diferentes públicos por permitir flexibilidade em relação aos horários de estudo e a localização geográfica dos participantes. Os últimos censos divulgados pela Associação Brasileira de Educação a Distância (ABED) demonstram oscilações nos números de matrículas na EAD do Brasil. Por exemplo, entre os anos de 2012 e 2016 a quantidade de ingressos em cursos/disciplinas a distância foi de 5.772.466 (2012), 4.044.315 (2013), 3.868.706 (2014), 5.048.912 (2015) e 3.734.887 (2016). A oscilação observada é justificada por suposta subnotificação por parte de algumas das instituições consideradas nos levantamentos. Apesar disso, os referidos censos ressaltam o crescente interesse da classe estudantil por esta categoria de ensino (ABED, 2012, 2013, 2014, 2015, 2016). Contudo, pesquisadores apresentam estudos que demonstram a "sensação de isolamento"provocada nos estudantes, pela distância física dos demais envolvidos, como uma das variáveis com influência negativa na aprendizagem e na decisão do aprendiz de evadir (PARK; CHOI, 2009; KIM; KWON; CHO, 2011; DONDONIS DAUDT; BEHAR, 2013; BASTOS; SILVA, 2010). Ainda que os últimos levantamentos não coloquem a evasão como uma das principais preocupações, das instituições de ensino brasileiras, é unanimidade o anseio por inovações pedagógicas voltadas ao engajamento dos estudantes (ABED, 2016). Nesse contexto, os Ambientes Virtuais de Aprendizagem (AVA) são amplamente utilizados como forma de possibilitar interações entre estudantes e professores (JAIN, 2015). Podendo ser destacada a ferramenta fórum por ter como principal característica favorecer à aprendizagem colaborativa (ANDERSON; KANUKA, 1997; XIA; FIELDER; SIRAGUSA, 2013). O ambiente de aprendizado colaborativo é propício para a criação de laços afetivos e diminuição da sensação de isolamento (HUGHES et al., 2002). Além de estimular, uma das habilidades essenciais para o profissional contemporâneo, o pensamento crítico/reflexivo (PANITZ, 1999; GOKHALE, 1995).

Embora os fóruns educacionais sejam potencialmente colaborativos, as postagens inseridas nas discussões são em sua maioria direcionadas do estudante para o professor/tutor (MELO FERREIRA et al., 2013). Considerando a importância das interações do tipo estudante-estudante para a construção de conhecimento colaborativo tal como para a promoção da afetividade, (AN; SHIN; LIM, 2009) investigaram o impacto da ação docente nas interações estudantis

em discussões assíncronas. Os resultados evidenciaram um aumento nas interações mediante obrigatoriedade de resposta a postagens de colegas e da redução do número de intervenções do mediador.

Mas, apenas o aumento do número de interações não garante a colaboração (MURPHY, 2004). Para isso, é preciso desenvolver nos estudantes a habilidade de trabalhar em grupo de forma construtiva, isto é, atuar em um ambiente coletivo com ações direcionadas e coordenadas como: reconhecer a presença social e acolher as perspectivas dos demais participantes. Por isso, quando o objetivo é favorecer uma discussão colaborativa, é importante prover uma avaliação formativa (XIA; FIELDER; SIRAGUSA, 2013). Tendo o professor a função de informar ao estudante sua *performance* ao longo da discussão para, assim, possibilitar a reflexão e consequentemente a melhoria dos resultados (BLOOM et al., 1983; PERRENOUD, 2016). Estudos têm buscado propor/utilizar modelos direcionados à identificar indícios de colaboração entre aprendizes em fóruns de discussão (ROURKE, 2001; MURPHY, 2004; GUIMARAES; ESMIN, 2014). Iniciativas como essas trazem consigo grandes contribuições para a Educação, de modo geral, pois indicam horizontes de como mensurar a participação colaborativa em discussões assíncronas. Apesar disso, como tais modelos são teóricos e existe um aumento natural no número de postagens a medida em que as discussões são aprofundadas, a aplicação desses em um contexto real implica em um trabalho dispendioso para professores/tutores (DRINGUS; ELLIS, 2005; SCHEUER; MCLAREN, 2008; ALMATRAFI; JOHRI; RANGWALA, 2017).

Diante disso, surgem três possíveis problemas resultantes da adoção de modelos teóricos em cursos com uma grande quantidade de estudantes:

- O professor/tutor fica sobrecarregado podendo prejudicar sua atuação docente não apenas no fórum, mas no acompanhamento geral dos estudantes no AVA;
- A instituição opta por não utilizar os fóruns com a justificativa de não ser um ambiente produtivo;
- A instituição aumenta o investimento no curso para direcionar mais professores/tutores para acompanhar o andamento dos fóruns.

O anseio por utilizar todo o potencial colaborativo dos fóruns, sem desencadear os problemas listados anteriormente, cria uma demanda por soluções computacionais focadas em automatizar a identificação do cenário colaborativo. Para, com isso, possibilitar aos professores/tutores (mediador) a condução de avaliações formativas e aos estudantes o ajuste de ações, ao longo do processo, de modo a adquirir/aperfeiçoar a colaboração enquanto habilidade.

Nesse sentido, a área denominada de *Learning Analytics* (LA), destinada a exibir relatórios contendo análises da *performance* dos aprendizes no processo de ensino e aprendizagem, apresenta-se como relevante (LEITNER; KHALIL; EBNER, 2017). De tal forma que esta pesquisa busca por respostas para o seguinte problema: Como classificar automaticamente interações colaborativas em fóruns para promover a geração de LA?

Como a maioria das contribuições dos estudantes em fóruns são textuais, a hipótese levantada é que o desenvolvimento de uma abordagem fundamentada no modelo de identificação de colaboração proposto por MURPHY (2004) e baseada em técnicas de Mineração de Textos (MT), Aprendizagem de Máquina (AM) e Computação Evolucionária (CE), possibilita a extração automática de LA colaborativas em discussões assíncronas escritas em português.

1.1 Objetivos

1.1.1 Objetivo geral

Esta dissertação de mestrado tem como objetivo geral desenvolver uma abordagem, direcionada à extração automática de LA sobre a colaboração em fóruns educacionais, para apoiar professores no acompanhamento de fóruns e no incentivo à colaboração entre estudantes.

1.1.2 Objetivos específicos

Para atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

- Realizar uma revisão sistemática da literatura sobre as técnicas de Mineração de Textos mais utilizadas para a extração e classificação de textos em fóruns educacionais;
- Realizar uma combinação de técnicas de Mineração de Textos, Aprendizagem de Máquina e Computação Evolucionária para extrair LA relacionadas à colaboração;
- Realizar experimentos, com a abordagem proposta, em bases de dados compostas por postagens extraídas de disciplinas ministradas na EAD;
- Desenvolver um ambiente de teste para a aplicação da proposta.

1.2 Organização do trabalho

Esta dissertação encontra-se organizada em 7 capítulos. O presente capítulo de introdução apresenta o problema de pesquisa e esboça os objetivos geral e específicos. Em seguida, no capítulo 2, são elucidadas os conceitos necessários para a compreensão do trabalho: Fóruns de Discussões Educacionais, LA, MT, AM e CE.

O objeto colocado como estudo, nesta pesquisa, é a colaboração em fóruns de discussão. Dessa forma, como a maioria das contribuições dos estudantes nestas ferramentas são textuais, o capítulo 3 apresenta, por meio de uma revisão da literatura, o atual cenário sobre a aplicação de técnicas de Mineração de Textos em Fóruns Educacionais. No capítulo 4 é apresentada uma abordagem, cujo objetivo é extrair LA relacionadas à colaboração em fóruns educacionais,

além de uma ferramenta fórum criada para possibilitar a avaliação da referida abordagem com potenciais usuários.

O capítulo 5, por sua vez, descreve o processo de avaliação da abordagem em quatro bases de dados e em um ambiente real. No capítulo seguinte (capítulo 6) é feito o posicionamento da abordagem apresentada, no capítulo 4, frente aos trabalhos já realizados na mesma linha. Para isso, são listadas as principais características dos trabalhos relacionados para, em seguida, ser feita uma análise comparativa.

Por fim, no último capítulo, constam as considerações finais com enfoque nos seguintes pontos: limitações da pesquisa aqui apresentada, artigos publicados e as perspectivas de continuidade para trabalhos futuros.

2

Fundamentação Teórica

Este capítulo encontra-se dividido em duas partes. Na primeira, seções 2.1 e 2.2, são apresentadas visões gerais sobre a colaboração em fóruns de discussão e LA. No segundo momento, seções 2.3, 2.4 e 2.5, são expostos os principais conceitos das áreas consideradas no processo de extração das LA relacionadas à colaboração, a saber: MT, AM e CE.

2.1 Fóruns de discussões educacionais

Na modalidade de ensino presencial as interações entre os atores envolvidos no processo de ensino e aprendizagem ocorrem face a face. Já no ensino à distância os canais mais utilizados para a comunicação são as ferramentas integradas aos AVA, tais como: *chats*, *wikis* e fóruns (KEAR, 2007).

De acordo com ONAH; SINCLAIR; BOYATT (2014), os fóruns têm sido utilizados no âmbito educacional desde o início dos anos 90 com a promessa de possibilitar o engajamento, a motivação e a reflexão aos estudantes. Nessa ferramenta os aprendizes podem discutir, numa perspectiva de interação assíncrona, sobre assuntos curriculares ou temas pré-definidos pelos mediadores (LIN; HSIEH; CHUANG, 2009). Não sendo necessária uma interação simultânea (síncrona), podendo a discussão ocorrer com a inserção de postagens em diferentes dias e horários. Para CHENG et al. (2011), esta flexibilidade implica na oportunidade dos estudantes refletirem sobre os assuntos debatidos antes de inserirem seus pontos de vista. Além de resultar em um ambiente menos ameaçador para participantes introvertidos (CHENG et al., 2011).

O AVA Moodle[1], amplamente difundido no contexto da EAD, conta com cinco tipos de fóruns: **fórum de notícias** - criado automaticamente pelo Moodle e bastante utilizado para o compartilhamento de avisos relacionados ao curso/disciplina em questão; **fórum geral** - possibilita aos estudantes responder questões levantadas e criar novos tópicos de discussão. Por sinal, a liberdade do estudante de criar novos tópicos acarreta em uma maior dedicação do professor/tutor para manter o alinhamento do debate; **fórum simples** - permite aos estudantes debaterem sobre um único tópico colocado pelo professor; **fórum onde cada usuário inicia um**

[1] `https://docs.moodle.org/all/pt_br/Fóruns`

novo tópico - possui similaridade com o fórum geral, no entanto, propicia a cada estudante a criação de apenas um tópico de discussão e; **fórum de pergunta e resposta** - oculta as respostas de todos participantes até que o estudante insira a sua resposta.

2.1.1 Qual o papel do mediador nas discussões?

A participação ativa do professor/tutor (mediador) nos fóruns é extremamente importante. Sendo atribuído a ele o papel de incentivar os estudantes e contribuir para o aprofundamento das discussões. Para isso, segundo MAZZOLINI; MADDISON (2007), o mediador precisa inserir comentários que estimulem reflexões. Uma maneira de fazer isso é inserir postagens com visões distintas às postagens dos aprendizes (CHEN; CHIU, 2008). É também função do moderador expor, para os envolvidos, quais as expectativas e as regras vigentes durante o debate (XIA; FIELDER; SIRAGUSA, 2013).

Seguindo o mesmo pensamento, em LIM; CHEAH (2003) foram listadas funções para o moderador de acordo com cada momento da discussão: **pré-discussão** - definir o tempo de duração da discussão, os assuntos a serem discutidos e as regras de conduta; **durante a discussão** - colocar questões de sondagem e pontos de vista contraditórios para incentivar a reflexão, fornecer *feedbacks*, reconhecer as contribuições dos participantes, incentivar a inserção de postagens, resolver conflitos entre os participantes e manter os estudantes engajados e trabalhando em grupo; **pós-discussão** - tirar conclusões de todos os tópicos e solicitar que os estudantes avaliem as contribuições dos demais participantes.

2.1.2 Identificação de colaboração em fóruns de discussão

Na aprendizagem colaborativa os alunos trabalham juntos para atingir propósitos/objetivos geralmente definidos pelo docente (LAAL; LAAL, 2012; LAAL et al., 2013). Com isso, estudos têm destacado os benefícios sociais, psicológicos e acadêmicos provocados nos aprendizes em decorrência da adoção desse modelo de aprendizagem. (LAAL; GHODSI, 2012). DILLENBOURG (1999) define a aprendizagem colaborativa como situações em que duas ou mais pessoas tentam aprender algo juntas. Para ele, trabalhar em conjunto não requer a mesma localização geográfica dos envolvidos ou comunicação síncrona. Por outro lado, é preciso haver esforço e divisão sistemática do trabalho centro da colaboração.

Ao considerar o potencial colaborativo dos fóruns, MURPHY (2004) propôs um modelo teórico direcionado ao reconhecimento de colaboração em discussões assíncronas. Tal modelo foi experimentado em algumas pesquisas e tem servido como embasamento no campo que estuda a colaboração em fóruns (FERGUSON, 2009; MINHOTO; MEIRINHOS, 2012; GUIMARÃES; CAÇÃO; COUTINHO, 2013; ALRUSHIEDAT; OLFMAN, 2013). O referido autor defende que a colaboração acontece a partir de ações propositais, coordenadas e com a intenção de "produzir algo", "resolver um problema" ou "descobrir algo" em conjunto.

Como mostra a Figura 2.1, o modelo proposto por MURPHY (2004) é dividido em

Figura 2.1: Estrutura do modelo proposto por MURPHY (2004)

seis processos. Sendo esperado que o estudante expresse características colaborativas nessa ordem: (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. Para viabilizar a identificação desses processos, ao longo de debates textuais em fóruns, foram elencados indicadores como mostra a Tabela 2.1.

Processos	**Indicadores**
Presença social	Partilha de informação pessoal; Reconhecimento da presença do grupo; Elogiar/expressar apreciação pelos outros participantes; Expressão de sentimentos e emoções; Afirmação de objetivos ou propósitos; Expressão de motivação relativamente ao projeto ou à participação.
Articulação de perspectivas individuais	Afirmação de opinião pessoal ou crenças sem fazer referência às perspectivas do outro; Sumarização ou referência a conteúdo sem fazer referência às perspectivas do outro.
Acolhimento ou reflexão das perspectivas dos outros	Discordância direta ou desafio às afirmações avançadas por outro participante; Introdução de novas perspectivas; Coordenação de perspectivas.
Co-construção de perspectivas e significados partilhados	Partilha de informação e recursos; solicitação de classificação ou elaboração; Colocação de questões retóricas; Solicitação de *feedback*; Provocação de pensamento e discussão; resposta a questões; aconselhamento.
Construção de objetivos e propósitos partilhados	Proposta de objetivo ou propósito partilhado; Trabalho conjunto em direção a um objetivo partilhado.
Produção de artefatos partilhados	Documentos ou outros artefatos produzidos pelo trabalho conjunto de elementos do grupo.

Tabela 2.1: Modelo de identificação de colaboração proposto por (MURPHY, 2004)

Porém, a verificação de tais aspectos colaborativos pressupõe a busca por pistas textuais

em cada postagem inserida em um dado debate. Assim, em um cenário onde o número de postagens aumenta com o avanço da discussão, a visualização dos indicadores de colaboração pelo professor/tutor pode ser facilitada por meio de técnicas *Learning Analytics* (LA).

2.2 *Learning Analytics*

Identificar a *performance* dos estudantes no âmbito educacional possibilita o replanejamento do processo de ensino e aprendizagem considerando os sucessos e insucessos dos discentes. Isso pode ser feito, por exemplo, a partir da avaliação de notas e da percepção dos professores ao fim do semestre. Contudo, a necessidade de prover um acompanhamento contínuo acarretou no surgimento do campo de pesquisa LA (ELIAS, 2011). Sendo esse dedicado a avaliar o comportamento dos estudantes no âmbito educacional, para fornecer informações que possibilitem melhorias contínuas do processo de ensino e aprendizagem (LEITNER; KHALIL; EBNER, 2017).

Com base em LANG et al. (2017), diante de um cenário onde pesquisadores das áreas de Educação e Computação têm demonstrado crescente interesse em aplicar técnicas computacionais para extrair análises de aprendizagem de grandes massas de dados, é importante pontuar a importância de fundamentar tais iniciativas em teorias de aprendizagem como forma de dar maior relevância às análises obtidas do contexto educacional. Os autores supracitados definem basicamente três abordagens para LA: **Análise de rede** - representa os atores do contexto educacional numa rede formada por nós ligados sob diferentes perspectivas: vínculos de afiliação, amizade, interação profissional, interação comportamental ou compartilhamento de informações. **Análise dos processos** - analisa as ações dos envolvidos no processo de ensino e aprendizagem por meio de *logs* do sistema e; **Análise do conteúdo** - extrai análises de conteúdos gerados pelos aprendizes, sendo comumente utilizadas técnicas de Mineração de Textos.

Como forma de contribuir para o direcionamento das pesquisas nessa área, alguns trabalhos têm buscado evidenciar a exclusiva ligação da LA com aspectos da aprendizagem dos estudantes (GAŠEVIĆ; DAWSON; SIEMENS, 2015; SIEMENS et al., 2011). Nesse sentido, segundo SIEMENS; LONG (2011), LA consiste da medição, coleta, análise e geração de relatórios sobre a *performance* dos aprendizes, buscando o refinamento das práticas pedagógicas e dos ambientes de ensino. De forma mais específica, a etapa de coleta é responsável por extrair dados das interações dos aprendizes. A etapa de análise busca dar sentido aos dados coletados por meio da utilização de técnicas de mineração de dados, tais como: agrupamento e classificação (DYCKHOFF et al., 2012). Para, na última etapa, as informações serem exibidas aos interessados, geralmente, em *dashboards* (DYCKHOFF et al., 2012; FREITAS et al., 2017).

Dessa forma, como a maioria das contribuições dos estudantes em fóruns educacionais são textuais (FUKS et al., 2005; DIONÍSIO et al., 2017), para extrair análises das interações dos aprendizes nesses ambientes, técnicas focadas em processar textos (MT) são pontuadas como relevantes.

2.3 Mineração de textos

De acordo com FELDMAN; SANGER (2007), a Mineração de Textos (MT) é um processo de descoberta de conhecimento inspirado na área de mineração de dados, porém, com ênfase em dados textuais. Em resumo, tem como finalidade possibilitar a extração de informações significativas a partir de dados semiestruturados ou não estruturados (textos em linguagem natural) (WITTEN, 2004). As principais áreas do conhecimento que contribuem com a MT são: Aprendizagem de Máquina AM, Processamento de Linguagem Natural (PLN), Estatística, Recuperação de Informação, Ciência Cognitiva, Mineração de dados e Mineração na Web (OLIVEIRA JÚNIOR; ESMIN, 2012).

A importância da MT pode ser evidenciada pela predominância de dados organizacionais em formato de texto, aproximadamente 80%, tornando difícil a extração manual de conhecimento útil de grandes volumes de dados sem a utilização de técnicas computacionais (MORAIS; AMBRÓSIO, 2007). Existem basicamente duas abordagens, de MT, as estatísticas e as semânticas (ARANHA; PASSOS, 2006). As estatísticas utilizam, por exemplo, as frequências das palavras para determinar a importância de termos em uma coleção de textos. Por outro lado, as semânticas utilizam técnicas da área de PLN a qual busca entender como o ser humano compreende e usa determinada língua de modo a reconhecer e sintetizar a fala humana, efetuar análise léxico-morfológica e entre outros (CHOWDHURY, 2003). As técnicas de MT utilizadas na abordagem proposta são listadas a seguir:

Tokenization: identifica e separa cada unidade de um texto (palavras individuais) em *tokens* (HABERT et al., 1998). Podendo haver casos de separação **triviais** - quando as palavras estão separadas por caracteres em branco, casos **menos triviais** - quando as palavras encontram-se separadas por hífen e **casos difíceis** - quando recai sobre problemas de ambiguidade (BRANCO; SILVA, 2003).

