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

Valter dos Santos Mendonça Neto

ANÁLISE AUTOMÁTICA DA PRESENÇA COGNITIVA EM DISCUSSÕES ONLINE ESCRITAS EM PORTUGUÊS

Dissertação de Mestrado

Recife
Novembro de 2018

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

Valter dos Santos Mendonça Neto

ANÁLISE AUTOMÁTICA DA PRESENÇA COGNITIVA EM DISCUSSÕES ONLINE ESCRITAS EM PORTUGUÊS

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

Orientador: *Prof. Dr. Rafael Dueire Lins*
Co-Orientador: *Prof. Dr. Rafael Ferreira Leite de Mello*

Recife
Novembro de 2018

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

M539a Mendonça Neto, Valter dos Santos
Análise automática da presença cognitiva em discussões online escritas em português / Valter dos Santos Mendonça Neto. – 2018.
93 f. : il.

Orientador: Rafael Dueire Lins.
Coorientador: 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.

1. Ensino à distância 2. Aprendizagem 3. Aprendizagem baseada na inquirição 4. Crítica textual I. Lins, Rafael Dueire, orient. II. Mello, Rafael Ferreira Leite de, coorient. III. Título

CDD 004

Dissertação de Mestrado apresentada por **Valter dos Santos Mendonça Neto** ao programa de Pós-Graduação em Informática Aplicada do Departamento de Computação da Universidade Federal Rural de Pernambuco, sob o título **Análise Automática da Presença Cognitiva em Discussões Online Escritas em Português**, orientada pelo **Prof. Dr. Rafael Dueire Lins** e aprovada pela banca examinadora formada pelos professores:

Prof. Dr. Rafael Dueire Lins
Departamento de Computação/UFRPE

Prof. Dr. Rinaldo José de Lima
Departamento de Computação/UFRPE

Profa. Dra. Taciana Pontual da Rocha Falcão
Departamento de Computação/UFRPE

Recife
Novembro de 2018

Aos meus pais, esposa, irmãos, familiares, amigos, professores e, principalmente, a Deus, por acreditar em mim e sempre renovar as minhas forças para superar todas as dificuldades, sem nunca deixar de lembrar-me que nasci vencedor.

Agradecimentos

A Deus por me direcionar em todos os momentos da minha vida e por ter me dado a oportunidade de chegar até aqui na conclusão do mestrado.

À minha família sempre presente dando o suporte que tanto preciso.

À minha esposa, Rafaela Mendonça, pelo companheirismo, amizade, paciência, compreensão, alegria e amor.

Aos meus irmãos em Cristo que sempre me cobraram e incentivaram a concluir o curso.

Aos meus alunos que me estimulam a ser um profissional cada vez melhor, sem eles jamais poderia ter feito este trabalho.

Aos meus orientadores, os professores Rafael Lins e Rafael Ferreira pela colaboração, incentivo e atenção durante os momentos de orientação.

Aos professores e colegas de profissão pelo apoio dado durante o curso.

A todos os professores do programa que contribuíram de forma significativa para minha formação.

Aos meus companheiros de mestrado Vitor Rolim, Alane Lima, Geraldo Gomes, Rafaela Almeida pelos esclarecimentos às minhas dúvidas e pelo apoio nos momentos de dificuldades.

O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo.

—JOSÉ DE ALENCAR

Resumo

Esta dissertação de mestrado apresenta um método que permite a análise automatizada das mensagens trocadas em fóruns *online* de ensino a distância escritas em português brasileiro. Em particular, analisa o problema da codificação de mensagens de discussão para níveis de presença cognitiva, um importante construto do modelo de Comunidade de Investigação amplamente utilizado na aprendizagem *online*. Embora existam técnicas de codificação para presença cognitiva na língua inglesa, a literatura ainda é pobre em métodos para outras línguas, como o português. O método aqui proposto faz uso de um conjunto de 127 características extraídas de diferentes recursos, disponíveis para análise textual através de técnicas de Mineração de Texto, para criar um classificador *random forest* com a finalidade de extrair automaticamente as fases cognitivas. O modelo desenvolvido atingiu 76% de acurácia e o κ de 0,55, o que representa uma concordância moderada, e está acima do nível de puro acaso. Este trabalho também fornece uma análise da natureza da presença cognitiva, observando as características de classificação que foram mais relevantes para distinguir as diferentes fases da presença cognitiva e um estudo comparativo sobre as principais características identificadas nas fases da presença em diferentes contextos.

Palavras-chave: Presença Cognitiva, Modelo de Comunidade de Investigação (CoI), Discussões *Online*, Classificação de Texto

Abstract

This M.Sc. dissertation presents a method that allows the automated analysis of the messages exchanged in online distance-learning forums written in Brazilian Portuguese. In particular, it analyzes the problem of coding discussion transcripts for levels of cognitive presence, an important construct in a widely Community of Inquiry model used in online learning. Although there are coding techniques for cognitive presence in the English language, the literature is still poor in methods for other languages, such as Portuguese. The method proposed here makes use of a set of 127 features extracted from different resources, available for textual analysis through Text Mining techniques, to create a random forest classifier to automatically extract the cognitive phases. The model developed reached 76% accuracy and Cohen's κ of .55, which represents a moderate agreement, and is above the level of pure chance. This paper also provides an analysis of the nature of cognitive presence, observing the classification characteristics that were most relevant to distinguish between the different phases of cognitive presence and a comparative study regarding the main characteristics identified in the phases of presence in different contexts.

Keywords: Cognitive Presence, Community of Inquiry (CoI) model, Online Discussion, Text Classification

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

Sumário

1

Introdução

Apresentada como uma alternativa efetiva para a formação e qualificação de profissionais, a Educação a Distância (EaD) vem crescendo significativamente no Brasil. De acordo com os dados estatísticos do Censo EAD.BR 2017 (ABED, 2018), no que se refere as matrículas em cursos EaD no Brasil, foram contabilizados 7.773.828 alunos em cursos a distância.

Segundo MORAN (2009), a EaD é uma modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorre por meios e tecnologias de informação e comunicação, com estudantes e professores desenvolvendo atividades educativas em lugares ou tempos diversos. Dentre as ferramentas disponibilizadas para esse tipo de modalidade, os Ambientes Virtuais de Aprendizagem (AVAs) têm se destacado nos últimos anos (MCGILL; KLOBAS, 2009), assim como sua representatividade dentre as ferramentas aplicadas no contexto da aprendizagem (CONDE et al., 2012). Nesses ambientes, é possível encontrar vários recursos que podem permitir interações sociais entre alunos, assim como entre alunos e seus professores. Dentre os recursos disponíveis, os fóruns de discussão assíncronos são amplamente utilizados para encorajar a participação dos alunos no curso, respondendo a perguntas e compartilhando conteúdos educacionais (HEW; CHEUNG, 2008).

De acordo com ABAWAJY (2012), os fóruns consistem em espaços para discussões e trocas de ideias sobre assuntos definidos por seus participantes, possibilitando uma experiência de aprendizagem favorável ao processo pedagógico. Assim, as discussões *online* desempenham um papel importante na experiência educacional dos alunos, principalmente em cursos de aprendizagem totalmente *online*, já que nesses há ausência de interações face a face. Dessa forma, analisar como é realizada a interação entre os alunos e deles com seus professores, torna-se muito relevante para o processo educacional, pois o aprendizado efetivo no ensino a distância ocorre quando os participantes envolvidos nesse processo conseguem formar uma comunidade de investigação (GARRISON; ANDERSON; ARCHER, 2000).

O modelo da Comunidade de Investigação (*Community of Inquiry* (CoI)) enfatiza a natureza social da aprendizagem *online* moderna e é um dos modelos pedagógicos mais pesquisados e validados no domínio da educação à distância (GARRISON; ANDERSON; ARCHER, 2000). Ele define três componentes (conhecidos como presenças) que moldam o aprendizado *online*

dos alunos: presença social, presença de ensino e presença cognitiva. Dentre elas, a presença cognitiva, que está relacionada com o desenvolvimento das habilidades de pensamento crítico e reflexões dos aluno, é considerada o componente central do modelo (GARRISON; ANDERSON; ARCHER, 2001).

Para medir/avaliar os processos relacionados a construção do conhecimento nas discussões *online*, a Análise de Conteúdo Quantitativo (QCA) é apresentada como um método amplamente adotado no contexto das três presenças de CoI (ROURKE et al., 2001b; STRIJBOS et al., 2006), fornecendo inferências válidas e confiáveis a partir da análise de dados textuais (BAUER, 2007). O modelo CoI define três esquemas de codificação QCA, um para cada presença, que podem ser aplicados para analisar as mensagens de discussão dos alunos *online*.

A QCA, embora amplamente adotada nas ciências sociais dentro da comunidade de investigação, tem sido usada principalmente para retrospecção e pesquisa após o término dos cursos, sem muito efeito no que se refere ao processo de aprendizagem e aos resultados reais do aluno (STRIJBOS, 2011). DRINGUS; ELLIS (2005), ao analisarem as dificuldades relacionadas a tarefa de avaliação de fóruns de discussão, ressaltam que o avaliador diante do grande número de informações possui capacidade limitada para processar os dados e torná-los significativos para o processo avaliativo. Assim, os autores sugerem que, a partir de técnicas de mineração de textos, o esforço para avaliar os fóruns pode ser reduzido, além de contribuir para a diminuição das inconsistências nas avaliações. Neste sentido, GAŠEVIĆ; KOVANOVIĆ; JOKSIMOVIĆ (2017) mostram os bons resultados da utilização dos métodos de análise automática de texto para análise de aprendizagem, apontando para o potencial destes em tornar a avaliação das presenças CoI CoI mais fácil e menos trabalhosa, visando utilizar o modelo CoI CoI para direcionar intervenções e afetar os resultados de aprendizagem dos alunos (KOVANOVIĆ; GAŠEVIĆ; HATALA, 2014).

Na literatura é possível encontrar estudos promissores para automatizar a avaliação da presença cognitiva (MCKLIN, 2004; CORICH; HUNT; HUNT, 2006; KOVANOVIĆ et al., 2014, 2017; WATERS et al., 2015), mas o foco dessas abordagens tem sido exclusivamente em cursos em inglês, limitando seu uso apenas para o contexto da língua inglesa. Da mesma forma, a disponibilidade de ferramentas que possibilitem a análise de texto para outros idiomas além do inglês é ainda mais limitada, comprometendo de forma significativa a precisão dos sistemas desenvolvidos para esses idiomas. Além disso, os diferentes dados demográficos e o contexto do curso em cursos fora do âmbito da língua inglesa, podem ter um impacto relevante no poder preditivo da análise desenvolvida. É importante frisar que a crescente necessidade de uma educação de alta qualidade nos países em desenvolvimento implica na importância em examinar como tais descobertas podem ser reproduzidas em cursos em outros idiomas além do inglês e como as descobertas resultantes de análises podem ser utilizadas para apoiar estudantes em países não falantes do inglês como língua mãe.

Este trabalho de dissertação apresenta uma proposta para facilitar o processo de análise da presença cognitiva no contexto da língua portuguesa, uma vez que os métodos utilizados são difíceis e trabalhosos, ou seja, realizados de forma manual. A hipótese consiste em utilizar

ferramentas e técnicas de análise de texto amplamente utilizadas na literatura para desenvolver um método para automatizar a avaliação da presença cognitiva de mensagens de discussões *online* escritas em português.

O estudo baseou-se nos trabalhos anteriores realizados em cursos de língua inglesa (KOVANOVIĆ et al., 2014, 2016; WATERS et al., 2015) e adotou uma abordagem de classificação semelhante, embora com algumas modificações devido às diferenças entre as ferramentas de análise de texto em inglês e português. É importante ressaltar que, apesar da abordagem adotada possuir elementos dos estudos supracitados, automatizar a análise da presença cognitiva para língua portuguesa não é uma tarefa trivial, devido as limitações apresentadas quanto as ferramentas e métodos para análise textual disponíveis para língua, comparado a outras línguas, como o inglês. Assim, buscou-se os recursos disponíveis para o português, como LIWC-PT e Coh-Metrix-PT, e que possibilitassem bons resultados (*Word Embedding*, por exemplo).