***Term Frequency-Inverse Document Frequency* (TF-IDF)**: é uma medida destinada à identificação do grau de importância de uma palavra dentro um conjunto de documentos (SALTON; YANG, 1973). Assim, o seu cálculo consiste de três passos, como mostra as Equações 2.1, 2.2 e 2.3.

$$\text{TF} = \frac{\text{número de vezes em que o termo aparece em dado documento}}{\text{número total de termos presentes no documento}} \tag{2.1}$$

$$\text{IDF} = 1 + log_{\text{e}} \frac{\text{numero total de documentos}}{\text{quantidade de documentos que apresentam determinado termo}} \tag{2.2}$$

$$\text{TF-IDF} = TF \times IDF \tag{2.3}$$

Como pode ser visto, *Term frequency* (TF) é o número de vezes em que uma palavra

aparece dentro do universo de documentos; *Inverse document frequency* (IDF) é o número total de documentos em que uma determinada palavra aparece dividido pelo número total de documentos existentes e; TF-IDF a multiplicação do TF pelo IDF de cada palavra.

Sentence segmentation: identifica o início e fim de cada frase presente em um texto. No caso da língua portuguesa, considera como fim de uma sentença os seguintes caracteres: ponto final (.), interrogação (?), exclamação (!) e reticências (...) (SILVA, 2007). A Tabela 2.2 exibe um exemplo de segmentação e tokenização de um texto. No exemplo, as unidades textuais (*tokens*) são separadas por colchetes "[]"e o início e fim de cada sentença por parênteses "()".

Texto original	Texto segmentado em sentenças e tokenizado
Com o grande avanço tecnológico e a alta competitividade, cada vez mais são utilizados os sistemas de informação, tanto para a satisfação dos clientes como também para melhorias e inovação. Para que isso ocorra de maneira significativa é necessário uma aprendizagem organizacional e flexibilidade para mudanças	([Com], [o], [grande], [avanco], [tecnologico] [e], [a], [alta], [competitividade], [,], [cada], [vez], [mais], [sao], [utilizados], [os], [sistemas], [de], [informacao], [,], [tanto], [para], [a], [satisfacao], [dos], [clientes], [como], [tambem], [para], [melhorias], [e], [inovacao], [.]) ([Para], [que], [isso], [ocorra], [de], [maneira], [significativa], [e], [necessario], [uma], [aprendizagem], [organizacional], [e], [flexibilidade], [para], [mudanças], [.])

Tabela 2.2: Exemplo de segmentação e tokenização

POS Tagger: analisa cada palavra ou termo contido em uma sentença para, em seguida, atribuir a cada item uma classe gramatical (VIEIRA; LIMA, 2001). Os etiquetadores podem ser projetados com base em regras, em modelos probabilísticos ou em regras e modelos probabilísticos (híbridos) (SANTOS; PAIVA; BITTENCOURT, 2016). As etiquetas atribuídas aos termos de uma sentença podem variar de acordo com o componente gramatical. A Tabela 2.3 sintetiza as etiquetas disponíveis no componente gramatical CoGrOO (SILVA; FINGER, 2013).

Etiquetadas disponíveis no CoGroo
n - substantivo; **prop** - nome próprio; **art** - artigo; **pron** - pronome; **pron-pers** - pronome pessoal; **pron-det** - pronome determinativo; **pron-indp** - substantivo/pron-indp; **adj** - adjetivo; **n-adj** - substantivo/adjetivo; **v** - verbo; **v-fin** - verbo finitivo; **v-inf** - verbo infinitivo; **v-pcp** - verbo particípio; **v-ger** - verbo gerúndio; **num** - numeral; **prp** - preposição; **adj** - adjetivo; **conj** - conjunção; **conj-s** - conjunção subordinativa; **conj-c** - **conjunção coordenativa**; **intj** - interjeição; **adv** - advérbio; **xxx** - outro

Tabela 2.3: Etiquetas disponíveis no CoGroo

O CoGroo é direcionado para a análise de textos redigidos em português brasileiro, especificamente, no que se refere à identificação de erros gramaticais e atribuição de etiquetadas relacionadas à função gramatical de cada termo. A Tabela 2.4 mostra um exemplo de análise.

Remoção de *stopwords*: é um técnica responsável por remover palavras com pouca significância em um texto (stopwords), tais como: artigos, conjunções e preposições (LO; HE;

Postagem original	Postagem analisada
Os sistemas de informação são comumente utilizados em diversos seguimentos da sociedade	([Os - art], [sistemas - n], [de - prp], [informação - n], [são - v-fin], [comumente - adv], [utilizados - v-pcp], [em - prp], [diversos - pron-det], [seguimentos - n], [de - prp], [a - art], [sociedade - n])

Tabela 2.4: Exemplo de análise gramatical

OUNIS, 2005). Nesta tarefa, normalmente são utilizadas, para filtragem, listas de palavras encontradas na web[2].

Stemming: reduz as palavras aos seus respectivos radicais. Por exemplo, as palavras "engenheiro"e "engenharia"ficam "engenh". Um algoritmo bastante utilizado para o Português é o Removedor de Sufixos da Língua Portuguesa (RSLP). Este foi proposto por ORENGO; HUYCK (2001) e conta em seu funcionamento com 199 regras divididas nos seguintes passos: redução plural, redução feminina, redução de advertência, redução aumentativa/diminutiva, redução do sufixo nominal, remoção de vogais e remoção de acentos. Na Tabela 2.5 é possível visualizar um texto após a aplicação da remoção de *stopwords* e *stemming*.

Exemplo de remoção de *stopwords*	Exemplo de *stemming*
('grande', 'avanco', 'tecnologico', 'alta', 'competitividade', 'cada', 'vez', 'utilizados', 'sistemas', 'informacao', 'tanto', 'satisfacao', 'clientes', 'melhorias', 'inovacao') ('ocorra', 'maneira', 'significativa', 'necessario', 'aprendizagem', 'organizacional', 'flexibilidade', 'mudancas')	('grand', 'avanc', 'tecnolog', 'alt', 'competit', 'cad', 'vez', 'utiliz', 'sistem', 'informaca', 'tant', 'satisfaca', 'client', 'melh', 'inovaca') ('ocorr', 'man', 'signific', 'necessari', 'aprendiz', 'organizac', 'flexibil', 'mudanc')

Tabela 2.5: Exemplo de remoção de *stopwords* e *stemming*

N-grams: é uma sequência de n elementos de uma sequência textual. Na literatura é comum encontrar estudos que tratam *n-grams* a nível de caractere ou palavra (AL-SHALABI; OBEIDAT, 2008). Com base no cumprimento (tamanho do n), o *n-gram* é nomeado de *unigram* para $n = 1$, *bigram* para $n = 2$, *trigram* para $n = 3$ e assim por diante. Por exemplo, os *bigrams* da frase "fóruns de discussões educacionais na EAD"são: "fóruns de", "de discussões", "discussões educacionais", "educacionais na"e "na EAD".

2.4 Aprendizagem de máquina

Como descrito por SMOLA; VISHWANATHAN (2008), o ato de aprender consiste de um processo de aquisição de novos conhecimentos como: o desenvolvimento de habilidades motoras e cognitivas. De forma análoga, a AM é uma área da Computação destinada a estudar meios que possibilitem aos computadores aprender e, a partir de um conjunto de dados de treinamento,

[2]http://snowball.tartarus.org/algorithms/portuguese/stop.txt

estimar saídas para dados desconhecidos (NILSSON, 1996). Tal aprendizagem pode acontecer de forma supervisionada, quando se tratando de dados já rotulados, ou não-supervisionada para dados não rotulados. A seguir são apresentados os algoritmos de aprendizagem supervisionada experimentados na abordagem proposta.

Rede Neural – algoritmo baseado em um sistema nervoso que conta em seu funcionamento com a presença de neurônios artificiais interligados entre si por meio de sinapses (na computação pesos). Tais neurônios recebem entradas com pesos associados representando a força do sinal sináptico. A partir das entradas e de seus respectivos pesos, um somatório ponderado é realizado no núcleo do neurônio e com base em um limiar de ativação é verificado se a entrada será ou não propagada para neurônios das camadas adjacentes à camada atual (WANKHEDE, 2014).

Árvore de decisão – algoritmo conhecido pelos bons resultados obtidos com sua aplicação, além da fácil compreensão do processo seguido até a classificação. Na árvore de decisão cada nó interno representa um teste a ser realizado para uma das características passadas como entrada. Nesse contexto, os nós filhos do nó atual são os possíveis resultados dos testes a serem realizados e os nós folhas o resultado final. Existem diversos algoritmos disponíveis para se trabalhar com árvores de decisão, tais como: J48, ID3, C 4.5 entre outros (HAND; MANNILA; SMYTH, 2001).

Naive Bayes - algoritmo baseado no teorema de *bayes* e tem como principal característica a análise dos atributos de uma classe de forma que um atributo ignora possíveis influências/dependências sobre outro atributo no processo de inferência (RISH, 2001). Por isso, é conhecido como um classificador ingênuo, mas com vários relatos na literatura sobre sua competitividade para com outros classificadores considerados sofisticados.

Support Vector Machine - algoritmo baseado na teoria do aprendizado estatístico (VAPNIK, 2013) e que consegue lidar com problemas lineares e não-lineares. No caso de conjuntos de treinamento lineares são construídos hiperplanos a fim de conseguir separar, no plano cartesiano, os elementos a serem classificados. Um hiperplano é considerado como separação ótima se consegue separar os vetores das classes sem erro e com distância máxima para com os vetores mais próximos. Para conjuntos não-lineares são utilizadas as chamadas funções *kernels*, tais como: *Polynomial*, *Rbf* e *Sigmoid*. Essas permitem o mapeamento do conjunto de treinamento (não-linear) para um espaço dimensional maior, denominado de espaço de características, tornando possível a separação linear.

2.5 Computação evolucionária

Os algoritmos evolucionários são inspirados na teoria da seleção natural onde dado um conjunto de soluções prevalecerão as soluções mais aptas ao ambiente (problema em questão). Apesar de não garantirem soluções ótimas, tais algoritmos são considerados boas opções para reduzir o custo computacional ao mesmo tempo em que se obtêm soluções aproximadas (EIBEN;

SMITH et al., 2003). Segundo DIANATI; SONG; TREIBER (2002), existem diferentes tipos de algoritmos evolucionários, dentre eles: Algoritmo Genético (AG), Programação genética e estratégias evolucionárias. Os AG têm sido bastante utilizados para encontrar melhores conjuntos de parâmetros para algoritmos de aprendizagem de máquina e em tarefas de agrupamento de textos (GREFENSTETTE, 1986; AFONSO, 2016). O funcionamento de um AG simples consiste da criação e evolução de uma população de soluções (indivíduos) para o problema em questão (MELANIE, 1999).

1	0	0	0	0	1	0	1

Tabela 2.6: Exemplo de um cromossomo binário

Cada indivíduo é representado por um cromossomo, sendo este composto de genes binários (0s e 1s), como exemplifica a Tabela 2.6. Com o objetivo de evoluir as soluções, o AG simples funciona obedecendo os seguintes passos:

1. **População inicial** – gera aleatoriamente uma população de soluções (indivíduos) para o problema em questão;
2. **Avaliação da população** – verifica o quanto uma solução (indivíduo) se encontra adaptada ao problema;
3. **Seleção de pais** – escolhe as soluções (indivíduos) mais adaptadas ao problema para participarem do cruzamento e/ou mutação;
4. **Cruzamento** – mistura do material genético das soluções selecionadas (indivíduos pais) com o objetivo de gerar novas soluções (indivíduos filhos) capazes de explorar novos espaços no campo de busca;
5. **Mutação** – insere diversidade na população por meio da alteração (mutação) de alguns genes dada uma probabilidade definida a priori;
6. **Seleção de sobreviventes** – com a geração de novas soluções (filhos) o número de indivíduos na população duplica e por isso é preciso selecionar os indivíduos que farão parte da próxima geração e;
7. **Atualização da população** – a partir da seleção de sobreviventes atualiza a população com os indivíduos que farão parte da nova geração.

As etapas listadas são executadas até atingirem um limite de execuções pré-determinado ou um critério de parada.

3

Revisão sistemática da literatura

Uma Revisão Sistemática da Literatura (RSL) permite a identificação e posterior análise das pesquisas mais relevantes, já realizadas, em um determinado campo de pesquisa (BABAR; ZHANG, 2009). Nessa perspectiva, o crescente interesse de pesquisadores pelo uso de técnicas de Mineração de Textos em fóruns educacionais justificou a realização de uma RSL para detalhar os artigos já publicados e as oportunidades de pesquisas nesse campo. Este capítulo apresenta uma RSL conduzida seguindo o percurso metodológico proposto por KITCHENHAM (2004) que divide a RSL em planejamento, execução e síntese dos resultados, como mostra as subseções a seguir.

3.1 Planejamento da revisão

Nesta etapa, foram definidas as estratégias a serem adotadas durante a realização da RSL, cujo objetivo foi responder a seguinte questão de pesquisa: "Qual o cenário atual de pesquisas sobre a aplicação de técnicas de Mineração de Textos em Fóruns Educacionais?". Como primeira estratégia, para responder a questão de pesquisa definida, foram delineadas quatro questões específicas para serem respondidas a partir da análise das publicações existentes, no campo de Mineração de Textos em Fóruns Educacionais, são elas:

- Quais as técnicas de PLN estão sendo utilizadas?
- Quais as ferramentas, com direcionamento para Processamento de Textos ou Aprendizagem de Máquina, estão sendo utilizadas?
- Quais os AVA estão sendo utilizados?
- Quais os objetivos educacionais norteiam as pesquisas publicadas?

Foram realizadas buscas nas páginas *Google Scholar*, *ACM Digital Library* e *IEEExplore* com os termos de busca "Fórum" + "Mineração de texto" e "Forum" + "Text mining". Após o resultado inicial foram levantados os principais periódicos e conferências das áreas de "Computação e Educação", "Inteligência Artificial" e "Processamento de Linguagem Natural"

que possibilitassem responder as questões levantadas. A partir disso, foi realizada uma busca manual, em 56 bases de dados (conferências e periódicos), nacionais e internacionais, nas áreas citadas. A lista com as bases consideradas pode ser visualizada no Apêndice A.

Critérios de inclusão	Critérios de exclusão
▪ Apresentam técnicas/modelos de mineração de textos para fóruns educacionais; ▪ Avaliam técnicas/modelos de mineração de textos para fóruns educacionais; ▪ Apresentam ferramentas baseadas em mineração de textos para fóruns educacionais.	▪ Trabalhos dedicados a apresentar/experimentar técnicas/modelos de mineração de textos, para fóruns, sem objetivos educacionais; ▪ Revisões/mapeamentos da literatura; ▪ Trabalhos publicados fora do período de 2007 - 2016.

Tabela 3.1: Critérios de inclusão e exclusão

Com o objetivo de selecionar em cada base de dados os artigos relevantes ao contexto da pesquisa, assumiram-se alguns critérios de inclusão e exclusão conforme exibe a Tabela 3.1.

3.2 Fases da revisão

Alinhado aos critérios de inclusão e exclusão estabelecidos, foram definidas duas fases para seleção das pesquisas (Tabela 3.2). A fase 1 consiste da leitura do título e/ou resumo de cada artigo e posterior aplicação dos critérios de inclusão/exclusão adotados. A fase 2, por sua vez, compreende uma leitura de todo o texto dos artigos selecionados, na fase 1, além da aplicação dos critérios de inclusão/exclusão.

Fase 1	Leitura do título e/ou resumo de cada artigo e aplicação dos critérios de inclusão e exclusão.
Fase 2	Leitura de todo o texto de cada artigo selecionado na etapa1 e aplicação dos critérios de inclusão e exclusão.

Tabela 3.2: Fases adotadas para seleção dos artigos

Como forma de reduzir as chances de trabalhos relevantes serem excluídos, organizou-se a inclusão/exclusão de cada pesquisa por pares. A execução da primeira fase resultou na seleção de 55 artigos, conforme exibe a Tabela 3.3. Logo após, ao executar a fase 2, dos 55 artigos selecionados na fase 1, apenas 30 artigos foram enquadrados como relevantes em relação ao contexto desta pesquisa.

Dessa forma, o passo seguinte foi extrair de cada artigo selecionado as informações necessárias para posterior análise, como segue: a referência completa; técnicas de PLN utilizadas;

Fase 1	55
Fase 2	30

Tabela 3.3: Número de artigos selecionados por fase

AVA utilizados; ferramentas de Processamento de Textos ou Aprendizagem de Máquina utilizadas e; o objetivo da pesquisa.

3.3 Síntese dos resultados

Nesta seção são apresentados os resultados obtidos. Como já foi dito, 30 artigos foram selecionados como relevantes para esta pesquisa. A Figura 3.1 apresenta a distribuição anual das publicações selecionadas.

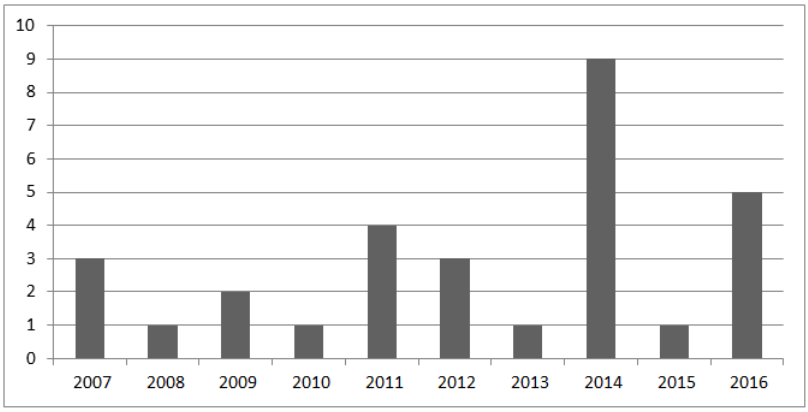


Figura 3.1: Distribuição anual das publicações selecionadas.

É possível perceber a incidência de publicações, no contexto de MT em Fóruns Educacionais, pois nos últimos 10 anos pelo menos uma pesquisa foi publicada anualmente.

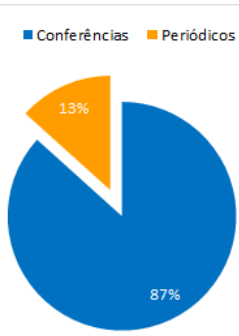
Instituições	Qt. de publicações
Universidade Federal do Rio Grande do Sul	6
Instituto Federal Fluminense	5
Universidade Federal Rural de Pernambuco	5
University of Hong Kong	2
University of Córdoba	2
University of Southern California	2

Tabela 3.4: Instituições com mais publicações

Atrelado a isso, a Tabela 3.4 mostra que apenas 6 instituições possuem duas ou mais publicações no contexto pesquisado e 16 apenas uma pesquisa cada. Sendo importante destacar a ocorrência de pesquisas realizadas envolvendo mais de uma instituição.

De acordo com a Figura 3.2, dos artigos selecionados, 4 (13%) foram extraídos de periódicos e 26 (87%) de conferências. Nesse contexto, chama a atenção o número representativo

de publicações selecionadas (8 publicações) do Congresso Brasileiro de Informática na Educação (CBIE).