As principais contribuições deste trabalho são:

- Um modelo para codificar as mensagens dos alunos, em relação as fases da presença cognitiva, escritas em português.
- Análise detalhada da relevância das características propostas no contexto da presença cognitiva e de suas fases.
- Estudo sobre as principais diferenças relacionadas à presença cognitiva em dois contextos diferentes.

1.1 Objetivos

1.1.1 Objetivo geral

O objetivo desta pesquisa é apresentar um método para análise automática da presença cognitiva em discussões assíncronas escritas em português utilizando técnicas de mineração de texto.

1.1.2 Objetivos específicos

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

- Verificar os métodos de análise da presença cognitiva em discussões *online* existentes na literatura;
- Analisar técnicas de mineração de texto que podem auxiliar na identificação da presença cognitiva;
- Criar um modelo para identificar a presença cognitiva em mensagens escritas em Português oriundas de ambientes *online*;

- Avaliar o modelo em uma base de dados composta por postagens extraídas de disciplinas ministradas na EaD;
- Realizar uma análise comparativa das características da presença cognitiva em duas turmas de cursos EaD de contextos diferentes.

1.2 Organização do trabalho

Este trabalho está dividido da seguinte forma: o Capítulo 2 expõe a fundamentação teórica, em que são apresentados os conceitos fundamentais para o entendimento do trabalho realizado. No Capítulo 3 são citados os principais trabalhos relacionados a este estudo, mostrando a aplicação da presença cognitiva e de suas respectivas fases para a construção do conhecimento.

O Capítulo 4 detalha a metodologia utilizada para a construção do método proposto nesta dissertação, com o intuito de automatizar a análise da presença cognitiva utilizando Mineração de Texto. O Capítulo 5 mostra os resultados obtidos com a aplicação do método em um *corpus* construído com mensagens de fóruns de discussões *online*.

O Capítulo 6 mostra as discussões geradas pela análise dos resultados. Por fim, o Capítulo 7 apresenta as considerações finais e contribuições da pesquisa, artigos submetidos/aceitos, as limitações da pesquisa e os trabalhos futuros.

2

Fundamentação Teórica

O presente capítulo apresenta os conceitos básicos necessários para o entendimento desta dissertação e foi dividido em três seções. Na primeira irá abordar a Comunidade de Investigação (CoI) e as três dimensões que a compõem. A segunda seção apresenta os fundamentos básicos aplicados neste trabalho relacionados a mineração de textos. Por fim, na terceira seção são expostas as ferramentas aqui utilizadas para o desenvolvimento da pesquisa.

2.1 O Modelo de Comunidade de Investigação (CoI)

Um modelo conceitual de análise focado no processo social de construção conjunta e colaborativa do conhecimento em ambientes de comunicação assíncrona baseada em texto foi proposto na referência GARRISON; ANDERSON; ARCHER (2000). Esse modelo surgiu de um estudo para investigar os efeitos na qualidade do processo de aprendizagem e os seus resultados, mediante o ensino baseado na comunicação mediada por computador, em textos-base de ambientes virtuais de aprendizagem de nível superior. Para tais autores, a educação virtual permite elevados níveis de interação entre professores e alunos, principais participantes do processo educacional, e que, quando integrados em um trabalho conjunto, possibilitam o desenvolvimento de capacidades cognitivas mais complexas, gerando aprendizagens mais ricas, coerentemente organizadas e com objetivos constantes de exploração e aprofundamento.

Assim, o modelo de Comunidade de Investigação (CoI - *Community of Inquiry*) é um modelo teórico amplamente adotado para orientar a pesquisa e a prática da aprendizagem *online*, em que uma comunidade de investigação se constitui por meio de três dimensões (Figura 2.1) imprescindíveis para que haja uma experiência educacional de sucesso: a presença social, a presença de ensino e a presença cognitiva, que se influenciam mutuamente.

A seguir, será detalhada cada uma das dimensões, bem como suas fases e indicadores utilizados para detalhar as presenças.

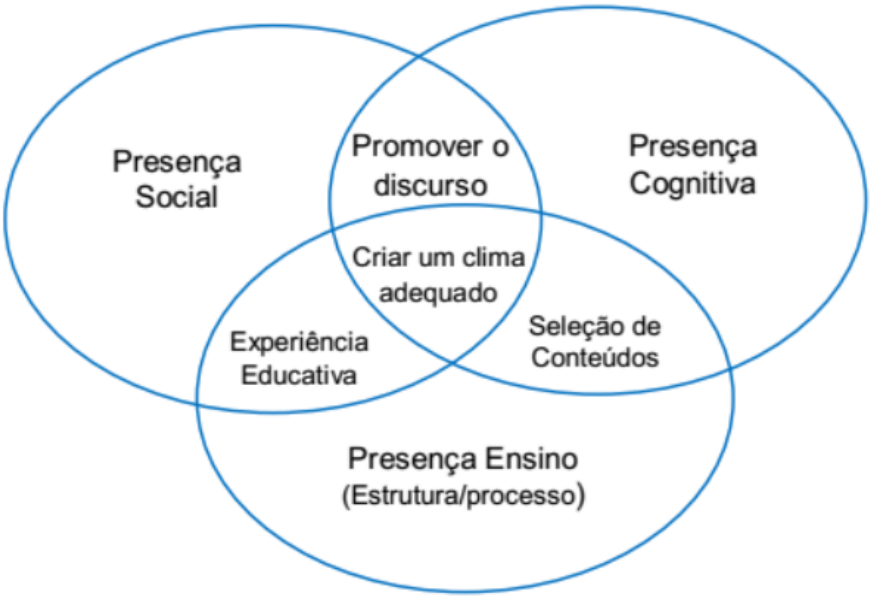

Figura 2.1: Modelo de uma comunidade de investigação
(adaptado de GARRISON; ANDERSON; ARCHER (2000))

2.1.1 Presença Social

Segundo GARRISON; ANDERSON; ARCHER (2000), a presença social consiste na capacidade que os participantes de uma comunidade de investigação têm em se projetar social e emocionalmente como pessoas 'reais' mediante os instrumentos de comunicação em uso. Os autores afirmam que a presença social deve ir mais além do simples estabelecimento da presença sócio-emocional e de relações pessoais. A coesão do grupo e da comunidade está relacionada com o quanto os participantes possuem objetivos em comum e demonstrem a existência de amizade entre eles.