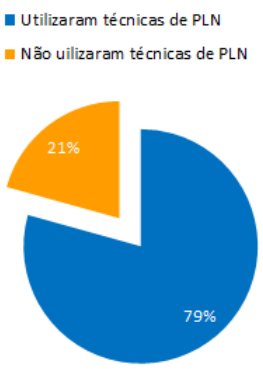


Períodos/conferências	Nº de estudos
RENOTE	2
Computers & Education	1
Artificial Intelligence Review	1
Artificial Intelligence in Education	2
IEEE International Conference on Advanced Learning Technologies	4
CBIE	8
IEEE Frontiers in Education Conference	2
BRACIS	3
International Conference on Interactive Collaborative Learning	1
International Conference on Educational Data Mining	2
IEEE Global Engineering Education Conference	1
ASEE Annual Conference and Exposition	1
International Conference on Advances in Web based Learning	1
IEEE International Conference on Data Engineering	1
Total	30

Figura 3.2: Quantidade de artigos selecionados por congresso/periódico

Outro ponto interessante é que as pesquisas encontradas foram publicadas em periódicos ou conferências nas áreas de Computação Educação e Inteligência Artificial. Esta afirmação é baseada no fato de não terem sido extraídos artigos dos periódicos e das conferências de PLN consideradas.

Em resposta à primeira questão específica de pesquisa, "Quais as técnicas de PLN estão sendo utilizadas?", a Figura 3.3 exibe um detalhamento das técnicas utilizadas nas pesquisas selecionadas. Com isso, percebe-se que seis pesquisas (21%) não utilizaram/especificaram técnicas de PLN no estudo enquanto, vinte e três (79%) relataram o uso de uma ou mais técnicas.



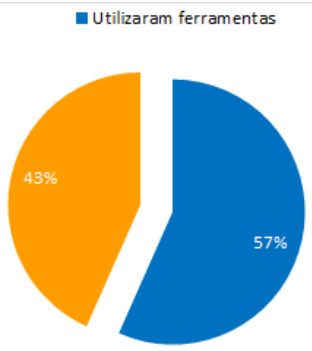
Técnicas	Nº de estudos	% de estudos
Stopwords	15	52%
TF-IDF	13	45%
Stemming	9	31%
LSA	5	17%
N-grams	4	14%
POS Tagger	3	10%
Tokenization	3	10%
Segmentation	2	7%
Lemmatization	1	3%

Figura 3.3: Técnicas de PLN utilizadas nas pesquisas selecionadas

Ainda, na Figura 3.3, é exibido um detalhamento das técnicas utilizadas, na seguinte ordem: remoção de *Stopwords* – remove palavras sem importância semântica para o texto (AZEVEDO; BEHAR; REATEGUI, 2011; LUI; LI; CHOY, 2007; OLIVEIRA JÚNIOR; ESMIN, 2012), **TF-IDF** – pondera a importância de cada palavra para um conjunto de documentos (MELO FERREIRA et al., 2013), ***Stemming*** – reduz palavras ao radical (ROLIM; FERREIRA; COSTA, 2016), ***Latent Semantic Analysis* (LSA)** – analisa as relações entre documentos de texto (YOO; KIM, 2014), **N-grams** – monta grupos de palavras de modo a possibilitar a verificação de possíveis dependências (RAVI; KIM, 2007), ***POS Tagger*** – etiqueta palavras com suas

respectivas classes gramaticais (LAU et al., 2007), ***Tokenization*** – remove caracteres especiais e divide o texto em *tokens* a partir do caractere espaço (SILVA et al., 2015), ***Segmentation*** – divide o texto seguindo sua estrutura semântica, por exemplo, palavras e orações (LIN; HSIEH; CHUANG, 2009) e ***Lemmatization*** – transfere as palavras para sua forma de dicionário (LAU et al., 2007).

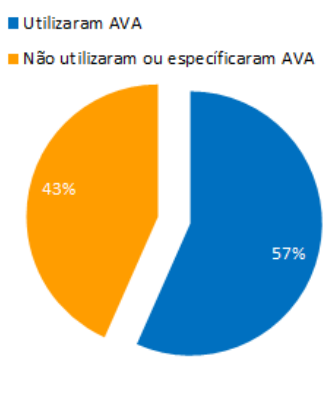
Em relação à segunda pergunta de pesquisa, "Quais as ferramentas, com direcionamento para Processamento de Textos ou aprendizagem de máquina, estão sendo utilizadas?" de acordo com a Figura 3.4, apenas dezessete pesquisas (59%) utilizaram ferramentas direcionadas para Processamento de Texto ou Aprendizagem de Máquina. Dentre as ferramentas usadas, podem ser citadas: o *Waikato Environment for Knowledge Analysis* (WEKA) – biblioteca composta por diversos classificadores de Aprendizagem de Máquina (LIN; HSIEH; CHUANG, 2009) e; o dicionário para análise de sentimentos de textos em Inglês *Linguistic Inquiry and Word Count* (LIWC) (YOO; KIM, 2014).



Ferramentas	Nº de estudos	% de estudos
WEKA	5	17%
MineraFórum	2	7%
Dbpedia Spotlight tool	1	3%
Cluto-agglomerative clustering	1	3%
LIWC	1	3%
CKIP	1	3%
RapidMiner	1	3%
SO-CAL	1	3%
HowNet	1	3%
UCINET	1	3%
SOBEK	1	3%
WordNet	1	3%
SMART Retrieval System	1	3%

Figura 3.4: Ferramentas de PLN ou AM utilizadas

A Figura 3.5, por sua vez, expõe o percentual de pesquisas que fizeram uso de Ambientes Virtuais de Aprendizagem. A propósito, é considerado como uso de AVA, nas pesquisas, os seguintes cenários: cenário 1 - aplicação de sistema/modelo de MT em um dado AVA ou; cenário 2 – extração de postagens, geradas em um dado AVA, para testar sistemas de MT.



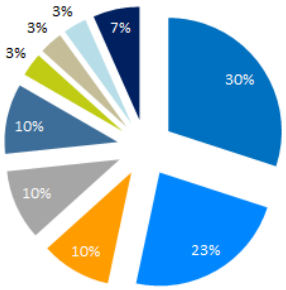
AVA	Nº de estudos	% de estudos
Moodle	12	40%
ROODA	3	10%
Fórum.Edu	1	3%
Coursera	1	3%

Figura 3.5: Ambientes Virtuais de Aprendizagem utilizados

Dessa forma, em resposta à terceira questão de pesquisa: "Quais os AVA estão sendo

utilizados?", os AVA utilizados nas pesquisas selecionadas são: o Moodle utilizado em 12 estudos (RUBIO; VILLALON, 2016; LOPEZ et al., 2012), a plataforma de Ensino a Distância da UFRGS ROODA em 3 estudos (AZEVEDO; BEHAR; REATEGUI, 2011; SILVA et al., 2015), o Fórum.Edu no estudo de (DIONÍSIO et al., 2016) e o Coursera em (WEN; YANG; ROSE, 2014).

Por fim, para responder a quarta questão de pesquisa: "Quais os objetivos educacionais norteiam as pesquisas publicadas?", na Figura 3.6 são exibidos os objetivos educacionais identificados nas pesquisas selecionadas. Como pode ser visualizado, as 30 pesquisas selecionadas foram agrupadas em 9 objetivos macro.



Objetivos das publicações	Nº de publicações
Identificar o gênero das postagens	9
Avaliar postagens automaticamente	7
Analisar o sentimento dos estudantes	3
Fornecer uma visão dos assuntos que estão sendo discutidos	3
Extrair informações das postagens para prever o desempenho dos estudantes	3
Possibilitar a visualização das postagens de acordo com o eixo temático	1
Recomendar automaticamente conteúdos externos com base nos assuntos que estão sendo discutidos	1
Fornecer respostas automáticas	1
Identificar aspectos relacionados à colaboração	2

Figura 3.6: Objetivos educacionais das pesquisas selecionadas

Em cada objetivo listado, na Figura 3.6, estão agrupadas pesquisas relacionadas. Por exemplo, o objetivo de pesquisa "Identificar o gênero das postagens" agrupa 9 estudos. A seguir são exemplificados alguns dos estudos selecionados para cada objetivo:

- **Identificar o gênero das postagens** - constam, neste objetivo, pesquisas direcionadas para a identificação de perguntas não respondidas em uma discussão online (RAVI; KIM, 2007); dedicadas para a identificação de postagens sobre dúvidas (ROLIM; FERREIRA; COSTA, 2016).

- **Avaliar postagens automaticamente** - pesquisas com foco em avaliar a relevância das contribuições do estudante em relação à discussão corrente (RUBIO; VILLALON, 2016). Analisar o sentimento dos estudantes: pesquisas direcionadas para a análise dos sentimentos dos estudantes com relação ao curso (WEN; YANG; ROSE, 2014).

- **Fornecer uma visão dos assuntos que estão sendo discutidos** - pesquisas que propõem a extração de tópicos, relevantes, para serem discutidos no decorrer do fórum (NUNES et al., 2014); sugerem a criação de mapas conceituas com o objetivo de sintetizar o andamento da discussão (LAU et al., 2007).

- **Extrair informações das postagens para prever o desempenho dos estudantes** - pesquisas voltadas para prever as notas finais dos estudantes, ou o desempenho

em projetos específicos, a partir das postagens proferidas em fóruns de discussão (LOPEZ et al., 2012; YOO; KIM, 2014);

- **Possibilitar a visualização das postagens de acordo com o eixo temático** - Estudos sobre a possibilidade de navegação temática e recomendação das postagens (LI; DONG; HUANG, 2008). Recomendar automaticamente conteúdos externos com base nos assuntos que estão sendo discutidos: pesquisas com viés para a recomendação de links relacionados ao assunto discutido (FONSECA; FARIAS; SILVA, 2014).

- **Fornecer respostas automáticas** - pesquisas sobre o uso de informações textuais, extraídas dos fóruns, para orientar agentes pedagógicos na resposta a perguntas dos estudantes (KIM; SHAW, 2014).

- **Identificar aspectos relacionados à colaboração** - pesquisas com foco na identificação de Presença Social e sensação de Pertencimento em Fóruns Educacionais (BASTOS; BERCHT; WIVES, 2011).

3.4 Discussões

Os resultados obtidos evidenciam o interesse da comunidade de Informática na Educação, nos últimos 10 anos, por pesquisas direcionadas à aplicação de técnicas de Mineração de Textos em Fóruns Educacionais. De acordo com resultados, as técnicas de PLN mais utilizadas para potencializar a mineração textual é a remoção de stopwords e o TF-IDF. Destaca-se também o uso de técnicas/ferramentas de Aprendizagem de Máquina nos estudos selecionados, tais como o WEKA.

Como ambiente para a aplicação das técnicas propostas nos estudos, o Moodle tem sido amplamente escolhido. Além disso, é possível perceber basicamente duas correntes de pesquisas na área de MT em Fóruns Educacionais. Em uma o foco é automatizar processos, de modo a facilitar a mediação pedagógica nos fóruns, é o caso de pesquisas focadas em avaliar qualitativamente postagens, extrair tópicos relevantes para serem discutidos ao longo do debate e organizar as postagens de acordo com o tema abordado. Na segunda corrente, são conduzidas pesquisas sobre como a aplicação de técnicas de MT pode auxiliar os professores/tutores em ações preventivas contra baixos desempenhos de aprendizagem e evasão, podem ser citados como exemplos os trabalhos dedicados para a previsão do desempenho e a análise de sentimentos dos estudantes.

Apesar do visível avanço neste campo de pesquisa, uma análise do atual cenário considerando as características colaborativas do fórum, mostra algumas oportunidades de estudos. Por exemplo, práticas não colaborativas como a predominância de postagens direcionadas do estudante exclusivamente para o mediador, são comuns no atual contexto da EAD, abrindo mar-

gem para a condução de investigações focadas em utilizar técnicas de Mineração de Textos para fornecer indicadores acerca dos aspectos colaborativos das discussões. Por sinal, das pesquisas selecionadas duas se posicionam nesse aspecto, porém focadas apenas na identificação de um dos aspectos colaborativos, a presença social.

Também, saber o perfil de aprendizagem dos estudantes ou o interesse por determinados conteúdos, identificar os níveis de plágio nos fóruns são algumas das oportunidades visualizadas para o emprego de técnicas de MT em Fóruns Educacionais. Tais pontos, uma vez integrados no ambiente educacional, podem subsidiar a montagem de pares de discussões ou ações corretivas/preventivas dos professores/tutores.

Portanto, os resultados obtidos serviram para evidenciar a importância de abordagens como a apresentada no próximo capítulo. Além de influenciarem na escolha das técnicas de PLN adotadas na construção da proposta.

4

Extração de LA relacionadas à colaboração

Neste capítulo é apresentada uma abordagem, baseada em técnicas de MT, AM e CE, para extrair padrões de interações colaborativas que promovam LA em fóruns educacionais conduzidos em língua portuguesa. Foi adotado o modelo teórico de reconhecimento de colaboração proposto por MURPHY (2004) o qual é detalhado no capítulo de fundamentação teórica, especificamente, na seção 2.1. Nesta pesquisa foram consideradas cinco construtos do referido modelo: (1) solicitar *feedback*; (2) responder a questões; (3) elogiar/expressar apreciação pelos outros participantes; (4) partilhar informação e recursos e; (5) reconhecer a presença do grupo.

Na abordagem proposta a extração de LA sobre a colaboração nos fóruns é tratada como um problema de classificação de textos supervisionada. Com isso, para cada construto do modelo adotado foi projetado um módulo para extrair padrões comportamentais de colaboração que promovam LA. Com exceção das características (1) e (2) que por serem consideradas complementares compõem juntas um único módulo, conforme listado a seguir:

- **Módulo 1** - Solicitar *feedback* ou responder a questões;
- **Módulo 2** - Elogiar/expressar apreciação pelos outros participantes;
- **Módulo 3** - Partilhar informações e recursos;
- **Módulo 4** - Reconhecer a presença do grupo.

Vale ressaltar que o módulo 1 utiliza em conjunto técnicas de MT, AM e CE, o módulo 2 MT e CE, enquanto os demais (3 e 4) fazem uso de técnicas de MT. Por isso, como mostra a Figura 4.1, a execução dos módulos 1 e 2 requer uma etapa inicial de treinamento como forma de criar modelos capazes de predizer futuras entradas.

Como exibe a Figura 4.1, o primeiro módulo classifica as postagens em três classes: **zero** - a postagem possui indícios de solicitação de *feedback*; **um** - a postagem é neutra ou; **dois** - a postagem possui indícios de resposta a questões. Os outros módulos consistem de classificação binária, isto é, possibilitam duas saídas: **zero** - característica ausente na postagem ou; **um** - característica presente na postagem.

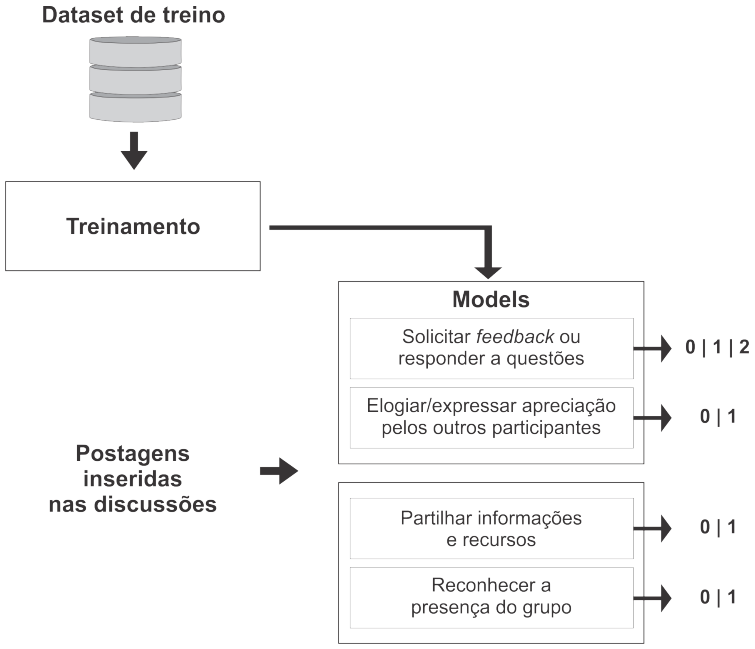


Figura 4.1: Estrutura da abordagem proposta

As próximas seções apresentam como são extraídas as LA colaborativas nos fóruns.

4.1 Módulo 1 - Solicitar *feedbacks* ou responder a questões

O modelo de MURPHY (2004) enquadra como solicitação de *feedback* postagens contendo pedido de retorno sobre alguma outra postagem inserida na discussão, conforme exemplificado na frase: "Eu queria saber se alguém tem alguma sugestão sobre isso?". E como resposta a questões as postagens com respostas a questionamentos levantados por outro participante. Dessa forma, com o objetivo de extrair LA relacionadas à solicitação de *feedback* (Dúvida) e respostas a questões em fóruns educacionais, este módulo classifica as postagens nas categorias Dúvida (D) - postagem com solicitação de *feedback*, Neutra (N) - comentário neutro e Resposta (R) - postagem com resposta a uma pergunta levantada por outro aprendiz ou pelo mediador da discussão.

Na literatura já existem trabalhos direcionados a classificar postagens nas classes mencionadas acima (Dúvida, Neutra e Resposta). Porém, o trabalho com melhor resultado (ROLIM; FERREIRA; COSTA, 2016) utiliza o número de interrogações como elemento no vetor de características das postagens classificadas. Mas, na língua portuguesa, perguntas podem ser diretas ou indiretas (sem interrogação), por isso em escala maior o referido trabalho pode apresentar problemas.

Diferentemente, este módulo busca distinguir as postagens a partir da distribuição gramatical das palavras presentes. Onde, para isso, foram utilizadas técnicas de MT no processo de extração de características das postagens, o algoritmo *Support Vector Machine* (SVM) para classificação e um AG para encontrar o melhor conjunto de parâmetros para o SVM. Assim,

divide-se a execução em duas etapas: extração de características e classificação das postagens. Na etapa de extração de características são retiradas das postagens as informações necessárias para o processo de classificação. Com isso, inicialmente é utilizado o componente gramatical CoGroo (SILVA; FINGER, 2013) para etiquetar todas as palavras das *n* postagens em análise de acordo com as suas respectivas funções gramaticais.

Na sequência são criados os vetores de palavras das classes Pergunta, Neutra e Resposta considerando apenas as palavras com as seguintes funções gramaticais: Verbo, Substantivo, Pronome, Adjetivo e Advérbio. Desse modo, as palavras são retiradas das postagens e coladas dentro do vetor referente à sua classe de postagem (Dúvida, Neutra ou Resposta) e no subconjunto da função gramatical que desempenham (Verbo, Substantivo, Pronome, Adjetivo ou Advérbio) como demonstrado na Figura 4.2.

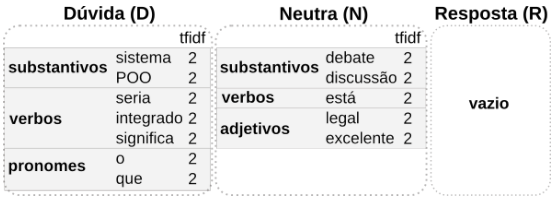


Figura 4.2: Exemplo criação dos sacos de palavras.

Após a distribuição das palavras nos vetores estas são ordenadas decrescentemente de modo a posicionar as palavras com maior representatividade, para cada classe gramatical, no início dos vetores. O processo de ordenamento considera o TF-IDF de cada palavra o qual é calculado como segue: TF – número de vezes que uma palavra aparece desempenhando a mesma função gramatical dentro do universo de postagens da classe a qual pertence (Dúvida, Neutra ou Resposta); IDF – número de vezes que uma palavra aparece desempenhando a mesma função gramatical, dentro do universo de postagens da classe a qual pertence (Dúvida, Neutra ou Resposta), dividido pelo número total de postagens que possuem palavras com a mesma função gramatical e classe de postagem (Dúvida, Neutra ou Resposta) da palavra em questão.

Por fim são gerados vetores de características para cada postagem com 15 posições (Figura 4.3).