A presença social constata-se na resolução de atividades colaborativas, entretanto, para que, em uma comunidade *online*, os relacionamentos sejam desenvolvidos, é necessário tempo para que se alcance um nível de conforto, de confiança e um sentimento de pertencimento.

Assim, ela se torna necessária para o desenvolvimento da presença cognitiva, pois fornece suporte no trabalho de facilitar o processo de pensamento crítico, uma vez que os participantes de uma comunidade precisam estar socialmente envolvidos e emocionalmente motivados para interagir no contexto virtual. Além disso, mostra-se um elemento importante para o sucesso da experiência educacional (GARRISON; ANDERSON; ARCHER, 2000).

De acordo com CRUZ (2013), no contexto educacional, o objetivo da presença social é proporcionar a existência de condições para que a investigação ocorra e que as interações se mostrem de qualidade, com natureza reflexiva e um bom encadeamento de ideias. Para isso, GARRISON; ANDERSON; ARCHER (2000) desenvolveram um esquema de classificação da presença social que resultou em três categorias fundamentais: Expressões de Afetividade, Comunicação Aberta e Coesão de Grupo.

A primeira categoria compreende as expressões de emoção, sentimentos, crenças e valores, apresentados através de repetições de pontuação, frase inteira escrita em caixa alta ou ainda uso de *emoticons*. Além disso, considera-se também o senso de humor das pessoas, através de piadas, ironias ou mesmo sarcasmos. E ainda as revelações pessoais, em especial, sobre aspectos não relacionados ao curso em si, como gostos pessoais, particularidades da vida pessoal, do trabalho e até expressões de vulnerabilidades.

A Comunicação Aberta é descrita como troca de comunicação recíproca e respeitosa. GARRISON; ANDERSON; ARCHER (2000) citam a consciência mútua e reconhecimento da contribuição, como exemplos desta categoria. A consciência mútua contribui moldando as atividades de aprendizagem de cada participante e o reconhecimento é o processo que serve para estimular o desenvolvimento e a manutenção da troca de relacionamentos. Além disso, os autores afirmam que a interatividade se revela em ambientes assíncronos quando se utiliza o recurso de respostas para postar mensagens; através da citação de mensagens de outros participantes, no momento em que um comentário é direcionado a uma pessoa em particular; e por meio de referências explícitas às mensagens de outros participantes.

Os autores exemplificam a terceira categoria, a Coesão de Grupo, por meio de práticas que constroem e sustentam o senso de envolvimento do grupo. O princípio da categoria é que a qualidade do discurso é facilitada e otimizada, quando os estudantes se sentem como membros do grupo e não como indivíduos isolados. Os autores apresentam os vocativos, o emprego de pronomes inclusivos e as saudações, como importantes expressões de coesão, uma vez que possibilitam a participação e a afinidade e apresentam-se como indício de uma tentativa de estreitar o relacionamento com o destinatário.

A Tabela 2.1 apresenta as categorias e os indicadores sugeridos por ROURKE et al. (2001a), que auxiliam na identificação, em discursos escritos, da presença social. Dessa forma, a presença social apresenta-se como um apoio contextual para as relações de afetividade, de interação e de coesão.

A seguir, é apresentada a presença de ensino, apontada pelos autores do modelo como fundamental para constituir e manter uma comunidade educacional de investigação, além de integrar com sucesso as demais presenças através da comunicação baseado em texto.

Categorias	Indicadores	Definição
Expressões de Afetividade	Expressar emoções	Expressões convencionais ou não convencionais de emoções, incluindo pontuação repetida, uso de maiúsculas, símbolos (*emoticons*).
	Uso de humor	Piadas, ironias, sarcasmo.
	Informações sobre si	Apresenta detalhes da vida fora da classe ou expressa alguma vulnerabilidade.
Comunicação Aberta	Continuar uma conversa	Usar o comando "responder" do software, em vez de começar uma conversa nova.
	Citar mensagens de outros	Usar os recursos do software para citar as mensagens inteiras dos outros, cortar e colar partes de outras mensagens.
	Fazer referência às mensagens de outros	Fazer referência ao conteúdo de outras mensagens.
	Fazer perguntas	Os alunos fazem perguntas a outros alunos ou ao professor.
	Saudar, expressar apreço, expressar concordância	Elogiar os outros ou os seus comentários; Expressar concordância com outros ou com o conteúdo de outras mensagens.
	Concordar	Concordar com outros ou com o conteúdo das suas mensagens.
Coesão de Grupo	Vocativos	Referir-se aos participantes pelo nome.
	Refere-se ao grupo usando pronomes inclusivos	Refere-se ao grupo por "nós", "nosso grupo".
	Saudações	Comunicação social: saudações, despedidas.

Tabela 2.1: Dimensões, categorias e indicadores da Presença Social (adaptado de ROURKE et al. (2001a))

2.1.2 Presença de Ensino

ANDERSON et al. (2001) definem a presença de ensino ou pedagógica como a ação de conceber, facilitar e orientar os processos cognitivo e social com o propósito de obter resultados

de aprendizagem significativos e de valor educacional. Para GARRISON; ANDERSON (2003), a participação do professor na presença de ensino é um elemento vital na experiência educativa, principalmente no ensino a distância, considerado um grande desafio quando comparado ao ensino presencial. Para tais autores, o professor precisa não só ser conhecedor do assunto do curso, mas também dever ter estratégia educacional, além de ser um facilitador social. Assim, cabe a ele o papel de gerenciar o ambiente e facilitar a aprendizagem do aluno, no sentido de proporcionar um espaço mais acessível, confortável e seguro para que se tenha as condições necessárias para o desenvolvimento da aprendizagem do estudante. Dessa forma, para a construção colaborativa do conhecimento, a presença de ensino é composta por três categorias: Desenho Instrucional e Organização; Facilitação do Discurso e Instrução Direta.