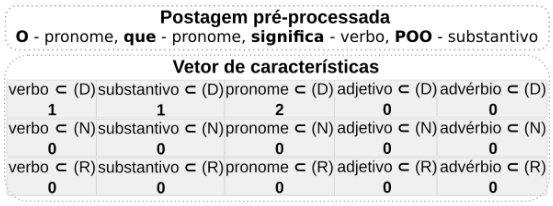


Figura 4.3: Exemplo de um vetor de características

Cada posição do vetor recebe o número de palavras, com mesma função gramatical, que constam na postagem em questão e também em um dos vetores de palavras (Dúvida, Neutra e Resposta). Por exemplo, a postagem utilizada na Figura 4.3 possui um verbo e um pronome coincidentes com palavras de mesmas funções gramaticais presentes no saco de palavras de

Dúvidas. Após a preparação dos vetores de características, a classificação das postagens é realizada por meio do classificador SVM. Este foi escolhido por obter bons resultados no domínio de classificação de textos (JOACHIMS, 2002; FERREIRA et al., 2016).

Figura 4.4: 2ª etapa do módulo – classificação das postagens.

Como exibido na Figura 4.4, o SVM recebe os vetores de características e realiza o treinamento utilizando cada conjunto de parâmetros gerado pelo AG até que uma condição de parada seja atingida. A propósito, o AG tem a sua população inicial representada por soluções (indivíduos) cujo material genético é constituído por oito genes que se referem aos parâmetros do SVM sintetizados na Tabela 4.1.

Mutação uniforme			**Mutação Gaussiana**				
Svm_type	*Probability*	*Kernel_type*	*Gamma*	*nu*	*Cache_size*	*Cost*	*Épsilon*

Tabela 4.1: Organização dos operadores de mutação para cada gene

O ciclo de execução do AG é formado pelas seguintes fases: (i) **geração da população inicial** – inicializa o algoritmo com indivíduos (soluções) com parâmetros (genes) aleatórios; (ii) **avaliação dos indivíduos** – executa um SVM com o conjunto de parâmetros de cada indivíduo da população.

Como o objetivo de cada indivíduo é parametrizar o SVM de modo a classificar corretamente as postagens, das classes Dúvida, Neutra e Resposta, a cada execução é calculada a *F-measure* que é utilizada como aptidão para cada indivíduo. Para realizar a otimização os indivíduos passam por processos de atualização, para isso foram utilizadas as seguintes etapas:

- **seleção de pais** – escolhe indivíduos para participarem do cruzamento, sendo o método de seleção adotado a roleta por possibilitar aos indivíduos com baixa aptidão oportunidade de reprodução;
- **cruzamento** – mistura do material genético dos indivíduos selecionados (pais) com o objetivo de gerar novas soluções (filhos) capazes de explorar novos espaços no campo de busca, para tal é usado o cruzamento uniforme (MELANIE, 1999);
- **mutação** – insere diversidade na população por meio da alteração (mutação) de alguns genes dada uma probabilidade definida a *priori*. Para isso, são utilizados dois operadores conforme exemplificado na Tabela 4.1;

- **mutação uniforme** – por possibilitar a escolha de dois genes aleatoriamente para posterior troca de posição;

- **mutação *gaussiana*** – tendo em vista a ampla utilizando em problemas contínuos onde é gerado um valor a partir de uma distribuição normal (0, α), sendo o desvio padrão 1 (um), para em seguida somar o valor gerado pela distribuição normal com o valor contido em um dos genes do indivíduo escolhido para mutação;

- **Seleção de sobreviventes** – com a geração de novas soluções (filhos) o número de indivíduos na população duplica e por isso é preciso selecionar os indivíduos que farão parte da próxima geração. Nesse aspecto utiliza-se o método *steady state* que consiste na seleção das melhores soluções dentro de um universo constituído pela população anterior e a população atual (MELANIE, 1999);

- **atualização da população** – atualiza a população de indivíduos, deixando as melhores soluções, com base na seleção realizada anteriormente.

4.2 Módulo 2 - Elogiar/Expressar apreciação pelos outros participantes

Este módulo busca extrair análises sobre a inserção de elogios/apreciações por parte de estudantes com direção a outros participantes em discussões textuais. Para isso, cada postagem é classificada como **True** (elogio/apreciação presente na postagem) ou **False** (elogio/apreciação ausente na postagem). Em resumo, o funcionamento deste módulo foi organizado em duas atividades: composição do vetor de características e classificação das postagens. Inicialmente as postagens são seguimentadas em sentenças (*sentence segmentation*) e, em seguida, a primeira sentença de cada postagem é tokenizada em unidades simples (*unigram*) a serem utilizadas na extração de informações para a montagem dos vetores de características, como descrito abaixo.

A primeira posição do vetor pode receber os valores: um – caso a postagem possua pelo menos uma palavra com polaridade positiva ou; zero – caso a postagem não possua palavra com polaridade positiva. A importância das palavras positivas na postagem é justificada por entender-se que elogios ou expressões de apreciação são manifestados por meio destas. Para verificar a polaridade das palavras presentes na postagem, optou-se por uma abordagem baseada em dicionários *léxicos*. Dessa maneira, foram escolhidos os dicionários sentiLEX e Oplexicon ambos direcionados para a análise de sentimentos em textos redigidos em língua portuguesa (CARVALHO; SILVA, 2015; SOUZA; VIEIRA, 2012). Por entender-se que um estudante pode elogiar uma ação desempenhada por outro (VERBO) ou, também, elogiar o outro participante em si (SUBSTANTIVO). Foram consideradas, dos referidos dicionários, apenas palavras etiquetadas como adjetivos e advérbios. Pois, segundo CÂMARA (1970), o adjetivo é o determinante do substantivo e o advérbio do verbo.

Mesmo com a identificação de palavras positivas, é importante também examinar a presença de palavras negativas. Isto acontece porque em uma frase “João não discutiu bem.”, onde um advérbio positivo “bem” modifica o verbo “discutiu” e um advérbio negativo “não” também modifica o mesmo verbo “discutiu”, o sentido negativo prevalece. Por isto, ainda fazendo uso dos dicionários supracitados, a segunda posição do vetor pode ser preenchida com os valores: um – caso a postagem possua pelo menos uma palavra com polaridade negativa ou; zero – caso a postagem não possua palavra com polaridade negativa.

É importante ressaltar que a associação de uma palavra positiva/negativa ao estudante somente acontece se o estudante estiver explicitamente apontado na postagem, isto é, devidamente nomeado. Nesse sentido, a terceira posição do vetor recebe os valores: um – caso a postagem possua nome próprio ou; zero – caso a postagem não possua nome próprio. Para fins de constatação é utilização o dicionário *léxico* NomesLEX[1] que contém 2027 nomes em português. Como o objetivo de associar possíveis palavras positivas e negativas encontradas na postagem a um nome próprio, a quarta e a quinta posição recebem respectivamente a distância entre o nome próprio encontrado e a palavras positiva e negativa mais próximas.

A análise de polaridade das palavras e a procura por nomes próprios considera apenas a primeira sentença de cada postagem. Isto ocorre devido a observações empíricas as quais evidenciaram uma maior concentração de elogios/apreciações no início das postagens. Outro fator considerado relevante para aferir se uma postagem possui ou não elogio a um estudante é o tamanho da postagem, partindo dessa premissa a sexta posição armazena o tamanho da postagem em análise. Resumindo, as posições do vetor de característica são:

1. **Posição 1** - Presença de palavras com polaridade positiva;
2. **Posição 2** - Presença de palavras com polaridade negativa;
3. **Posição 3** - Existência de nome próprio;
4. **Posição 4** - Distância de um elogio/apreciação para um nome próprio;
5. **Posição 5** - Distância de uma palavra negativa para o elogio/apreciação identificado;
6. **Posição 6** - Tamanho da postagem.

A Tabela 4.2 mostra o vetor de características para a postagem: "Muito legal João, encontrei outros dois comandos usados para controle de fluxo no java, que são: do...while, while e o switch.".

Posição 1	Posição 2	Posição 3	Posição 4	Posição 5	Posição 6
1	1	1	1	7	128

Tabela 4.2: Exemplo de um vetor de características

[1] http://xldb.fc.ul.pt/wiki/NomesLex-PT01

Na postagem usada para exemplificar, a palavra com polaridade positiva encontrada foi "legal", estando a referida a uma posição do nome próprio, enquanto com polaridade negativa foram identificadas "usados"e "controle". Após a montagem do vetor de características este é usado como entrada de um algoritmo de AM responsável por realizar a classificação, isto é, predizer se a postagem em análise contém ou não elogio/expressão de apreciação de um estudante para com outros.

4.3 Módulo 3 - Partilhar informações e recursos

Este módulo compreende partilhamento de informações e recursos como o ato de estudantes compartilharem links *Uniform Resource Locator (URL)*, para conteúdos externos, em fóruns educacionais. Dessa forma, busca identificar a presença de *URL* em postagens para, em seguida, analisar se as páginas web referenciadas possuem conteúdos concernentes aos discutidos no fórum em questão.

Para atingir tal objetivo, é criado um conjunto com as principais palavras-chave do conteúdo compartilhado pelo estudante (CC) e outro com palavras representativas da discussão conduzida no fórum analisado (DF). Isto acontece para possibilitar a identificação de possíveis relações, entre o conteúdo compartilhado e o assunto debatido, por meio da comparação das palavras presentes em ambos os conjuntos (CC e DF). A Figura 4.5 sintetiza as etapas adotadas para a criação do conjunto CC.

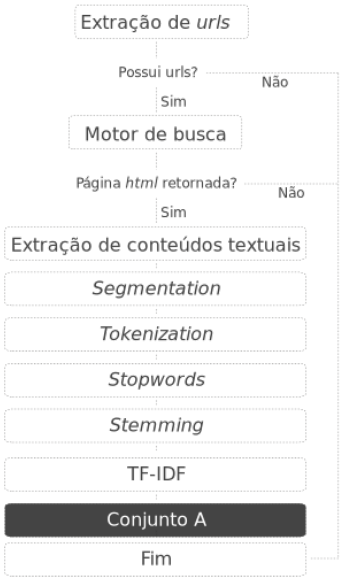


Figura 4.5: Etapas para criar o conjunto

Como pode ser observado, o processo de criação do conjunto CC consiste inicialmente em verificar a existência de *URL*, na postagem em análise *p*, por meio do método *ExtractUrls(p)* que utiliza a biblioteca *URLEXtract* [2] e ao receber como entrada uma postagem *p* retorna uma

[2]https://pypi.python.org/pypi/urlextract

lista vazia, caso não sejam encontradas *URLs*, ou com as *URLs* encontradas.

Quando retornadas *URLs*, estas são processadas utilizando o método, *ContentSearch(url)*, responsável por filtrar os conteúdos da página endereçada pela *URL* recebida como parâmetro. Sendo filtradas nesse processo as informações textuais presentes em todas as *tags* HTML. É importante também ressaltar que a cada execução deste procedimento é retornado um texto com todo o conteúdo extraído *c* a partir da *url* processada.

Em seguida, a texto obtido *c* é dado como entrada no método *RepresentativeWords(c)*. Este, por sua vez, efetua algumas operações afim de gerar, como saída, um conjunto com as principais palavras do texto recebido *c*. A primeira operação consiste em remover os acentos e caracteres especiais. Depois é executada uma segmentação em sentenças considerando como início e fim de cada sentença os sinais de pontuação: interrogação (?), exclamação (!) e ponto final (.) quando seguidos de letra maiúscula.

Posteriormente são removidas, de todas as sentenças, as palavras que não agregam significado para o texto (*stopwords*). Além disso, considerando a possibilidade de palavras com o mesmo significado serem tratadas de formas distintas por estarem flexionadas como, por exemplo, as palavras "aprendiz"e "aprendizes", é aplicado o algoritmo RSLP direcionado para redução das palavras aos seus respectivos radicais (*stemming*) (ORENGO; HUYCK, 2001).

A última operação do *RepresentativeWords(c)* consiste de utilizar a medida TF-IDF no conjunto de sentenças tokenizadas. Isto ocorre da seguinte maneira: TF – número de vezes que uma palavra aparece dentro de um conjunto de sentenças e; IDF – número de sentenças que uma palavra aparece dividido pelo número total de sentenças. Dessa forma, todas palavras são ranqueadas em ordem decrescente com base no respectivo TF-IDF. Finalmente, o resultado a chamada do *RepresentativeWords(c)* é 10% das palavras com melhor ranqueamento, sendo este usado para compor o conjunto CC.

Para a criação do conjunto DF todas as postagens, anteriores à postagem em análise, são agrupadas em um só texto o qual passa por todos os procedimentos explicados anteriormente e listados na Figura 4.5. A partir disso, é criado um segundo conjunto com 10% das palavras representativas do fórum em questão. Com isso, é verificada a interseção entre ambos os conjuntos. Caso o conjunto resultante não seja vazio, é identificado o Partilhamento de Informações relacionadas ao conteúdo proposto para discussão no fórum.

4.4 Módulo 4 - Reconhecer a presença de grupo

Este módulo considera o ato de um estudante dirigir-se a outros estudantes, por meio de suas postagens, como um forte indício de que o mesmo reconhece estar inserido em um grupo. Assim sendo, objetiva identificar o reconhecimento de Presença de Grupo utilizando quatro regras elaboradas em conjunto com uma professora da Universidade Aberta do Brasil (UAB).

Conforme ilustra a Figura 4.6, ao receber como entrada uma postagem, inicialmente é executada uma análise morfológica utilizando a ferramenta CoGroo (SILVA; FINGER, 2013). O

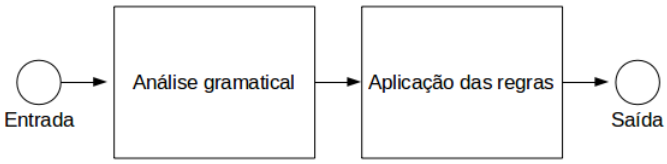


Figura 4.6: Estrutura do módulo de Reconhecimento de Presença Grupo

resultado deste procedimento é uma postagem tokenizada e com funções gramaticais associadas a cada *token*. Após isso, a postagem é examinada sob a ótica das regras criadas:

- **Regra 1** - verifica a existência de palavras etiquetadas como interjeição nas três primeiras posições da postagem em análise;
- **Regra 2** - procura cumprimentos no início da postagem, sendo consideradas as expressões “bom dia”, “boa tarde” e “boa noite”. A estratégia de buscar interjeições e cumprimentos no início das postagens é justificada por estas serem expressões costumeiramente introduzidas no começo da fala;
- **Regra 3** - busca localizar na postagem em análise: pronomes na primeira pessoa do plural. Mais especificamente, pronome pessoal caso reto (nós), pronome pessoal oblíquo (conosco) e pronomes pessoais de tratamento (vocês e senhores);
- **Regra 4** - procura por substantivos relacionados ao coletivo de estudantes. Para isso, cria um conjunto A “{colegas, estudantes, amigos, companheiros, galera}” com os substantivos comumente utilizados por estudantes para fazer referência aos demais participantes em uma discussão. Um conjunto B “substantivos presentes na postagem em análise”. Por fim, verifica a intersecção entre os conjuntos A e B. Ao término desta operação, se a intersecção entre A e B resultar em um conjunto com um ou mais elementos, a postagem analisada possui palavras com referência ao coletivo de estudantes. Em caso de intersecção vazia, entre A e B, o passo seguinte é montar um conjunto C contendo sinônimos de cada substantivo presente na postagem em questão. A preparação do conjunto de sinônimos conta com um *léxico* para Português, o WordNet.PT[3]. Em seguida, verifica-se a intersecção entre os conjuntos C e A e caso obtenha-se como resultante um conjunto com um ou mais elementos, a postagem analisada possui palavras que representam o coletivo de estudantes. Na tentativa de verificar se os substantivos encontrados referem-se ao grupo de estudantes do fórum em questão ou fazem menção a externos, é analisada a existência de pronomes demonstrativos na posição anterior ao substantivo encontrado, por exemplo: "estes companheiros são interessantes", "aqueles amigos buscam contribuir"e entre outros.

[3]https://github.com/own-pt/openWordnet-PT

Se, pelo menos, uma das regras listadas for atendida é identificada a Presença de Grupo na postagem analisada.

5

Avaliação

A presente pesquisa aplica conhecimentos advindos de áreas da Computação (Mineração de Textos, Aprendizagem de Máquina e Computação Evolucionária) em um contexto educacional. Mais especificamente, com o objetivo de extrair LA relacionadas à colaboração entre estudantes em fóruns educacionais. Dessa forma, faz-se necessária a avaliação da abordagem proposta sob a ótica da Computação e da Educação. Por este motivo, o processo de avaliação foi dividido em duas etapas. Na primeira, foram realizados experimentos em bases de dados oriundas de fóruns educacionais e o desempenho analisado a partir de métricas computacionais. De forma complementar, no segundo momento, a abordagem foi experimentada em um ambiente real a fim de verificar seus impactos para a mediação pedagógica e colaboração entre os discentes.

5.1 Experimentos nas bases de dados

Nesta seção são apresentados os resultados obtidos com a realização dos experimentos nas bases de dados consideradas. Para facilitar a compreensão, nas subseções abaixo são detalhadas as bases de dados e as métricas utilizadas para medir o desempenho de cada módulo da abordagem. Em seguida são expostos os resultados, sendo importante ressaltar que por já haverem diversos trabalhos publicados com o mesmo enfoque do primeiro módulo (Solicitar de *feedback* ou responder a questões) este foi avaliado de forma mais detalhada e comparado com o estudo que, até então, apresentava o melhor resultado na literatura.

5.1.1 Bases de dados

Foram utilizadas, nos experimentos, um total de quatro bases de dados compostas por postagens extraídas dos cursos a distância de Ciência da Computação e Sistemas da Informação da Universidade Federal de Alagoas (UFAL). Como o objeto em estudo é a interação entre estudantes, as bases foram alimentadas apenas com postagens inseridas por estudantes em fóruns do tipo pergunta e resposta (ver Seção 2.1). O fórum que originou a primeira base contava com estudantes do primeiro período de graduação, enquanto, os demais com aprendizes do segundo período. As bases são referenciadas ao longo do texto como BD1, BD2, BD3 e BD4.

Os BD1 e BD2, com 490 e 600 postagens, são os mesmos utilizados em um trabalho relacionado ao módulo 1 (Solicitação de *feedback* ou resposta a questões) possibilitando a posterior comparação de resultados (ROLIM; FERREIRA; COSTA, 2016). O BD3 conta com 679 postagens das disciplinas de Análise de dados, Sistemas de Informação, Internet e Web e Programação Orientada a Objetos. O BD4, por sua vez, comporta um total de 599 postagens obtidas de um conjunto de disciplinas equivalente ao conjunto que originou o BD3.

Para possibilitar a realização dos experimentos, as bases de dados foram avaliadas e anotadas de acordo com os aspectos colaborativos considerados na abordagem proposta. Nesse sentido, como BD1 e o BD2 foram utilizados no estudo de (ROLIM; FERREIRA; COSTA (2016)) não foi necessário anotá-los por serem etiquetados como Dúvida, Neutra e Resposta, anotação esta compatível com os experimentos a serem realizados no módulo de Solicitação de *feedback* ou resposta a questões. O BD3 foi anotado para ser utilizado nos experimentos do módulo 1, conforme a distribuição das postagens exibida na Tabela 5.1.

Solicitação de *feedback* ou resposta a questões			
Bases de dados	**Dúvida**	**Neutra**	**Resposta**
BD1	198	104	188
BD2	200	200	200
BD3	60	96	523

Tabela 5.1: Distribuição das postagens dos BD 1, 2 e 3

Por outro lado, o BD4 foi anotado de modo a oportunizar sua utilização nos experimentos dos demais módulos: Elogiar/expressar apreciação pelos outros participantes, Partilha de Informações e Recursos e Reconhecimento de Presença de Grupo. A Tabela 5.2 mostra como ficou a organização das postagens.