De acordo com ANDERSON et al. (2001), a primeira categoria, o Desenho Instrucional e Organização, está relacionada como o planejamento e projeto dos aspectos de estrutura, processo, interação e avaliação do curso *online*. Essas tarefas, desempenhadas pelos instrutores, requerem deles explicitação e transparência. GARRISON; ARBAUGH (2007) destacam como exemplo de atividades que compõem esta categoria a recriação de apresentações *PowerPoint* e de anotações de aulas; a elaboração de mini-lições em áudio/vídeo, indicações pessoais relacionadas ao material do curso; a produção de um cronograma para atividades individuais e de grupo e fornecendo orientações quanto ao uso do meio de forma eficaz. Essas são atividades particularmente importantes, pois o sucesso dos cursos *online* depende muito de uma estrutura clara e consistente, que dê o suporte necessário aos instrutores envolvidos e às discussões dinâmicas.

A segunda categoria, a Facilitação do Discurso, refere-se ao meio pelo qual os estudantes estão comprometidos na interação e na construção do conhecimento mediante os recursos oferecidos pelo curso. Além disso, esta categoria deve ser coerente com os resultados que apoiam a importância da interação dos alunos para a efetividade da aprendizagem *online* (GARRISON; ARBAUGH, 2007). GOMES; PESSOA (2012a) reforçam que ao facilitar o discurso, reconhece-se o papel da comunidade de aprendizagem como impulsor da construção de significado, assim como na promoção da compreensão dos elementos que fazem parte dessa comunidade. Portanto, cabe ao discente a responsabilidade de rever e comentar as respostas dos alunos; sugerir novos questionamentos e fazer comentários, com o objetivo de conduzir a discussão na direção desejada, de forma eficaz; estimular alunos menos ativos e limitar o número de participações daqueles que dominam as discussões, impedindo que esses prejudiquem o processo de aprendizagem em grupo (ANDERSON et al., 2001).

Na terceira e última categoria desta dimensão, a Aprendizagem Direta, os professores proveem a liderança intelectual e acadêmica, e compartilham seu conhecimento sobre o assunto com os alunos. Considera-se que o papel do professor vá além de simples facilitador do conteúdo, eles devem analisar os comentários, inserir fontes de informação, direcionar as discussões para um sentido desejado e instruir os alunos para a aquisição de novos conhecimentos (ANDERSON et al., 2001). Os autores da referência GARRISON; ARBAUGH (2007) reforçam que a responsabilidade do professor está focada em facilitar a reflexão e o discurso, através da apresentação

do conteúdo, usando os meios disponíveis para a avaliação e *feedback*. Além disso, afirmam que para eficácia desse tipo de comunicação, a presença social do professor é fundamental, exigindo dele conhecimentos técnicos e pedagógicos para relacionar as contribuições, identificar conceitos equivocados e introduzir o conhecimento consolidado em referências tais como livros, artigos científicos e recursos educativos disponíveis na *Web*.

A Tabela 2.2 apresenta uma síntese dos três quadros descritivos elaborados por ANDERSON et al. (2001), referentes as categorias da presença de ensino e seus indicadores e descritivos.

Categorias	**Indicadores**	**Definição**
Desenho Instrucional e Organização	Estabelecer currículos, tecnologia e ferramentas	Fase de planejamento para concepção do ambiente, processo de desenvolvimento das atividades, da avaliação e formas de interação.
	Desenhar métodos	Criação de estratégias que visem subsidiar os membros na aprendizagem, como comentários personalizados do professor, dos colegas, tutoriais, minipalestras, entre outros.
	Estabelecer prazos	Estipular prazos para a realização das atividades.
	Utilizar as mídias de forma eficaz	Orientação quanto ao uso dos recursos disponíveis, por exemplo explicitando como devem ocorrer as postagens em fóruns.
	Estabelecer a etiqueta da *Web*	Dicas para uso apropriado das mídias: dar instruções acerca do tamanho das mensagens, e tipo de linguagem na interação.

Facilitação do Discurso	Identificar áreas de acordo/desacordo	Identificar discordância de Opiniões/Conflito Cognitivo.
	Busca por consenso/ compreensão	Encontro de pontos que coincidem quando duas opiniões aparentemente contrárias estão sendo expressas, por exemplo.
	Encorajar, reconhecer ou reforçar as contribuições dos alunos	O professor ou os alunos apoiam e incentivam a participação, comentando e estimulando as respostas dos colegas.
	Estabelecimento de clima propício para a aprendizagem	Favorecer um ambiente acolhedor e que sobretudo respeite as opiniões de todos para a concretização da aprendizagem.
	Encorajar os participantes, promover a discussão	Questionar, interrogar e suscitar possíveis respostas dos alunos do fórum.
	Avaliar a eficácia do processo	Fornecer feedback construtivo dos comentários, tendo em vista o objetivo das discussões.
Aprendizagem Direta	Apresentar conteúdos ou questões	Facilitar a aprendizagem. O professor ou os alunos compartilham seus conhecimentos com o grupo.
	Focar a discussão em assuntos específicos	Dirigir a atenção para determinados conceitos ou assuntos que são necessários para moldar ou alcançar a construção do conhecimento.
	Resumir a discussão	Sintetizar as ideias principais das contribuições dos alunos.
	Confirmar a compreensão por meio de feedbacks avaliativos e exploratórios	Comentar a participação dos membros.
	Diagnosticar falhas de compreensão	Comentários feitos pelo professor ou pelos alunos sobre as atividades da aprendizagem, que mostrem possíveis equívocos.
	Introduzir conhecimento de diversas fontes (livros, artigos)	Fornecimento de diversas fontes de pesquisa para que os alunos possam aprofundar seus conhecimentos sobre o assunto abordado.
	Dar resposta às questões técnicas	Instruções diretas sobre o funcionamento do sistema, manipulação de software e operação de outras ferramentas ou recursos.

Tabela 2.2: Dimensões, categorias e indicadores da Presença de Ensino (adaptado de ROURKE et al. (2001a) e LISBÔA; COUTINHO (2012))

A presença de ensino é um fator fundamental para a satisfação dos alunos, a percepção da aprendizagem e o sentimento de comunidade. Além disso, quando integrada à presença social, fornece suporte à presença cognitiva.

2.1.3 Presença Cognitiva

A presença cognitiva é a medida pela qual os estudantes são capazes de construir e confirmar o significado mediante uma reflexão e discurso sustentado (GARRISON; ANDERSON; ARCHER, 2001). Esta presença visa a promoção da análise, a construção e a confirmação do significado e da compreensão dentro de uma comunidade de aprendizagem por meio da reflexão e do discurso nos fóruns (ARAUJO; OLIVEIRA NETO, 2013).

ARAUJO (2014) afirma que, no modelo de investigação, a presença cognitiva é um componente fundamental, pois fornece evidências da qualidade das discussões e possibilita uma avaliação processual da organização do pensamento crítico e das reflexões.

Com o intuito de descrever e compreender a presença cognitiva em processos educacionais e o desenvolvimento do processo de pensamento crítico, GARRISON; ANDERSON; ARCHER (2001) definiram um modelo prático de investigação (Figura 2.2) dividido em quatro fases.

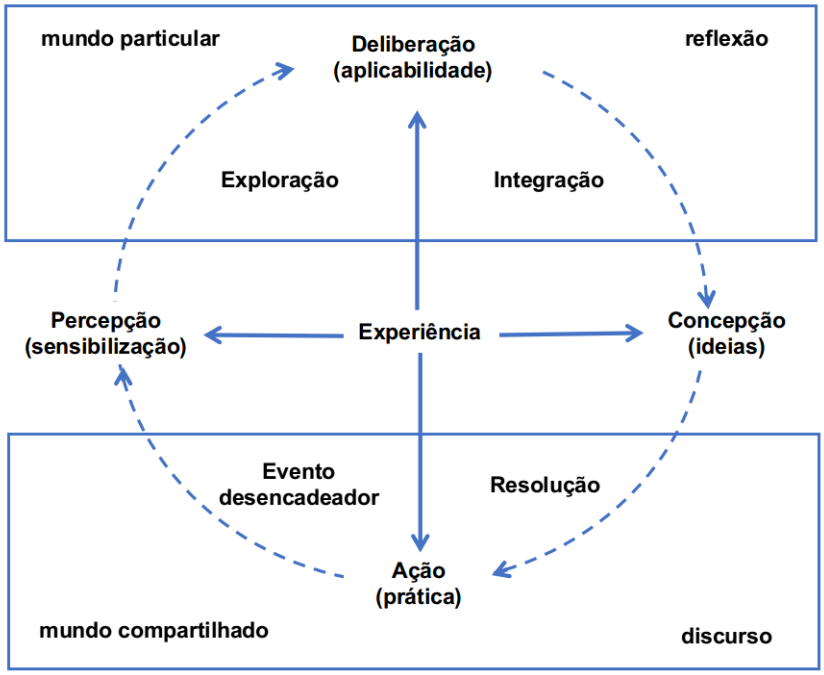

Figura 2.2: Modelo Prático de Investigação
(adaptado de GARRISON; ARBAUGH (2007))

Primeira fase: Evento desencadeador (*Triggering Event*) - Representa o início das discussões e reflete a fase inicial do processo de investigação crítica. Nesta fase, um problema ou uma questão, proveniente da experiência ou da análise do contexto educacional, é apresentada no fórum de discussões. Em fóruns de discussão mais democráticos, além do professor ou

moderador, o aluno também pode lançar um evento desencadeador, expondo uma dúvida, ou um desafio de aprendizagem. Entretanto, os autores reforçam a importância da atuação do professor neste início de interação. O direcionamento adequando nesse momento inicial evita distorções da discussão, uma vez que, dependendo do tipo de evento desencadeador, o diálogo podem ser desviados objetivo da atividade.

Segunda fase: Exploração (*Exploration*) - Reporta o momento em que os alunos alternam entre o mundo particular, reflexivo e a exploração social das ideias. Nesta fase, os alunos devem perceber ou compreender a natureza do problema para, em seguida, começar a explorar fontes para coletar informações mais relevantes. Esta exploração acontece através de movimentos dentro da própria comunidade de investigação que estimula a reflexão crítica e o diálogo. O fim desta fase é caracterizado por alunos mais seletivos quanto ao que é relevante para a questão ou problema. A fase de exploração é caracterizada pela crescente divergência, o questionamento, *brainstorming*, e a troca de informações.

Terceira Fase: Integração (*Integration*) - É caracterizada como a fase em que os alunos constroem significados baseados em ideias desenvolvidas durante a exploração. Durante a passagem da fase exploratória para a de integração, os alunos iniciam o processo de avaliação da aplicabilidade para o problema descrito ou evento desencadeador em questão. De acordo com os autores, detectar evidências da integração de ideias e construção de significados, mostra-se uma tarefa difícil nesta fase. Assim, é a partir da comunicação estabelecida na comunidade que se pode inferir a construção cognitiva. Além disso, alertam para a necessidade da presença ativa de ensino atuando no diagnóstico de conceitos errados, fornecendo perguntas, comentários, informações adicionais, com o intuito de garantir o desenvolvimento cognitivo e modelar o processo de pensamento crítico.

Quarta Fase: Resolução (*Resolution*) - Nesta fase ocorre a construção do conhecimento e sua possível aplicação em problemas práticos. Em um contexto não educacional, essa fase representaria a implementação de uma solução proposta ou teste da hipótese por meio de uma aplicação prática. Entretanto, em um contexto educacional esta abordagem mostra-se mais complexa, pois requer outros agentes dispostos a realizar os testes das hipóteses e a construir um consenso entre os participantes da comunidade de investigação. Para os autores, progredir para a quarta fase exige do aluno consciência sobre as expectativas e as oportunidades para aplicar o conhecimento recém-criado. No campo educacional, o final desta fase pode requerer que o estudante passe para um novo problema ou novos eventos desencadeantes. Assumindo-se que eles já adquiriram conhecimento relevante, eles são levados a aplicar esses conhecimentos na resolução de novas situações-problema, fazendo com que o ciclo lógico do modelo de investigação se inicie novamente.