BD4		
Característica colaborativa	**Não identificada**	**identificada**
Elogiar/expressar apreciação pelos outros participantes	516	83
Partilha de Informações e Recursos	583	16
Reconhecimento de Presença de Grupo	580	19

Tabela 5.2: Distribuição das postagens do BD4

A atividade de anotação das bases contou com quatro avaliadores externos alocados da seguinte maneira: dois avaliadores para o BD3 e dois para o BD4. Como forma de minimizar eventuais inconsistências, foi solicitada a atuação de um terceiro nos momentos de divergência de avaliação, isto é, quando houve diferentes percepções para uma mesma postagem. Na tentativa de padronizar as avaliações, cada avaliador recebeu uma planilha contendo as postagens e os aspectos a serem analisados. Além disso, também foi disponibilizado um documento, em formato digital, com o detalhamento do modelo de identificação de colaboração adotado.

5.1.2 Métricas de avaliação

A abordagem proposta concentra-se em inferir se postagens de fóruns educacionais possuem ou não determinados aspectos colaborativos. Isto implica em uma tarefa de classificação a qual, em áreas como classificação de textos, é comumente avaliada por meio de medidas como *recall*, *precision* e *F-measure* (Equações 5.1, 5.2 e 5.3) (FERRI; HERNÁNDEZ-ORALLO; MODROIU, 2009).

$$precision = \frac{VP}{VP + FP} \tag{5.1}$$

$$recall = \frac{VP}{VP + FN} \tag{5.2}$$

$$F\text{-}measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{5.3}$$

Para calcular a *precision* e *recall* é utilizada como auxílio a chamada de matriz de confusão conhecida por tabelar os resultados obtidos da seguinte maneira: Verdadeiro positivo (VP) – número de elementos positivos classificados como positivos; Verdadeiro negativo (VN) – número de elementos positivos classificados como falsos; Falso positivo (FP) – número de elementos falsos classificados como positivos e; Falso negativo (FN) – número de elementos falsos classificados como falsos. Assim, a *F-measure* é definida como uma medida ponderada da *precision* e *recall* (ZHANG; ZHANG, 2009).

5.1.3 Solicitar *feedback* ou responder a questões

A estrutura apresentada para o módulo 1 (seção 4.1) implica em alguns questionamentos: considerar as funções gramaticais na atividade de montar os sacos de palavras (Dúvida, Neutra e Resposta) e os vetores de características das *n* postagens melhora o desempenho do módulo 1? O uso do AG para ajustar os parâmetros do SVM implica em um melhor índice de classificação das postagens? O fato do módulo 1 não considerar caracteres como interrogações reduz seu desempenho em comparação a trabalhos que utilizam? Com base nos questionamentos listados, na etapa de levantamento de hipóteses foram elaboradas 5 hipóteses (Tabela 5.3). Tais hipóteses foram pareadas (nula H_0 e alternativa H_a) de tal forma que se uma for considerada falsa a outra será verdadeira.

Como forma de verificar as hipóteses listadas na Tabela 5.3, foram criados dois cenários de execução para classificar as postagens do BD1, BD2 e BD3:

- **Cenário 1 (C1)** - classificar as postagens utilizando a configuração padrão do módulo 1 (Seção 4.2);
- **Cenário 2 (C2)** - classificar as postagens por meio do módulo 1 sem considerar as funções gramaticais na criação dos vetores de palavras e dos vetores de características.

	Hipótese nula		Hipótese alternativa
$\mathbf{H_{0a}}$	A *F-measure* obtida ao considerar as funções gramaticais é menor ou igual à *F-measure* obtida sem considerar as funções gramaticais.	$\mathbf{H_{a1}}$	A *F-measure* obtida ao considerar as funções gramaticais é maior que a *F-measure* obtida sem considerar as funções gramaticais.
$\mathbf{H_{0b}}$	A *F-measure* obtida com o uso do AG é menor ou igual à *F-measure* obtida sem o uso do AG.	$\mathbf{H_{a2}}$	A *F-measure* obtida com o uso do AG é maior que a *F-measure* obtida sem uso do AG.
$\mathbf{H_{0c}}$	A *F-measure* obtida com o uso do AG e das funções gramaticais é menor ou igual à *F-measure* obtida apenas com o uso das funções gramaticais.	$\mathbf{H_{a3}}$	A *F-measure* obtida com o uso do AG e das funções gramaticais é maior que a *F-measure* obtida apenas com o uso das funções gramaticais.
$\mathbf{H_{0d}}$	A *F-measure* obtida com o uso do AG e das funções gramaticais é menor ou igual à *F-measure* obtida apenas com o uso do AG.	$\mathbf{H_{a4}}$	A *F-measure* obtida com o uso AG e das funções gramaticais é maior que a *F-measure* obtida apenas com o uso do AG.
$\mathbf{H_{0e}}$	A *F-measure* obtida com o módulo 1 é menor ou igual à *F-measure* obtida com o algoritmo proposto em ROLIM; FERREIRA; COSTA (2016).	$\mathbf{H_{a5}}$	A *F-measure* obtida com o módulo 1 é maior que a *F-measure* obtida com o algoritmo proposto em ROLIM; FERREIRA; COSTA (2016).

Tabela 5.3: Hipóteses levantadas

Na etapa de experimentos cada cenário é executado 30 vezes com o AG ativado e 30 vezes com o AG desativado em cada configuração de corte (corte 0, corte 5 e corte 10) dos BD1, BD2 e BD3, totalizando 540 execuções para cada cenário.

O tamanho do corte refere-se a quanto em percentual se utiliza das palavras contidas nos vetores das classes Dúvida, Neutra e Resposta para montar os vetores de características. Por exemplo: corte 0 – utiliza 100% das palavras contidas nos vetores de palavras; corte 5 – utiliza 95% das palavras contidas nos vetores de palavras e; corte 10 – utiliza 90% das palavras contidas nos vetores de palavras. As execuções que utilizaram o AG fizeram uso dos seguintes ajustes: 100 gerações, 30 indivíduos por geração, taxa de cruzamento de 50% e uma probabilidade de mutação de 2% para cada gene. Os resultados alcançados são comparados com os de ROLIM; FERREIRA; COSTA (2016), até então, melhor resultado na literatura. Para isso, o referido algoritmo foi executado, com sua melhor configuração, 30 vezes em cada configuração de corte (corte 0, corte 5 e corte 10) dos BD1, BD2 e BD3 totalizando 270 execuções.

Na etapa de testes de hipóteses foi adotado um teste paramétrico levando em consideração que os elementos (resultado final de cada execução) das amostras geradas com os experimentos (ciclos de 30 execuções) são independentes (LARSON; FARBER; PATARRA, 2004). Dentro da classe de teste selecionada (testes paramétricos) encontram-se os testes *t-student* e z. Para selecionar o teste z, por exemplo, é preciso atender a alguns critérios, tais como: as amostras que participarão dos testes devem ter sido selecionadas aleatoriamente, possuir distribuição normal e desvio populacional conhecido para amostras com tamanho < 30.

Nesse caso, visualiza-se a seleção das amostras de forma aleatória, isto é, as amostras foram geradas por meio de ciclos de 30 execuções do módulo 1 não havendo nenhuma heurística para geração ou seleção dos resultados. Também possui aproximação normal, com base no teorema do limite central, por possuir amostras com tamanho 30. Por atender aos critérios o teste z foi escolhido para comparação de médias μ de duas amostras. Assim, como as hipóteses exibidas na Tabela 5.3 são representadas por H_0: $\mu_1 \ \mu_2$ e H_a: $\mu_1 > \mu_2$ os testes são unicaudais à direita.

Para montar as amostras de acordo com a necessidade de cada teste de hipótese foram criadas 5 amostras a partir das execuções dos cenários 1 e 2, são elas: amostra de execuções com o módulo do AG ativado (540 execuções); amostra de execuções com o módulo do AG desativado (540 execuções); amostra de execuções considerando as funções gramaticais e com o módulo do AG ativado (270 execuções); amostra de execuções considerando as funções gramaticais (540 execuções); amostra de execuções sem utilizar as funções gramaticais (540 execuções).

A seguir são apresentados os resultados obtidos a partir dos experimentos. Vale ressaltar que as tabelas exibem em negrito as médias das *f-mesuares* obtidas (a cada ciclo de 30 execuções) e seus respectivos desvios padrões. Por fim são realizados os testes de hipótese.

5.1.3.1 Resultados do Cenário 1

Os resultados obtidos com classificação das postagens por meio do Cenário 1 são exibidos na Tabela 5.4. Nesse Cenário, sem fazer uso do AG o módulo 1 chegou a atingir no BD1 *F-measure* média de 0,946 (corte 0), 0,961 (corte 0) e 0,953 (corte 0) nas classes Dúvida, Neutra e Resposta. Para as mesmas classes o BD2 atingiu *F-measures* de 0,937 (corte 0), 0,943 (corte) e 0,979 (corte 0).

		Corte 0			**Corte 5**			**Corte 10**		
		D	**N**	**R**	**D**	**N**	**R**	**D**	**N**	**R**
Sem o AG	BD1	**0,946**	**0,961**	**0,953**	**0,913**	**0,927**	**0,916**	**0,917**	**0,922**	**0,917**
		0,0041	0,0050	0,0035	0,0033	0,0047	0,0020	0,0032	0,0048	0,0024
	BD2	**0,937**	**0,943**	**0,979**	**0,160**	**0,521**	**0,176**	**0,158**	**0,520**	**0,170**
		0,0041	0,0032	0,0018	0,0107	0,0021	0,0055	0,01	0,0022	0,0060
	BD3	**0,874**	**0,979**	**0,987**	**0,826**	**0,960**	**0,980**	**0,757**	**0,954**	**0,975**
		0,008	0,0014	0,0009	0,0073	0,0026	0,0006	0,0119	0,0034	0,0009
Com o AG	BD1	**0,978**	**0,976**	**0,985**	**0,957**	**0,957**	**0,961**	**0,949**	**0,948**	**0,954**
		0,0021	0,0024	0,0026	0,0021	0,0030	0,0023	0,0024	0,0043	0,0020
	BD2	**0,981**	**0,985**	**0,983**	**0,912**	**0,927**	**0,957**	**0,883**	**0,905**	**0,939**
		0,003	0,001	0,002	0,011	0,011	0,004	0,009	0,009	0,004
	BD3	**0,978**	**0,997**	**0,988**	**0,957**	**0,964**	**0,992**	**0,930**	**0,964**	**0,989**
		0,005	0,0006	0,003	0,002	0,0019	0,0004	0,002	0,002	0,0005

Tabela 5.4: Resultados cenário 1

Seguido pelo o BD3 que, em todos os cortes, obteve resultados superiores a 0,90. Utilizando o AG para evoluir os parâmetros do SVM os resultados foram potencializados em

todos os cortes, com o AG, mostrando que no Cenário 1 o sistema não se restringiu a obter bons resultados da forma isolada (*outliers*) ao contrário disso manteve-se consistente em todas as bases de dados.

5.1.3.2 Resultados do Cenário 2

O Cenário 2 sem a utilização do AG (Tabela 5.5) obteve as melhores médias de *F-measure* quando o corte estava em 0, sendo interessante destacar alguns resultados, tais como: a classe Dúvida no BD1 com 0,870 (corte 0), a classe Neutra nos BD1 e BD2 com respectivas taxas de 0,98 (corte 0) e 0,99 (corte 0) e a classe Resposta no BD3 com 0,952 (corte 0).

		Corte 0			**Corte 5**			**Corte 10**		
		D	**N**	**R**	**D**	**N**	**R**	**D**	**N**	**R**
Sem o AG	BD1	**0,870**	**0,98**	**0,88**	**0,137**	**0,373**	**0,173**	**0,146**	**0,373**	**0,171**
		0,00	0,0005	0,0003	0,0153	0,0012	0,0071	0,0016	0,0056	0,0022
	BD2	**0,285**	**0,621**	**0226**	**0,424**	**0,738**	**0,497**	**0,375**	**0,705**	**0,4329**
		0,2296	0,1223	0,2471	0,2147	0,1441	0,3285	0,2223	0,1390	0,3146
	BD3	**0,283**	**0,990**	**0,952**	**0,225**	**0,963**	**0,947**	**0,259**	**0,947**	**0,947**
		0,0044	0,0029	0,0005	0,0182	0,0003	0,0010	0,0337	0,00	0,0010
Com o AG	BD1	**0,896**	**0,995**	**0,911**	**0,838**	**0,950**	**0,849**	**0,844**	**0,942**	**0,876**
		0,0014	0,0	0,0011	0,0026	0,0099	0,0047	0,022	0,0022	0,012
	BD2	**0,750**	**0,949**	**0,834**	**0,721**	**0,905**	**0,825**	**0,697**	**0,877**	**0,788**
		0,0401	0,0114	0,0179	0,0394	0,024	0,825	0,019	0,018	0,023
	BD3	**0,549**	**0,994**	**0,964**	**0,476**	**0,960**	**0,973**	**0,498**	**0,954**	**0,959**
		0,0064	0,0	0,0002	0,0180	0,0004	0,0014	0,016	0,003	0,0007

Tabela 5.5: Resultados cenário 2

Ao analisar a Tabela 5.5, de modo geral, conclui-se que os melhores resultados no Cenário 2 se deram nas execuções com o AG chegando às classes Dúvida, Neutra e Resposta a atingir médias de *F-measure* nessa ordem: 0,896 (corte 0), 0,995 (corte 0) e 0,973 (corte 5).

5.1.3.3 Resultados do Algoritmo estado da arte

O algoritmo proposto por ROLIM; FERREIRA; COSTA (2016) foi executado em sua melhor configuração. Ao analisar os resultados (Tabela 5.6) observa-se que as melhores médias de *F-measure* aconteceram no BD1 e BD2.

Por outro lado, o algoritmo demonstrou dificuldade de classificação no BD3 especialmente nas classes Dúvida e Neutra. De acordo com (PHUA; ALAHAKOON; LEE (2004)), algoritmos de aprendizagem de máquina são sensíveis a bases desbalanceadas, implicando em priorização das classes com maior número de elementos. Por isso, entende-se que a distribuição dos dados do BD3 pode ter influenciado no desempenho do algoritmo de (ROLIM; FERREIRA; COSTA (2016)). Entretanto, é importante ressaltar que o módulo 1 deste estudo também classificou as postagens do BD3. Além disso, o maior número de postagens do tipo Respostas é comum

		Corte 0			Corte 5			Corte 10		
		D	**N**	**R**	**D**	**N**	**R**	**D**	**N**	**R**
Rolim *et al.*	BD1	**0,954**	**0,940**	**0,927**	**0,880**	**0,764**	**0,837**	**0,847**	**0,700**	**0,803**
		0,0095	0,0010	0,011	0,0071	0,014	0,009	0,0081	0,0028	0,011
	BD2	**0,963**	**0,943**	**0,961**	**0,886**	**0,847**	**0,851**	**0,875**	**0,803**	**0,829**
		0,0074	0,010	0,009	0,011	0,014	0,0011	0,009	0,009	0,010
	BD3	**0,914**	**0,914**	**0,983**	**0,81**	**0,654**	**0,941**	**0,812**	**0,606**	**0,939**
		0,0189	0,010	0,003	0,009	0,006	0,006	0,008	0,026	0,003

Tabela 5.6: Resultados do algoritmo de ROLIM; FERREIRA; COSTA (2016)

se tratando de fóruns de discussão educacionais. Outro fator que pode ter impactado no resultado é a existência de algumas postagens no BD3 onde os estudantes escrevem uma pergunta e a respondem na mesma postagem. Dessa forma, como o algoritmo de (ROLIM; FERREIRA; COSTA (2016)) considera a frequência de interrogações para aferir se uma postagem trata-se de uma Dúvida, Comentário Neutro ou uma Resposta, postagens como as exemplificadas na Tabela 5.7 podem ter confundido o classificador.

Aluno 1	"O que é Java? É uma linguagem de programação orientada a objeto, atualmente pertence a Oracle, mas foi desenvolvida pela Sun Microsystems na década de 90"
Aluno 2	"O que é UML? UML é um acrônimo para a expressão Unified Modeling Language. Pela definição de seu nome, vemos que UML é uma linguagem que define uma série de artefatos que nos ajuda na tarefa de modelar e documentar os sistemas orientados a objetos que desenvolvemos".

Tabela 5.7: Exemplos de postagens da classe Resposta do BD3

Os resultados mencionados acima não confrontam os divulgados em ROLIM; FERREIRA; COSTA (2016) apesar de serem inferiores aos divulgados pelos autores supracitados. Atribui-se os resultados inferiores ao fato de ser sintetizado o desempenho médio do algoritmo com o número de execuções 30.

5.1.3.4 Testes de hipóteses

Os testes de hipóteses apresentados nesta seção foram realizados com um intervalo de confiança de 95% (Tabela 5.8) e um o ponto crítico superior, com o valor de 1,645, baseado na tabela de distribuição normal padrão.

Ao analisar os resultados obtidos, na Tabela 5.8, nota-se que a hipótese nula H_{0a} foi rejeitada em todas as classes (Dúvida, Neutra e Resposta) devido aos valores de probabilidade inferiores ao nível de significância de 0,05. Portanto, a *F-measure* obtida ao considerar as funções gramaticais para criar os sacos de palavras e os vetores de características é superior à obtida quando não consideradas as funções gramaticais. As probabilidades obtidas na hipótese H_{0b} em todas as classes de postagem foram de ≈0,00 e por serem inferiores à probabilidade

		Intervalo de confiança 95%		***z-valor***	***p-valor***
		Inferior	**Superior**		
H_{0a}	Dúvida	-∞	1,645	26,72304	≈ 0,00
	Neutra	-∞	1,645	7,9119809	≈ 0,00
	Resposta	-∞	1,645	11,677742	≈ 0,00
H_{0b}	Dúvida	-∞	1,645	20,00186	≈ 0,00
	Neutra	-∞	1,645	18,378388	≈ 0,00
	Resposta	-∞	1,645	14,46111	≈ 0,00
H_{0c}	Dúvida	-∞	1,645	16,06112	≈ 0,00
	Neutra	-∞	1,645	13,309292	≈ 0,00
	Resposta	-∞	1,645	13,806955	≈ 0,00
H_{0d}	Dúvida	-∞	1,645	20,42291	≈ 0,00
	Neutra	-∞	1,645	3,2434821	≈ 0,00
	Resposta	-∞	1,645	4,3306514	0,0006

Tabela 5.8: Resultados dos testes de hipóteses

de 0,05 implicam na rejeição da hipótese com 95% de confiança. Dessa forma, a hipótese alternativa é aceita mostrando que os resultados alcançados pelo módulo 1 com o uso do AG são melhores em comparação aos resultados obtidos com sem a utilização do AG. A terceira hipótese H_{0c} também foi rejeitada em todas as classes de postagens. Tal rejeição evidencia um melhor desempenho do módulo 1 ao utilizar conjuntamente funções gramaticais e o AG ao invés de considerar somente as funções gramaticais. A hipótese H_{0d}, por sua vez, atingiu probabilidades de ≈0,00, 0,0006 e ≈0,00 todas inferiores 0,05. Com isso, a hipótese nula (H_{0d}) foi rejeitada mostrando que as *F-measures* alcançadas ao considerar as funções gramaticais e o AG são superiores às *F-measures* obtidas apenas com o uso do AG. Com base nos testes estatísticos, considerar as funções gramaticais para criar os sacos de palavras e os vetores de características influência positivamente no desempenho do classificador. Atribui-se isto ao fato de algumas palavras poderem desempenhar diferentes funções gramaticais em uma sentença. Por exemplo, a palavra "mais" pode ser etiquetada como Substantivo, Pronome, Preposição, Advérbio ou Conjunção. Também foi evidenciado o potencial do AG para evoluir os parâmetros do SVM, bem como a potencialização do módulo 1 com a atuação conjunta do AG e das funções gramaticais. Por fim, os resultados considerando as funções gramaticais e utilizando o AG foram comparados com os do algoritmo de ROLIM; FERREIRA; COSTA (2016). Com p-valores de ≈0,00, para todas as classes de postagens, a hipótese H_{0e} também foi rejeitada (Tabela 5.9).

		Intervalo de confiança 95%		***z-valor***	***p-valor***
		Inferior	**Superior**		
H_{0g}	Dúvida	-∞	1,645	21,71310	≈ 0,00
	Neutra	-∞	1,645	25,22902854	≈ 0,00
	Resposta	-∞	1,645	24,64754215	≈ 0,00

Tabela 5.9: Resultados dos testes comparativos com o algoritmo de ROLIM; FERREIRA; COSTA (2016)

Diante disso, conclui-se com 95% de confiança que o módulo 1 apresentou um melhor desempenho em termos de *F-measure*, no processo de identificação de gêneros (Dúvida, Neutra e Resposta) de postagens de fóruns de discussão, levando em consideração as três bases de dados utilizadas.

5.1.4 Experimentos das outras LA

Os experimentos descritos nesta seção foram realizados no BD4 o qual é desbalanceado, devido à natureza do fórum, dificultando a identificação dos aspectos colaborativos da abordagem proposta.

O módulo de expressar apreciação pelos outros participantes foi experimentado com os algoritmos Perceptron MLP, J48, *Naive Bayes* e SVM. Conforme mostra a Tabela 5.10, todos os classificadores obtiveram resultados equiparáveis para ambas as classes com exceção do SVM que atingiu para a classe 1 *F-measure* máxima de 0.218. Apesar da pequena vantagem do *Multilayer Perceptron* (MLP) em relação aos demais, é importante destacar que o treinamento do J48 e do *Naive Bayes* é mais rápido e, por isso, esses são excelentes opções considerando a proposta de extrair LA colaborativas de acordo com a inserção de postagens em fóruns educacionais. Além disso, é importante destacar que o J48, por ser uma árvore de decisão, gera um modelo de classificação interpretável pelo ser humano. Por isso, para experimentos futuros utilizaremos o J48.

	Característica encontrada			**Característica não encontrada**		
	Precision	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
MLP	0.887	0.759	0.818	0.962	0.984	0.973
J48	0.896	0.723	0.800	0.957	0.986	0.971
Naive Bayes	0.849	0.747	0.795	0.960	0.979	0.969
SVM	0.611	0.133	0.218	0.876	0.986	0.928

Tabela 5.10: Resultados do Módulo de Apreciação pelos Outros Participantes

O módulo de partilha de informações e recursos, além de buscar por conteúdos compartilhados nas postagens (*links*), procura verificar a ligação do conteúdo com a discussão corrente no fórum. Nesse sentido, com base nas técnicas de MT utilizadas, foram criados os seguintes cenários de execução:

- **Cenário 1** - considerando a remoção de *stopwords*, *stemming* e *unigram*;
- **Cenário 2** - considerando a remoção de *stopwords* e *unigram*;
- **Cenário 3** - considerando a remoção de *stopwords*, *stemming* e *bi-gram*;
- **Cenário 4** - considerando a remoção de *stopwords* e *bi-gram*;

Cada cenário foi executado com o conjunto composto pelo conteúdo compartilhado pelo estudante (CC) sendo montado de duas formas: (1) extraindo textos de todas as *tags* da página web compartilhada e (2) extraindo textos apenas da *tag* **<title>** da página web compartilhada. A Tabela 5.11 mostra os resultados obtidos com o conjunto CC resultante da opção (1). Como pode ser observado, as *F-measures* conseguidas para a classe Característica não encontrada (*False*) estão acima de 0.98 mostrando que boa parte das postagens, sem a presença de partilhamento de conteúdos relacionados à discussão, foram classificadas corretamente.

	Característica encontrada			**Característica não encontrada**		
	Precision	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
Cenário 1	0.500	0.688	0.579	0.991	0.981	0.986
Cenário 2	0.550	0.688	0.611	0.991	0.985	0.988
Cenário 3	0.583	0.438	0.500	0.985	0.991	0.988
Cenário 4	0.778	0.438	0.560	0.985	0.997	0.991

Tabela 5.11: Resultados do Módulo Partilha de Informações e Recursos - Opção 1

Já os resultados da classe Característica encontrada (*True*) evidenciam os bons índices de recuperação das postagens positivas, nos cenários *unigram* (1 e 2), chegando a alcançar *recall* de 0.688. Por outro lado, os que fizeram uso de *bi-gram* alcançaram maiores precisões, especificamente, 0.583 e 0.778. Os resultados também demonstram uma maior harmonia entre precisão e *recall* nos cenários 2 e 4, quando não foi utilizada a técnica *stemming*, chegando a obter 0.611 e 0.560 de *F-measure* respectivamente. Para um sistema educacional o ideal é dispor de boa precisão e *recall*, no entanto, entre as duas medidas, é mais importante ter uma *recall* elevada, pois uma avaliação incorreta de uma postagem pode levar a desmotivação do estudante. Com isso, conclui-se que ao extrair textos de todas as *tags* das páginas web compartilhadas, a melhor configuração do módulo 2 é com a combinação das técnicas de remoção de *stopwords* e *unigram* (Cenário 2).

Na Tabela 5.12 são apresentados os resultados obtidos com o conjunto CC na opção (2). Chama atenção a precisão (0.917) e a *recall* (0.688) atingidas para a classe Característica encontrada (*True*) no cenário 2. Outro ponto a ser destacado são os baixos valores de *recall* nos cenários 3 e 4, sendo atribuído tais resultados a característica do *bi-gram* de criar sequências com dois termos aliado a pequena quantidade de textos presentes na *tag* **<title>**.

	Característica encontrada			**Característica não encontrada**		
	Precision	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
Cenário 1	0.500	0.688	0.579	0.991	0.981	0.986
Cenário 2	0.917	0.688	0.786	0.991	0.998	0.995
Cenário 3	1.000	0.188	0.316	0.978	1.000	0.989
Cenário 4	1.000	0.125	0.222	0.977	1.000	0.988

Tabela 5.12: Resultados do Módulo Partilha de Informações e Recursos - Opção 2

É importante ressaltar que as postagens do BD4 são relacionadas com os seguintes temas:

Estatística para análise de dados, Internet e web, Programação orientada a objetos e Sistemas da informação. Dessa forma, obteve-se o melhor resultado no cenário 2 sintetizado na Tabela 5.12. Neste, foi identificado partilhamento de informações em 12 postagens, sendo 11 classificações corretas para a classe Característica identificada (*True*). A partir da Tabela 5.13, é possível observar para cada postagem o conteúdo do fórum, a URL compartilhada pelo estudante e o conjunto resultante da intersecção entre o conjunto com as palavras representativas do fórum em questão e o conjunto referente ao conteúdo compartilhado.

Post	**Conteúdo do fórum**	***url* compartilhada**	**Conjunto resultante**
1	Estatística para análise de dados	`https://goo.gl/8f3Z4M`	dispersao
2	Estatística para análise de dados	`https://goo.gl/Q8DWoH`	tamanho, amostras
3	Estatística para análise de dados	`https://goo.gl/TjziHd`	teste
4	Estatística para análise de dados	`https://goo.gl/FwkYBT`	parametros
5	Estatística para análise de dados	`https://goo.gl/SjCmGE`	amostragem
6	Estatística para análise de dados	`https://goo.gl/GjA21T`	variaveis
7	Internet e web	`http://www.asciitable.com`	html, decimal
8	Programação orientada a objetos	`https://goo.gl/7ahhfw`	fundamentos, uml
9	Programação orientada a objetos	`https://goo.gl/qeZhjB`	java
10	Programação orientada a objetos	`https://goo.gl/QNKAJB`	orientada, objetos
11	Sistemas da informação	`https://goo.gl/UL36ho`	portal, corporativo
12	Programação orientada a objetos	`https://goo.gl/DQUJCB`	computador

Tabela 5.13: Postagens enquadradas na classe Característica identificada - Módulo 3

O último módulo, que busca identificar o reconhecimento de presença de grupo nas postagens, foi desenvolvido através de uma por abordagem baseada em regras. Por essa razão ele não foram realizados testes estatísticos para avaliar a performance, mas os resultados foram relatados abaixo. É importante destacar que esse tipo de avaliação é utilizada em outros trabalhos que utilizam solução baseada em regras (SILVA, 2012; DIONÍSIO et al., 2017).

Este módulo atingiu 0.473 de precisão e 1.00 de *recall* para a classe característica encontrada (quando existia algum indicativo de reconhecimento de presença) e 1.00 e 0.98 de precisão e *recall* na classe 2 (característica não encontrada).

5.2 Aplicação piloto

Esta seção apresenta uma ferramenta fórum, nomeada de iFórum, que utiliza a abordagem proposta (Seção 4) para analisar postagens inseridas nas discussões e fornecer LA relacionadas à colaboração entre os estudantes. A primeira parte da seção descreve a estrutura do iFórum e, mais adiante, são expostas análises da aplicação num cenário real.

5.2.1 iFórum

Como forma de possibilitar a comunicação entre o iFórum e a abordagem proposta na Seção 4, foi adotada a tecnologia *WebServices* conhecida por proporcionar interoperabilidade entre diferentes sistemas (CONCEIÇÃO, 2017). O *WebServices* criado abarca todos os módulos apresentados na seção 4, sendo estruturado seguindo a arquitetura em camadas (SOMMERVILLE, 2007). As camadas presentes no *WebServices* são: ***Business Logic Layer*** - camada de negócio e; ***Data Acess Layer*** - camada de acesso a dados.

A comunicação entre as aplicações, ilustrada na Figura 5.1, acontece por meio da linguagem de intercâmbio *JavaScript Object Notation* (JSON)[1]. Assim, as postagens inseridas nas discussões são enviadas para o *WebServices* o qual as processa e retorna um JSON contendo todos os rótulos atribuídos por cada módulo da abordagem (zero - característica não identificada, um - característica identificada), por exemplo: {"solicitacao_feedback": 1, "responde_questoes": 1, "elogio_apreciacao": 1, "partilha_informacoes": 1, "reconhece_presenca_grupo": 1}. Nesse exemplo, foram identificadas todas as características colaborativas na postagem analisada.

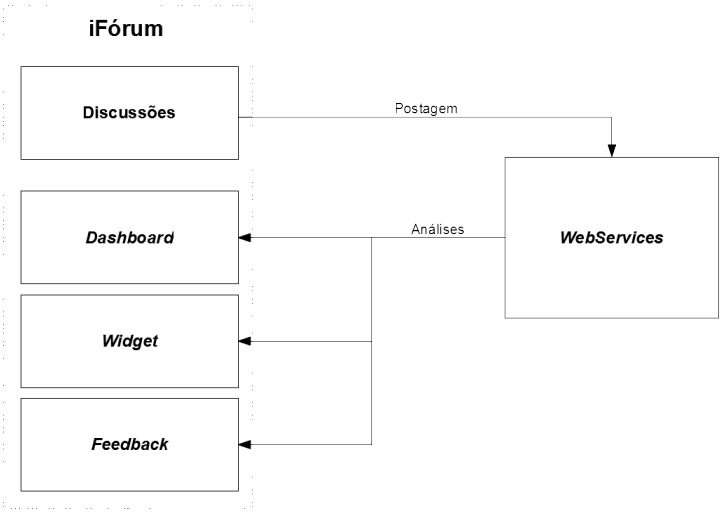


Figura 5.1: Funcionamento do iFórum

Depois de receber as análises, o iFórum municia: (1) um ***dashboard***, disponível apenas para os professores, com estatísticas da colaboração por estudante e por fórum; um ***widget***,

[1] http://www.json.org/>

chamado de termômetro da discussão, disponível para todos os estudantes e; exibe um ***feedback*** individual para o estudante na tentativa de motivá-lo a continuar/começar a colaborar como exemplifica a Tabela 5.14.

Características identificadas	**Exemplos de *feedbacks***
Presença de grupo	"De fato, não tem como debater sozinho. Por isso, é sempre bom reconhecer a presença dos colegas! Vamos em frente!"
Elogio/Apreciação	"Parabéns! Ao valorizar as contribuições dos seus colegas, você está contribuindo para o aprofundamento do debate."
Partilhamento de informações	"É muito interessante trazer conteúdos externos para enriquecer a discussão! Mas cuidado, é bom sempre analisar se o conteúdo externo pode contribuir para a discussão."

Tabela 5.14: Exemplos de *feedbacks*

A Figura 5.2 exibe um exemplo de uma página de discussão no iFórum onde os estudantes podem inserir e visualizar postagens de outros participantes. Ainda é possível perceber ao lado esquerdo o ***widget***, termômetro da discussão, configurado para aumentar e diminuir o seu nível de acordo com o atual cenário colaborativo do respectivo fórum.

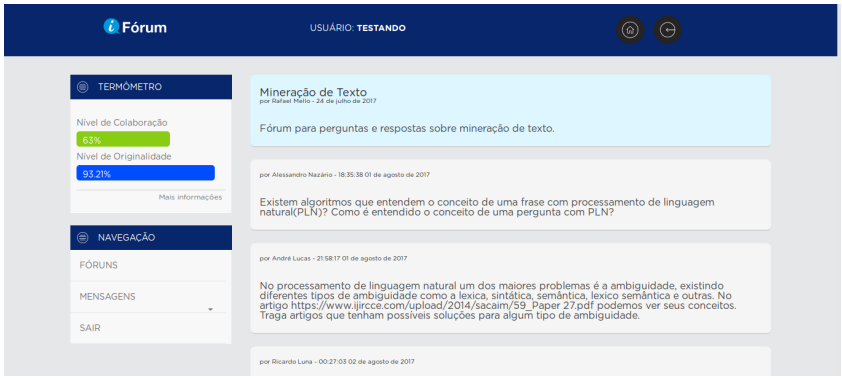


Figura 5.2: Exemplo de um fórum no ifórum

Ao invés de estimular um ambiente de competição entre os aprendizes, espera-se com a exibição de um nível geral de colaboração despertar o desejo de colaborar. Nesse sentido, para mostrar aos aprendizes como eles podem impulsionar a colaboração, na parte inferior do termômetro existe uma opção "mais informações" que ao ser acionada informa quais as características colaborativas são consideradas (Figura 5.3).

Para facilitar o acompanhamento da colaboração no fórum, o nível de participação de cada aluno é exibido em um *dashboard* exclusivo para o professor (administrador do fórum criado). A visualização deste pode ser feita por fóruns ou por estudante, como mostra as Figuras 5.4 e 5.5.

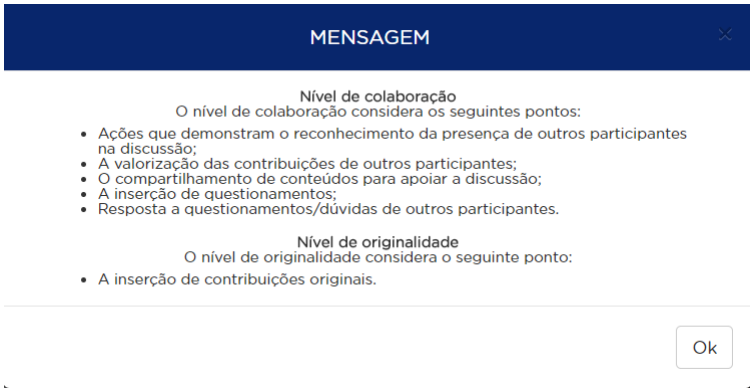


Figura 5.3: Exemplo de página informativa

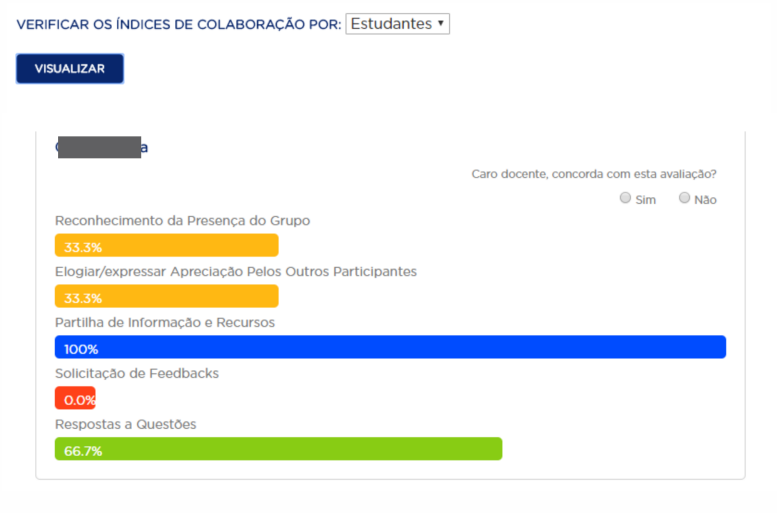


Figura 5.4: Exemplo de visualização das análises colaborativas por estudantes

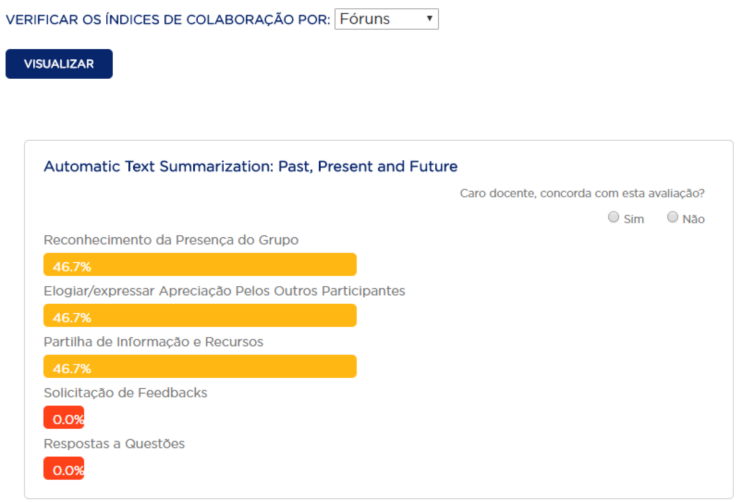


Figura 5.5: Exemplo de visualização das análises colaborativas por fóruns

5.2.2 Aplicação no ambiente real - (quase-experimento)

A avaliação do impacto do iFórum na mediação pedagógica e na colaboração entre estudantes aconteceu numa disciplina semipresencial, de Tópicos Avançados em Inteligência Artificial, com 12 estudantes do oitavo período do curso de Ciência da Computação da Universidade Federal Rural de Pernambuco (UFRPE). Como forma de preservar a identidade dos aprendizes estes serão referenciados a partir deste ponto de A1 à A12.

Como foi adotado apenas um grupo de estudantes, para poder estabelecer a relação causa-efeito do iFórum no objeto em estudo, foi conduzido um quase-experimento num período de duas semanas. Conforme caracteriza GIL (2015), a pesquisa quase experimental utiliza apenas um grupo (objeto de pesquisa) o qual é comparado/avaliado antes e após uma intervenção. Na primeira semana foram realizadas discussões por meio do fórum disponível no AVA Moodle e na segunda ocorreram discussões utilizando o iFórum. Em ambas as discussões ocorreram em torno de artigos científicos relacionados à disciplina em questão, sendo colocado em debate um artigo por semana. O primeiro artigo disponibilizado abordava *Natural Language Processing with Deep Learning* e o segundo *Automatic Text Summarization*. Além disso, nos dois momentos, foi utilizado o mesmo método de mediação: (1) Disponibilização do artigo tema; (2) Explicação sobre como deveria ocorrer as discussões e; (3) Intervenções do mediador apenas quando solicitadas.