A Tabela 2.3 apresenta as quatro fases da presença cognitiva, com seus respectivos indicadores e suas definições, adaptados da proposta de GARRISON; ANDERSON; ARCHER (2001), para auxiliar na identificação da presença em uma comunidade de aprendizagem virtual.

Categorias	Indicadores	Definição
Evento desencadeador	Reconhecer o problema	Apresentar uma informação sobre o assunto abordado que conduz a uma questão.
	Sensação de confusão ou perplexidade	Fazer perguntas; postar comentários que conduzam a discussão para novas direções.
Exploração	Divergência na comunidade *online*	Discordância de ideias, mas sem sustentação teórica.
	Divergência numa simples mensagem	Muitas ideias ou temas diferentes apresentados em uma única mensagem.
	Troca de Informações	Narrativas/descrições/fatos pessoais (não usados como argumento para sustentar um posicionamento ou conclusão).
	Sugestões para consideração	Comentários que denotem alguma restrição ou discordância de ideias (caracteriza explicitamente a mensagem como exploração). Ex: Isso parece correto?; Eu discordo; Estou muito enganado?
	Brainstorming	Acrescenta novas ideias, mas não as defende teoricamente, e nem tampouco desenvolve-as de forma sistematizada.
	Conclusões	Fornece sugestões e opiniões, mas não as fundamenta.
Integração	Convergência entre membros do grupo	Faz referência aos comentários dos colegas, concordando com suas ideias, acrescenta novas ideias e novos significados.
	Convergência na mesma mensagem	Tentar justificar, desenvolver e defender hipóteses.
	Ligar ideias, sintetizar	Integrar informação de várias fontes: livros, artigos, experiências pessoais.
	Criar soluções	Caracterização explícita de uma mensagem como uma solução pelo próprio participante.
Resolução	Aplicar ao mundo real	Aplicação prática dos conhecimentos adquiridos.
	Testar e defender soluções	Estabelecer relações com outros conhecimentos já existentes; adquirir competência de análise e reflexão crítica e ter poder de argumentação para sustentar as ideias que defende no que diz respeito ao problema proposto.

Tabela 2.3: Dimensões, categorias e indicadores da Presença de Ensino (adaptado de GARRISON; ANDERSON; ARCHER (2001) e LISBÔA; COUTINHO (2012))

É importante destacar o fato das três dimensões estarem interligadas, conforme apresentado na Figura 2.1, cada uma servindo de suporte para que a outra possa acontecer. GARRISON; ANDERSON; ARCHER (2001) afirmam que a presença social é essencial para que a presença cognitiva se consolide, pois prepara os alunos para aprenderem de maneira colaborativa, para discutirem ideias utilizando argumentos consistentes e dentro de princípios éticos, proporcionando dessa forma, a reflexão crítica, e por fim, a aprendizagem. Por sua vez, a presença de ensino tem como objetivo promover um ambiente favorável à partilha de saberes e construção de significados.

Segundo GARRISON; ARBAUGH (2007), a presença social constrói as bases para um discurso de nível superior; e a estrutura, organização e liderança quando associadas à presença de ensino, proporcionam um ambiente onde a presença cognitiva pode ser desenvolvida.

2.2 Mineração de Texto

A Mineração de Textos (MT) pode ser definida como um processo intensivo de conhecimento, no qual um usuário interage com uma grande quantidade de documentos, utilizando ferramentas para analisar os mesmos. O objetivo da MT é a extração de informações relevantes, através da identificação e exploração de padrões interessantes em bases textuais não estruturadas, ou semi-estruturadas (FELDMAN; SANGER, 2007). Segundo CHOWDHURY (2003), áreas como Aprendizagem de Máquina (AM) e Processamento de Linguagem Natural (PLN) contribuem para este processo de extração de conhecimento. A AM tem como objetivo estudar meios que possibilitem aos computadores adquirir conhecimento e, a partir de um conjunto de dados de treinamento, estimar saídas para dados desconhecidos (NILSSON, 1996). PLN estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais, ou seja fornecem aos computadores a capacidade realizar tarefas como: reconhecer o contexto, fazer análise sintática, semântica, léxica e morfológica, criar resumos, extrair informação, interpretar os sentidos, analisar sentimentos e até aprender conceitos com os textos processados (VIEIRA; LOPES, 2010).

A importância da MT está relacionada com a grande quantidade de dados produzidos e disponibilizados em formato digital na rede mundial de computadores nos dias atuais (TURNER et al., 2014), pois parte desses dados está no formato textual, como *e-mails*, relatórios, boletins, artigos, registros de pacientes e conteúdo de páginas *Web* (ROSSI, 2015). Realizar as tarefas de organizar, analisar, e extrair o conhecimento embutido nesses dados manualmente, é uma tarefa muito difícil de ser executada sem o auxílio de técnicas computacionais.

Em geral, para a extração do conhecimento através da mineração de textos, as seguintes etapas são realizadas: coleta de dados, pré-processamento, extração do conhecimento e avaliação e interpretação dos resultados (ARANHA; PASSOS, 2006). Essas etapas e as tarefas abordadas neste trabalho serão detalhadas nas próximas subseções.

2.2.1 Coleta de Dados

A coleta de dados é a etapa inicial e tem como objetivo construir a base de dados textual que será utilizada no processo de Mineração de Texto. Na literatura, esta base é conhecida como *corpus* e o conjunto destes é chamado de corpora.

Para a construção do *corpus*, uma grande quantidade de documentos é previamente coletada. Esses documentos podem ser coletados de várias fontes, como redes sociais, *emails*, campos textuais em banco de dados, páginas da *Web*, *chats* e fóruns de ambientes *online*, de acordo com a relevância dos mesmos ao domínio de estudo.

É uma etapa importante, que exige esforço e cuidado, a fim de se obter material de qualidade e que sirva de matéria-prima para a aquisição de conhecimento.

2.2.2 Pré-Processamento

Em muitas aplicações de MT, os dados da base precisam ser submetidos a técnicas de pré-processamento antes de aplicá-los a algum método computacional. Assim, esta fase tem por finalidade preparar o conjunto de dados textuais, coletados na etapa anterior, para servir de entrada para fase de extração de conhecimento.