É importante observar que não foram estabelecidas formas de controle em relação à familiaridade dos estudantes com o artigo disponibilizado. Isto é, o aluno pode atuar mais em uma discussão por sentir-se mais confortável com determinado tema. Somando-se a isto a pequena amostra de participantes, configuram-se as principais ameaças à validade deste quase-experimento.

Na primeira semana foram registradas apenas 10 postagens como mostra a Figura 5.6, a maioria delas contemplou apenas uma característica colaborativa "Responder a questões".

Em outras palavras, eram simplesmente um resumo do artigo em debate. Também não foram identificadas demonstrações de elogios/apreciação pelos outros participantes, partilhamento de informações além de qualquer postagem por parte dos alunos A11 e A12.

Ao contrário disso, na semana seguinte houve um aumento, chegando a serem inseridas um total de 30 postagens. Ficando, assim, evidente o crescimento da participação dos aprendizes com a utilização do iFórum. Porém, como inserir postagens não necessariamente significa colaborar, as informações dispostas na Figura 5.6 também demonstram elevação nos índices de colaboração na segunda semana. De forma mais específica, do total de 12 estudantes, 10 (80%) passaram a expressar mais características colaborativas.

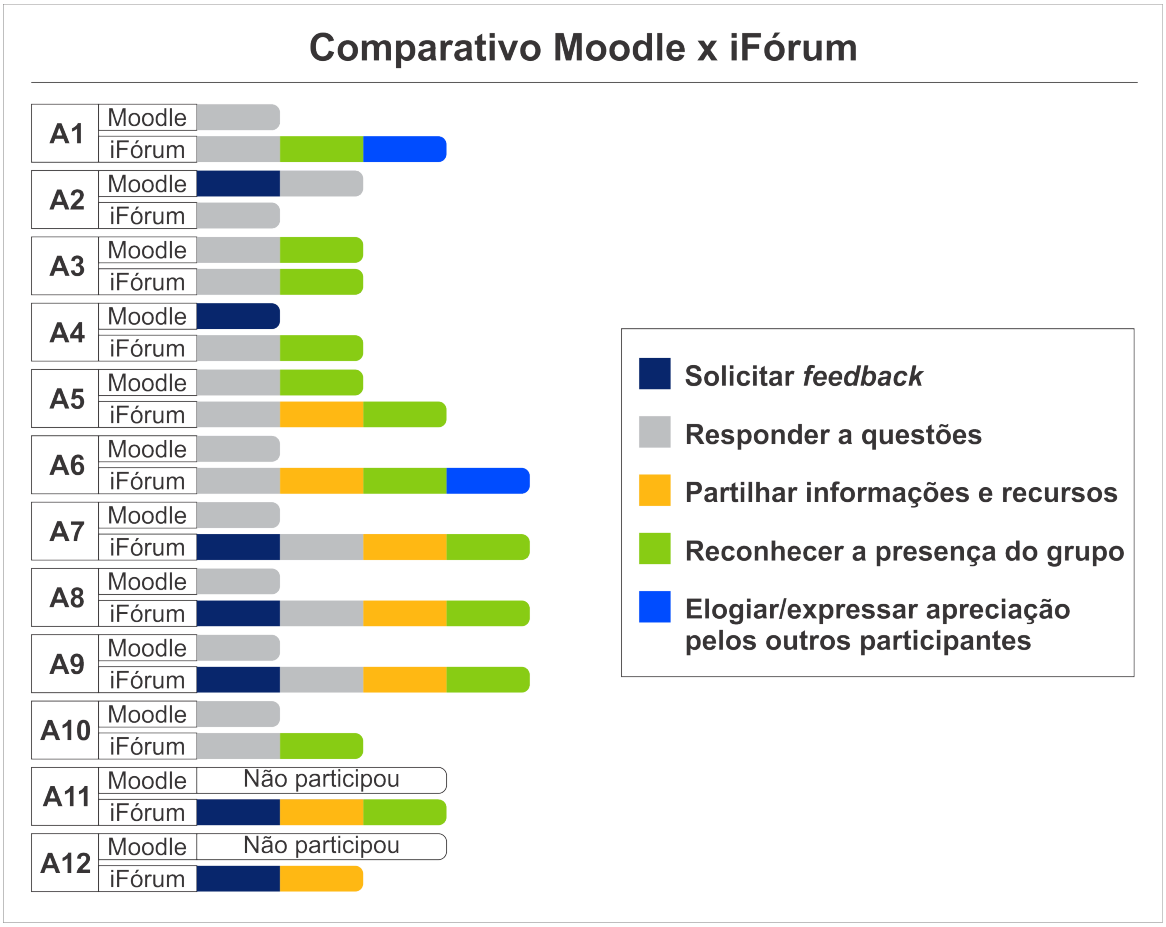


Figura 5.6: Comparativo da colaboração no Moodle e no iFórum

Apesar da visível melhoria da colaboração na segunda experiência, o aluno A2 teve uma melhor *performance* colaborativa no debate conduzido no Moodle e o A3 manteve a mesma *performance* em ambos os fóruns, inclusive, com as mesmas ações "Responder a questões"e "Reconhecer a presença do grupo".

O iFórum possui um *widget* municiado com o nível de colaboração do fórum além de *feedbacks* a serem exibidos aos aprendizes. No entanto, a forma como foi configurado este quase-experimento não possibilita inferir qual desses elementos (ou ambos) influenciou no cenário colaborativo. Apesar disso, a percepção do mediador sobre o iFórum foi bastante

positiva. Segundo ele, as estatísticas fornecidas foram levadas em consideração para atribuir a nota de participação dos alunos. Sendo a atuação nos fóruns responsável por compor 10% da nota final da disciplina.

6

Trabalhos relacionados

O potencial colaborativo dos fóruns de discussão somado aos baixos índices de colaboração entre estudantes, em atividades conduzidas na respectiva ferramenta, têm motivado a condução de pesquisas voltadas a apoiar atuações docentes pautadas em incentivar debates colaborativos. Como a maior parte do conteúdo gerado nos fóruns é textual, existem diversos estudos sobre o emprego de técnicas de Mineração de Textos para automatizar a identificação do cenário colaborativo e facilitar a tomada de decisões dos envolvidos.

Em LIN; HSIEH; CHUANG (2009) é apresentado um sistema voltado para a identificação automática do gênero das postagens de fóruns conduzidos em língua chinesa. Os gêneros considerados foram: Anúncio, Pergunta, Esclarecimento, Interpretação, Conflito, Afirmação e Outros (tradução nossa). Os autores utilizaram uma base de dados com 9178 postagens e o classificador J48 o qual obteve até 0,717 de *f-measure* nos experimentos realizados.

QU; LIU (2012) proporam um etiquetador que considera a dependência entre as postagens como forma de inferir se um *post* retrata um problema (pergunta), uma solução para o problema (resposta), uma avaliação da resposta (confirmação) ou diversos. Com isso, utilizando o modelo *conditional random field Conditional Random Field* (CRF), o etiquetador proposto alcançou *f-measure* de 0,80 na classificação das postagens da classe pergunta.

No cenário brasileiro é possível identificar alguns trabalhos nesse sentido, em GUIMARAES; ESMIN (2014) é realizada a classificação automática de postagens de fóruns educacionais nos seguintes gêneros: Anúncio, Dúvida, Esclarecimento, Interpretação e Outros. Para isso, foram extraídas e rotuladas 6367 postagens as quais foram utilizadas em um experimento com os classificadores *Naive Bayes* e SVM atingindo 0,608 e 0,603 de *f-measure* respectivamente.

ROLIM; FERREIRA; COSTA (2016) concentraram-se na identificação de dúvidas em fóruns educacionais e, como parte do processo, classificaram as postagens nas classes Dúvida, Neutra e Resposta. Para tal, conduziram experimentos em duas bases de dados com um total de 1090 postagens com os classificadores *Naive Bayes*, J48 e MLP e como resultados alcançaram 0,97 de *f-measure*.

Com um olhar para outros aspectos colaborativos, em (SILVA, 2012) é proposta uma abordagem destinada à identificação automática de presença social em fóruns e *chats*. Os autores

propuseram uma abordagem baseada em regras específicas para as bases de dados utilizadas no estudo e também em regras genéricas (podem ser aplicadas em outros fóruns). Para testá-las, foram conduzidos dois experimentos em cada base de dados. A abordagem em questão chegou a atingir com as regras específicas 93,95% (base 1) e 88,80% (base 2) de taxa de acerto e com as regras genéricas 58,90% (base 1) e 5,00% (base 2). A maior dificuldade dessa abordagem é a forte dependência de um anotador inicial para os termos de colaboração.

No mesmo sentido, em sua dissertação de mestrado GOMES (2012) defendeu a importância do processo de ensino-aprendizagem centrado em interações entre os alunos. Para tanto, apresentou um sistema baseado em PLN focado em analisar automaticamente interações em fóruns para, em seguida, gerar indicadores que permitam ao aluno avaliar seu próprio desempenho e ao tutor/professor mensurar o desempenho dos participantes da discussão. O sistema classifica os comentários nas seguintes classes: saudação, debate, motivação, social, informação, confirmação, negação, tarefa, esclarecimento, indagação e agradecimento. O classificar escolhido foi Bayesiano o qual atingiu, nos experimentos, uma precisão geral de 0,40.

A Tabela 6.1 exibe um comparativo dos trabalhos relacionados com a abordagem proposta. Para isso, contém os seguintes campos: **Estudos** - referência do trabalho; **Idioma** - língua na qual a pesquisa é direcionada; **Método** - se a solução computacional exige uma supervisão humana durante sua execução, podendo ser automático, semi-automático ou manual; **Aspectos colaborativos considerados** - quais as características colaborativas são abordadas no estudo e; **Avaliação em um ambiente real** - se a solução proposta foi avaliada em um ambiente educacional real.

Dessa forma, a partir das informações presentes na Tabela 6.1, é possível identificar dois diferenciais da abordagem proposta neste pesquisa frente aos estudos relacionados. A primeira refere-se ao fato da abordagem ter sido avaliada com seus potenciais usuários evidenciando, na prática, seu potencial educacional. O outro diferencial consiste das características colaborativas consideradas, pois a maioria dos trabalhos focam apenas em dois tipos de demonstrações de colaboração "Dúvidas"e "Respostas". Os trabalhos mais próximos são SILVA (2012) e GOMES (2012). Porém, estes adotaram outros modelos de identificação de colaboração, com algumas diferenças conceituais.

Estudos	Idioma	Método	Aspectos colaborativos considerados	Avaliação em um ambiente real
LIN; HSIEH; CHUANG (2009)	Chinês	Automático	Pergunta e esclarecimento	não
QU; LIU (2012)		Automático	Pergunta e resposta	não
GUIMARAES; ESMIN (2014)	Português	Automático	Dúvida e esclarecimento	não
ROLIM; FERREIRA; COSTA (2016)	Português	Automático	Dúvida e resposta	não
SILVA (2012)	Português	Semi-automático	Presença social	não
GOMES (2012)	Português	Automático	Saudação, debate, motivação, social, informação, confirmação, negação, tarefa, esclarecimento, indagação e agradecimento	não
Abordagem proposta	Português	Automático	Socilitar *feedback* ou responder a questões, Reconhecer a presença do grupo, Elogiar/expressar apreciação pelos outros participantes e Partilhar informação e recursos.	sim

Tabela 6.1: Tabela com a comparação entre os estudos

7

Considerações finais

Esta pesquisa teve como principal objetivo apresentar uma abordagem para a extração automática de LA relacionadas à colaboração em fóruns educacionais. Como forma de subsidiar a realização deste objetivo, foram elencados alguns objetivos específicos: (1) realizar uma revisão sistemática da literatura; (2) propor uma combinação de técnicas de MT, AM e CE extrair LA colaborativas; (3) realizar experimentos computacionais; (4) propor um ambiente de teste para a aplicação da proposta.

Todos os objetivos foram alcançados, uma revisão sobre Mineração de Textos em fóruns educacionais foi realizada como forma de dar uma visão geral do atual cenário de estudos nessa área. Onde, a partir disso, ficou visível o avanço do campo de pesquisa apesar de terem sido identificadas poucas pesquisas com o objetivo de fortalecer, a principal característica dos fóruns, a colaboração. Foi proposta uma abordagem, combinando técnicas de MT, AM e CE, com quatro características do modelo de identificação de colaboração de MURPHY (2004).

Os resultados obtidos nos experimentos evidenciaram o potencial da abordagem, pois, chegou-se a alcançar para as características colaborativas adotadas os seguintes resultados em termos de *f-measure*:

- Solicitação de *feedback*: 0,981
- Responder a questões: 0,988
- Elogiar/expressar apreciação pelos outros participantes: 0,818
- Partilhamento de informações e recursos: 0,786
- Reconhecimento de presença do grupo: 0,642

Posteriormente, a abordagem foi integrada a uma ferramenta fórum capaz de fornecer estatísticas sobre os índices de colaboração da discussão. Por fim, foi realizado um quase-experimento em um ambiente real, onde, ao utilizar a ferramenta mencionada os estudantes passaram a demonstrar mais características colaborativas. Além disso, as estatísticas fornecidas apoiaram o professor da disciplina na atribuição das notas de participação no debate. É importante

também ressaltar que não é objetivo desta pesquisa posicionar uma nova ferramenta de discussão, esta foi implementada com o objetivo de verificar o quanto a abordagem proposta pode contribuir para a mediação pedagógica e aumento da colaboração entre os aprendizes.

7.1 Contribuições

Ao considerar a relevância do objeto em estudo para a área de Informática na Educação, com a realização desta pesquisa, entende-se estar contribuindo nos seguintes pontos:

- Identificação das técnicas de Mineração de Textos mais utilizadas para extração e classificação de textos em fóruns educacionais;
- Construção de uma abordagem computacional, para a identificação automática de colaboração, dedicada à auxiliar professores/tutores no acompanhamento e incentivo à colaboração em fóruns educacionais;

7.2 Artigos submetidos/aceitos

De acordo com a evolução da pesquisa e a obtenção de resultados preliminares, artigos foram submetidos para conferências e periódicos qualificados na área de Ciências da Computação conforme listado a seguir:

- Fórum.Edu: Um Fórum Educacional Mobile que utiliza Mineração de Texto (DIONÍSIO et al., 2016) (aceito).
- Análise de Classificadores para Avaliação automática em Fóruns Educacionais (FERREIRA et al., 2016) (aceito).
- Mineração de Texto Aplicada à Identificação de Colaboração em Fóruns Educacionais (DIONÍSIO et al., 2017) (aceito).
- Mineração de Textos em Fóruns Educacionais: uma revisão da literatura (DIONÍSIO et al., 2017) (aceito).
- Um sistema baseado em PLN e AG para apoiar a mediação pedagógica em fóruns de discussão (DIONÍSIO et al., 2017) (submetido).
- *A Survey on Text Mining in Online Education* (FERREIRA et al., 2017) (submetido).
- *Adopting Learning Analytics to Promote Collaboration in Portuguese Educational Forums* (DIONÍSIO et al., 2018) (submetido).

7.3 Limitações da pesquisa

Apesar dos resultados promissores, é possível identificar algumas limitações nesta pesquisa. O modelo de identificação de colaboração proposto por MURPHY (2004) considera quatro tipos de perguntas em uma discussão: **solicitação de classificação ou elaboração** - perguntas pedindo aos participantes para elaborar ou explicar algo; **colocação de questões retóricas** - questões focadas em incentivar a reflexão; **solicitação de *feedback*** - pedidos de retorno sobre o que o participante acaba de postar e; **provocação de pensamento e discussão** - questões concebidas para provocar pensamento e discussão. O módulo 1 aborda a solicitação de *feedback* e resposta a questões, contudo, a maneira como foi proposto pode abrigar sem distinção os demais tipos de perguntas presentes no modelo adotado.

No módulo 2 (Elogiar/expressar apreciação pelos outros participantes), por exemplo, uma palavra só será identificada como um elogio/apreciação caso esteja etiquetada com polaridade positiva em um dos dicionários *léxicos* utilizados (ver seção 4.2). Dessa forma, considerando a enorme diversidade de sinônimos na língua Portuguesa, isto pode ocasionar erros de classificação. Como descrito na seção 4.2, o módulo 2 busca associar elogios/apreciações à estudantes. Contudo, tal associação somente acontecerá se o indivíduo elogiado estiver explicitamente indicado no texto (Nome próprio).

Com relação ao processo de avaliação, a utilização de bases de dados desbalanceadas dificultou a classificação das postagens da classe com menor número de elementos, no caso, as classes positivas para as características colaborativas. Também, o curto período de aplicação da abordagem no ambiente real somado ao pequeno número de estudantes participantes, reduzem a confiabilidade do quase-experimento.

7.4 Trabalhos futuros

Como continuação deste estudo, espera-se:

1. Investir na evolução da abordagem proposta de modo a contemplar outros aspectos do modelo de identificação de MURPHY (2004), tais como: Partilhamento de informação pessoal, Discordar/desafiar as declarações feitas por outro participante, Introdução de novas perspectivas e entre outras;

2. Experimentar a abordagem em bases de dados maiores e balanceadas;

3. Experimentar a abordagem, em ambientes reais, por maiores períodos e com um número maior de estudantes e;

4. Criar uma *Application Programming Interface* (API) que possibilite a integração com o AVA Moodle.

Referências

ABED. **Relatório analítico da aprendizagem a distância no Brasil**. [S.l.: s.n.], 2012.

ABED. **Relatório analítico da aprendizagem a distância no Brasil**. [S.l.: s.n.], 2013.

ABED. **Relatório analítico da aprendizagem a distância no Brasil**. [S.l.: s.n.], 2014.

ABED. **Relatório analítico da aprendizagem a distância no Brasil**. [S.l.: s.n.], 2015.

ABED. **Relatório analítico da aprendizagem a distância no Brasil**. [S.l.: s.n.], 2016.

AFONSO, A. R. Brazilian Portuguese Text Clustering Based on Evolutionary Computing. **IEEE Latin America Transactions**, [S.l.], v.14, n.7, p.3370–3377, 2016.

AL-SHALABI, R.; OBEIDAT, R. Improving KNN Arabic Text Classification with N-Grams Based Document Indexing. In: PROCEEDINGS OF THE 6 TH INTERNATIONAL CONFERENCE ON INFORMATICS AND SYSTEMS INFOS2008. **Anais...** [S.l.: s.n.], 2008. p.108–112.

ALMATRAFI, O.; JOHRI, A.; RANGWALA, H. Needle in a haystack: identifying learner posts that require urgent response in mooc discussion forums. **Computers & Education**, [S.l.], 2017.

ALRUSHIEDAT, N.; OLFMAN, L. Facilitating collaboration and peer learning through anchored asynchronous online discussions. , [S.l.], 2013.

AN, H.; SHIN, S.; LIM, K. The effects of different instructor facilitation approaches on students' interactions during asynchronous online discussions. **Computers & Education**, [S.l.], v.53, n.3, p.749–760, 2009.

ANDERSON, T.; KANUKA, H. On-line forums: new platforms for professional development and group collaboration. **Journal of Computer-Mediated Communication**, [S.l.], v.3, n.3, p.0–0, 1997.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.21529/RESI**, [S.l.], v.5, n.2, 2006.

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: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE). **Anais...** [S.l.: s.n.], 2011. v.1, n.1.

BABAR, M. A.; ZHANG, H. Systematic literature reviews in software engineering: preliminary results from interviews with researchers. In: EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT, 2009. ESEM 2009. 3RD INTERNATIONAL SYMPOSIUM ON. **Anais...** [S.l.: s.n.], 2009. p.346–355.

BASTOS, H. P. P.; BERCHT, M.; WIVES, L. K. Presença social e pertencimento em fóruns educacionais: manifestação e percepção de afetividade. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE). **Anais...** [S.l.: s.n.], 2011. v.1, n.1.

BASTOS, H. P. P.; SILVA, J. M. Fatores de evasão em curso a distância: relato de pesquisa sobre evadidos do curso "leitura instrumental em inglês a distância" no iff, rj. **RENOTE**, [S.l.], v.7, n.3, p.64–72, 2010.

BLOOM, B. S. et al. **Manual de avaliação formativa e somativa do aprendizado escolar**. [S.l.: s.n.], 1983.

BRANCO, A. H.; SILVA, J. Tokenization of Portuguese: resolving the hard cases. , [S.l.], 2003.

CÂMARA, J. M. **Estrutura da língua portuguesa**. [S.l.]: Editora Vozes, 1970.

CARVALHO, P.; SILVA, M. J. SentiLex-PT: principais características e potencialidades. **Oslo Studies in Language**, [S.l.], v.7, n.1, 2015.

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**, [S.l.], v.50, n.3, p.678–692, 2008.

CHENG, C. K. et al. Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. **Computers & Education**, [S.l.], v.56, n.1, p.253–261, 2011.

CHOWDHURY, G. G. Natural language processing. **Annual review of information science and technology**, [S.l.], v.37, n.1, p.51–89, 2003.

CONCEIÇÃO, L. d. S. Web service para acesso a dados da aplicação caronas. , [S.l.], 2017.

DIANATI, M.; SONG, I.; TREIBER, M. **An introduction to genetic algorithms and evolution strategies**. [S.l.]: Technical report, University of Waterloo, Ontario, N2L 3G1, Canada, 2002.

DILLENBOURG, P. **What do you mean by collaborative learning?** [S.l.]: Oxford: Elsevier, 1999.

DIONÍSIO, M. et al. Fórum. Edu: um fórum educacional mobile que utiliza mineração de texto. In: WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. **Anais...** [S.l.: s.n.], 2016. v.5, n.1, p.346.

DIONÍSIO, M. et al. Mineração de Textos em Fóruns Educacionais: uma revisão da literatura. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE). **Anais...** [S.l.: s.n.], 2017. v.28, n.1, p.21.

DIONÍSIO, M. et al. 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). **Anais...** [S.l.: s.n.], 2017. v.28, n.1, p.1437.

DIONÍSIO, M. et al. Um sistema baseado em PLN e AG para apoiar a mediação pedagógica em fóruns de discussão. , [S.l.], 2017.

DIONÍSIO, M. et al. *Adopting Learning Analytics to Promote Collaboration in Portuguese Educational Forums*. , [S.l.], 2018.

DONDONIS DAUDT, S. I.; BEHAR, P. A. A gestão de cursos de graduação a distância e o fenômeno da evasão. **Educação**, [S.l.], v.36, n.3, 2013.

DRINGUS, L. P.; ELLIS, T. Using data mining as a strategy for assessing asynchronous discussion forums. **Computers & Education**, [S.l.], v.45, n.1, p.141–160, 2005.

DYCKHOFF, A. L. et al. Design and implementation of a learning analytics toolkit for teachers. **Journal of Educational Technology & Society**, [S.l.], v.15, n.3, p.58, 2012.

EIBEN, A. E.; SMITH, J. E. et al. **Introduction to evolutionary computing**. [S.l.]: Springer, 2003. v.53.

ELIAS, T. Learning analytics. **Learning**, [S.l.], 2011.

FELDMAN, R.; SANGER, J. **The text mining handbook**: advanced approaches in analyzing unstructured data. [S.l.]: Cambridge university press, 2007.

FERGUSON, R. **The construction of shared knowledge through asynchronous dialogue**. 2009. Tese (Doutorado em Ciência da Computação) — The Open University.

FERREIRA, M. A. D. et al. **Análise de Classificadores para Avaliação automática em Fóruns Educacionais**. [S.l.]: ENIAC, 2016.

FERREIRA, R. et al. *A Survey on Text Mining in Online Education*. , [S.l.], 2017.

FERRI, C.; HERNÁNDEZ-ORALLO, J.; MODROIU, R. An experimental comparison of performance measures for classification. **Pattern Recognition Letters**, [S.l.], v.30, n.1, p.27–38, 2009.

FONSECA, L. C. C.; FARIAS, M. P.; SILVA, R. d. J. da. Sistema de Recomendação de Links para o fomento de discussões em fóruns de um Ambiente Virtual de Aprendizagem. **RENOTE**, [S.l.], v.12, n.2, 2014.

FREITAS, S. de et al. How to use Gamified Dashboards and Learning Analytics for Providing Immediate Student Feedback and Performance Tracking in Higher Education. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB COMPANION, 26. **Proceedings...** [S.l.: s.n.], 2017. p.429–434.

FUKS, H. et al. Informações estatísticas e visuais para a mediação de fóruns educacionais. **Brazilian Journal of Computers in Education**, [S.l.], v.13, n.3, 2005.

GAŠEVIĆ, D.; DAWSON, S.; SIEMENS, G. Let's not forget: learning analytics are about learning. **TechTrends**, [S.l.], v.59, n.1, p.64–71, 2015.

GIL, A. C. Métodos e técnicas de pesquisa social. In: **Métodos e técnicas de pesquisa social**. [S.l.: s.n.], 2015.

GOKHALE, A. A. Collaborative learning enhances critical thinking. , [S.l.], 1995.

GOMES, G. A. F. EU-TU: o emprego da classificação automática de mensagens em fóruns eletrônicos de discussões para análise do processo de ensino e aprendizagem centrado em interações. **Rio de Janeiro, PPGI/IM/iNCE/UFRJ**, [S.l.], 2012.

GREFENSTETTE, J. J. Optimization of control parameters for genetic algorithms. **IEEE Transactions on systems, man, and cybernetics**, [S.l.], v.16, n.1, p.122–128, 1986.

GUIMARAES, F. d. R. N.; ESMIN, A. A. A. Identificação Automática de Gêneros das Mensagens em Fóruns de Discussões do AVA. In: THE BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS) AND ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL. SÃO CARLOS-SP. **Anais...** [S.l.: s.n.], 2014.

GUIMARÃES, I. C.; CAÇÃO, O.; COUTINHO, V. Da interação à colaboração em comunidades e fóruns de discussão. **Internet Latent Corpus Journal**, [S.l.], v.3, n.1, p.49–64, 2013.

HABERT, B. et al. Towards tokenization evaluation. In: LREC. **Proceedings...** [S.l.: s.n.], 1998. v.98, p.427–431.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. [S.l.]: MIT press, 2001.

HUGHES, S. C. et al. Overcoming social and psychological barriers to effective on-line collaboration. **Educational Technology & Society**, [S.l.], v.5, n.1, p.86–92, 2002.

JAIN, P. Virtual learning environment. **International Journal in IT & Engineering**, [S.l.], v.3, n.5, p.75–84, 2015.

JOACHIMS, T. **Learning to classify text using support vector machines**: methods, theory and algorithms. [S.l.]: Kluwer Academic Publishers Norwell, 2002. v.186.

KEAR, K. Communication aspects of virtual learning environments: perspectives of early adopters. , [S.l.], 2007.

KIM, J.; KWON, Y.; CHO, D. Investigating factors that influence social presence and learning outcomes in distance higher education. **Computers & Education**, [S.l.], v.57, n.2, p.1512–1520, 2011.

KIM, J.; SHAW, E. Scaffolding student online discussions using past discussions: pedabot studies. **Artificial Intelligence Review**, [S.l.], p.1–16, 2014.

KITCHENHAM, B. Procedures for performing systematic reviews. **Keele, UK, Keele University**, [S.l.], v.33, n.2004, p.1–26, 2004.

LAAL, M. et al. What do we achieve from learning in collaboration? **Procedia-Social and Behavioral Sciences**, [S.l.], v.93, p.1427–1432, 2013.

LAAL, M.; GHODSI, S. M. Benefits of collaborative learning. **Procedia-Social and Behavioral Sciences**, [S.l.], v.31, p.486–490, 2012.

LAAL, M.; LAAL, M. Collaborative learning: what is it? **Procedia-Social and Behavioral Sciences**, [S.l.], v.31, p.491–495, 2012.

LANG, C. et al. **Handbook of Learning Analytics**. [S.l.]: SOLAR, Society for Learning Analytics and Research, 2017.

LARSON, R.; FARBER, B.; PATARRA, C. tradução técnica. **Estatística aplicada**. [S.l.]: Prentice Hall, 2004.

LAU, R. Y. et al. Towards fuzzy domain ontology based concept map generation for e-learning. In: INTERNATIONAL CONFERENCE ON WEB-BASED LEARNING. **Anais...** [S.l.: s.n.], 2007. p.90–101.

LEITNER, P.; KHALIL, M.; EBNER, M. Learning Analytics in Higher Education—A Literature Review. In: **Learning Analytics**: fundaments, applications, and trends. [S.l.]: Springer, 2017. p.1–23.

LI, Y.; DONG, M.; HUANG, R. Semantic organization of online discussion transcripts for active collaborative learning. In: ADVANCED LEARNING TECHNOLOGIES, 2008. ICALT'08. EIGHTH IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2008. p.756–760.

LIM, C. P.; CHEAH, P. T. The role of the tutor in asynchronous discussion boards: a case study of a pre-service teacher course. **Educational Media International**, [S.l.], v.40, n.1-2, p.33–48, 2003.

LIN, F.-R.; HSIEH, L.-S.; CHUANG, F.-T. Discovering genres of online discussion threads via text mining. **Computers & Education**, [S.l.], v.52, n.2, p.481–495, 2009.

LO, R. T.-W.; HE, B.; OUNIS, I. Automatically building a stopword list for an information retrieval system. In: JOURNAL ON DIGITAL INFORMATION MANAGEMENT: SPECIAL ISSUE ON THE 5TH DUTCH-BELGIAN INFORMATION RETRIEVAL WORKSHOP (DIR). **Anais...** [S.l.: s.n.], 2005. v.5, p.17–24.

LOPEZ, M. I. et al. Classification via clustering for predicting final marks based on student participation in forums. **International Educational Data Mining Society**, [S.l.], 2012.

LUI, A. K.-F.; LI, S. C.; CHOY, S. O. An evaluation of automatic text categorization in online discussion analysis. In: ADVANCED LEARNING TECHNOLOGIES, 2007. ICALT 2007. SEVENTH IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2007. p.205–209.

MAZZOLINI, M.; MADDISON, S. When to jump in: the role of the instructor in online discussion forums. **Computers & Education**, [S.l.], v.49, n.2, p.193–213, 2007.

MELANIE, M. An introduction to genetic algorithms. **Cambridge, Massachusetts London, England, Fifth printing**, [S.l.], v.3, p.62–75, 1999.

MELO FERREIRA, F. J. de et al. Um Modelo de Fórum de Discussão com Suporte às Interações entre Aprendizes utilizando Mapas Conceituais. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE). **Anais...** [S.l.: s.n.], 2013. v.24, n.1, p.416.

MINHOTO, P.; MEIRINHOS, M. As redes sociais na promoção da aprendizagem colaborativa: um estudo no ensino secundário. **Educação, Formação & Tecnologias-ISSN 1646-933X**, [S.l.], v.4, n.2, p.25–34, 2012.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineraçao de textos. **Relatório Técnico–Instituto de Informática (UFG)**, [S.l.], 2007.

MURPHY, E. Recognising and promoting collaboration in an online asynchronous discussion. **British Journal of Educational Technology**, [S.l.], v.35, n.4, p.421–431, 2004.

NILSSON, N. J. Introduction to machine learning. An early draft of a proposed textbook. , [S.l.], 1996.

NUNES, B. P. et al. A topic extraction process for online forums. In: ADVANCED LEARNING TECHNOLOGIES (ICALT), 2014 IEEE 14TH INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2014. p.541–543.

OLIVEIRA JÚNIOR, R. L. de; ESMIN, A. A. Monitoramento Automático de Mensagens de Fóruns de Discussão Usando Técnica de Classificação de Texto. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE). **Anais...** [S.l.: s.n.], 2012. v.23, n.1.

ONAH, D. F.; SINCLAIR, J.; BOYATT, R. Exploring the use of MOOC discussion forums. In: LONDON INTERNATIONAL CONFERENCE ON EDUCATION. **Proceedings...** [S.l.: s.n.], 2014. p.1–4.

ORENGO, V. M.; HUYCK, C. A stemming algorithm for the portuguese language. In: STRING PROCESSING AND INFORMATION RETRIEVAL, 2001. SPIRE 2001. PROCEEDINGS. EIGHTH INTERNATIONAL SYMPOSIUM ON. **Anais...** [S.l.: s.n.], 2001. p.186–193.

PANITZ, T. The Case for Student Centered Instruction via Collaborative Learning Paradigms. , [S.l.], 1999.

PARK, J.-H.; CHOI, H. J. Factors influencing adult learners' decision to drop out or persist in online learning. **Journal of Educational Technology & Society**, [S.l.], v.12, n.4, 2009.

PERRENOUD, P. **Os ciclos de aprendizagem**: um caminho para combater o fracasso escolar. [S.l.]: Artmed Editora, 2016.

PHUA, C.; ALAHAKOON, D.; LEE, V. Minority report in fraud detection: classification of skewed data. **Acm sigkdd explorations newsletter**, [S.l.], v.6, n.1, p.50–59, 2004.

QU, Z.; LIU, Y. Sentence dependency tagging in online question answering forums. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: LONG PAPERS-VOLUME 1, 50. **Proceedings...** [S.l.: s.n.], 2012. p.554–562.

RAVI, S.; KIM, J. Profiling student interactions in threaded discussions with speech act classifiers. **Frontiers in Artificial Intelligence and Applications**, [S.l.], v.158, p.357, 2007.

RISH, I. An empirical study of the naive Bayes classifier. In: IJCAI 2001 WORKSHOP ON EMPIRICAL METHODS IN ARTIFICIAL INTELLIGENCE. **Anais...** [S.l.: s.n.], 2001. v.3, n.22, p.41–46.

ROLIM, V.; FERREIRA, R.; COSTA, E. Identificação Automática de Dúvidas em Fóruns Educacionais. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE). **Anais...** [S.l.: s.n.], 2016. v.27, n.1, p.936.

ROURKE, L. Assessing Social Presence In Asynchronous Text-based Computer Conferencing. **Journal of Distance Education**, [S.l.], v.14, n.2, p.51–70, 2001.

RUBIO, D.; VILLALON, J. A Latent Semantic Analysis Method to Measure Participation Quality Online Forums. In: ADVANCED LEARNING TECHNOLOGIES (ICALT), 2016 IEEE 16TH INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2016. p.18–19.

SALTON, G.; YANG, C.-S. On the specification of term values in automatic indexing. **Journal of documentation**, [S.l.], v.29, n.4, p.351–372, 1973.

SANTOS, J.; PAIVA, R.; BITTENCOURT, I. I. Avaliação Léxico-Sintática de Atividades Escritas em Algoritmo Genético e Processamento de Linguagem Natural: um experimento no enem. **Revista Brasileira de Informática na Educação**, [S.l.], v.24, n.2, 2016.

SCHEUER, O.; MCLAREN, B. Helping teachers handle the flood of data in online student discussions. In: INTELLIGENT TUTORING SYSTEMS. **Anais...** [S.l.: s.n.], 2008. p.323–332.

SIEMENS, G. et al. **Open Learning Analytics**: an integrated & modularized platform. 2011. Tese (Doutorado em Ciência da Computação) — Open University Press Doctoral dissertation.

SIEMENS, G.; LONG, P. Penetrating the fog: analytics in learning and education. **EDUCAUSE review**, [S.l.], v.46, n.5, p.30, 2011.

SILVA, J. K. K. d. Automatização do processo de identificação de presença social em fóruns e chats. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE). **Anais...** [S.l.: s.n.], 2012.

SILVA, J. R. M. F. da. **Shallow processing of Portuguese**: from sentence chunking to nominal lemmatization. 2007. Tese (Doutorado em Ciência da Computação) — Master's thesis.

SILVA, L. A. et al. Mineração de Dados em publicações de Fóruns de Discussões do Moodle como geração de Indicadores para aprimoramento da Gestão Educacional. In: WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. **Anais...** [S.l.: s.n.], 2015. v.4, n.1, p.1084.

SILVA, W. D. C. M.; FINGER, M. Improving CoGrOO: the brazilian portuguese grammar checker. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 9. **Proceedings...** [S.l.: s.n.], 2013.

SMOLA, A.; VISHWANATHAN, S. V. N. Introduction to Machine Learning. **Cambridge University Pres**, [S.l.], 2008.

SOMMERVILLE, I. Engenharia de Software-8ª Edição (2007). **Ed Person Education**, [S.l.], 2007.

SOUZA, M.; VIEIRA, R. Sentiment analysis on twitter data for portuguese language. **Computational Processing of the Portuguese Language**, [S.l.], p.241–247, 2012.

VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer science & business media, 2013.

VIEIRA, R.; LIMA, V. L. Lingüística computacional: princípios e aplicações. In: XXI CONGRESSO DA SBC. I JORNADA DE ATUALIZAÇÃO EM INTELIGÊNCIA ARTIFICIAL. **Anais...** [S.l.: s.n.], 2001. v.3, p.47–86.

WANKHEDE, S. B. Analytical Study of Neural Network Techniques: som, mlp and classifier-a survey. **IOSR J. Comput. Eng.(IOSR-JCE)**, [S.l.], v.16, n.3, p.86–92, 2014.

WEN, M.; YANG, D.; ROSE, C. Sentiment Analysis in MOOC Discussion Forums: what does it tell us? In: EDUCATIONAL DATA MINING 2014. **Anais...** [S.l.: s.n.], 2014.

WITTEN, I. H. **Text Mining.** 2004.

XIA, C.; FIELDER, J.; SIRAGUSA, L. Achieving better peer interaction in online discussion forums: a reflective practitioner case study. **Issues in Educational Research**, [S.l.], v.23, n.1, p.97–113, 2013.

YOO, J.; KIM, J. Can online discussion participation predict group project performance? Investigating the roles of linguistic features and participation patterns. **International Journal of Artificial Intelligence in Education**, [S.l.], v.24, n.1, p.8–32, 2014.

ZHANG, E.; ZHANG, Y. F-measure. In: **Encyclopedia of Database Systems**. [S.l.]: Springer, 2009. p.1147–1147.

Apêndice

A

Bases de dados utilizadas na Revisão

Computação e Educação	Inteligência Artificial	Processamento de Linguagem Natural
ASEE Annual Conference and Exposition International Conference on Computers and Advanced Technology in Education; IEEE International Conference on Advanced Learning Technologies; International Journal on E-Learning; Journal of Distance Education Technologies; Journal of Education and Information Technologies; Frontiers in Education Conference; Journal of Educational Computing Research; International Conference on E-Learning; Proceedings of the Congress E-learning; Artificial Intelligence in Education; Intelligent Tutoring Systems; International Conference on Artificial Intelligence in Education; International Conference on advances in WebBased Learning; Computers and Education Journal; Conference on Educational Data Mining; Conference on Open Learning and Distance Education; Journal of Educational Technology Systems; Journal of Interactive Learning Research; Conference on Information Technology for Application; Proceedings of the Web-Based Education; Electronic Journal of E-Learning; IEEE Global Engineering Education Conference; CBIE LACLO; DESAFIE; TISE; RENOTE; RBIE.	AI Communications; Journal of Artificial Intelligence Research; Artificial Intelligence; Conference on Artificial Intelligence; International Conference on Web Information Systems Engineering; WWW Conference; European Conference on Artificial Intelligence; Knowledge Discovery and Data Mining; Computers in Human Behavior; International Conference on Data Engineering; International Conference on Machine Learning Applications; BRACIS; Expert Systems with Application; Artificial Intelligence Review.	Propor; Computer Speech and Language; NAACL; IJCAI; ACL; Interspeech; CoNLL; SigDial; Int'l Conf. on Natural Language Generation; EMNLP; COLING; SIGIR; EACL.

Tabela A.1: Lista de Periódicos e Conferências consideradas na revisão