Existem diversas técnicas que podem ser utilizadas e até mesmo combinadas, como tokenização, que consiste em identificar e separar cada unidade existente em um texto em *tokens* (pode ser uma palavra, um número, um sinal) (HABERT et al., 1998); remoção de *stopwords*, que remove termos com pouca significância em um texto (*stopwords*), tais como: artigos, preposições e conjunções (LO; HE; OUNIS, 2005) e *stemming*, que reduz as palavras de um texto para sua forma gramatical inicial, ou seja, para sua raiz (*stem*) (ex. “perdeu” e “perdemos” possui o radical “perd”) (RAMASUBRAMANIAN; RAMYA, 2013).

2.2.3 Extração do conhecimento

É a mineração propriamente dita. Nesta etapa tem-se a aplicação de algoritmos de extração automática de conhecimento na base de dados analisada, com o intuito de procurar informações desconhecidas até o momento, mas que possam ser úteis para o contexto de um problema específico. As técnicas utilizadas para a realização desta etapa na abordagem proposta são listadas a seguir.

2.2.3.1 Classificação de Texto

Segundo SEBASTIANI (2002), a classificação ou categorização textual tem como objetivo determinar se um documento específico pertence a uma ou mais classes. A tarefa é descrita como uma função: $\phi : DxC \rightarrow \left\{T, F\right\}$, onde $D = \left\{d_1, d_2, ...d_{|D|}\right\}$ é o conjunto que representa o domínio de documentos e $C = \left\{C_1, C_2, ...C_{|C|}\right\}$ é o conjunto pré-definido de classes (categorias). O valor T atribuído a $< d_j, c_i >$ indica uma decisão de classificar d_j como c_i, e F

indica que d_j não é classificado como c_i. A função que descreve como documentos devem ser categorizados é chamada de classificador.

Para realizar a classificação automática de textos, uma das formas existente é através de algoritmos de aprendizado de máquina. O objetivo desses algoritmos é aprender, generalizar, ou ainda extrair padrões ou características das classes definidas segundo os documentos textuais e rótulos (identificadores de classe) dos documentos indicados por um usuário ou especialista de domínio (ROSSI, 2015).

Dentre as categorias de algoritmos de aprendizado de máquina para realizar a classificação automática existentes, destaca-se o aprendizado supervisionado. Nesta abordagem, utiliza-se uma série de exemplos (instâncias), já classificados por um especialista de domínio (conjunto de treinamento), para induzir um modelo que seja capaz de classificar novas instâncias, com base no aprendizado obtido em uma fase de treinamento. Portanto, esse tipo de aprendizado também é conhecido por aprendizado indutivo supervisionado. Além disso, é importante que exista um conjunto de dados de treinamento de qualidade, para que o modelo gerado seja capaz de predizer novas instâncias de forma eficiente.

2.2.3.2 Extração de Características

Para LEE (2000) o aprendizado de uma máquina consiste na escolha ou adaptação de parâmetros da representação do modelo, dentro do paradigma escolhido, de forma que possibilite ao modelo criado classificar corretamente os exemplos conhecidos, assim como, os novos exemplos. Esses exemplos são representados por registros de características. As características consideradas potencialmente significativas para um dado contexto são também referenciadas como *features*, atributos, variáveis (WEISS; KULIKOWSKI, 1991). Assim, na classificação de textos, um documento de texto é interpretado como um conjunto de características. Essas características fornecem as informações necessárias para descrever e categorizar um determinado documento, sendo assim, essenciais para o aprendizado do classificador.

Para SHAH; PATEL (2016), extração de características é o processo de gerar novas características a partir de características existentes nos documentos. O objetivo da técnica é criar novas características que melhor representem os padrões das instâncias de documentos.

Os autores afirmam ainda que, o processo pode ser realizado através de métodos matemáticos ou a através da observação e raciocínio de um especialista sobre as características que podem ser geradas a partir das características já existentes no conjunto de documentos. Assim, a extração de características é um processo chave para o reconhecimento de padrões, de forma que quanto mais relevantes as características utilizadas para descrever os documentos, melhor será o classificador, mais confiável será a classificação.

Para realização desta tarefa, existem várias técnicas, como **N-gramas** - uma subsequência de n itens construídos a partir uma sequência de itens presentes em um texto, que podem ser unigramas ($n = 1$), bigramas ($n = 2$) e trigramas ($n = 3$) (BROWN et al., 1992); ***Part-of-Speech Tagger*** - analisa cada palavra ou termo contido em uma sentença para, em seguida, atribuir

a cada item uma classe gramatical (VIEIRA; LIMA, 2001); e ***dependency parsing*** - análise das dependências entre palavras individuais (MARNEFFE et al., 2006), e ferramentas, como ***DBpedia Spotlight*** - ferramenta de anotação semântica (MENDES et al., 2011) e ***Linguistic Inquiry and Word Count*** **(LIWC)** - dicionário léxico (TAUSCZIK; PENNEBAKER, 2010).

2.2.3.3 *Word embedding*

De modo geral, *word embeddings*, ou simplesmente *embeddings*, consiste em vetores reais distribuídos em um espaço multidimensional induzidos através de aprendizado não-supervisionado (TURIAN; RATINOV; BENGIO, 2010). Cada dimensão do vetor de uma palavra representa uma característica, de modo a obter distribuidamente propriedades semânticas, sintáticas ou morfológicas dessa palavra (COLLOBERT et al., 2011).

O trabalho de BENGIO et al. (2003) propôs pela primeira vez o uso de *word embedding* no modelo denominado *Neural Network Language Model* (NNLM), em que empregaram redes neurais para aprender automaticamente representações vetoriais, baseado nas ideias de representação distribuída proposta por (HINTON et al., 1986).

COLLOBERT; WESTON (2008) apresentam uma abordagem em que um modelo de língua baseado em uma rede neural tenta prever a próxima palavra do contexto, dadas as anteriores. As representações das palavras são aprendidas através do ajuste de pesos da rede usando o algoritmo de *backpropagation* (GOODFELLOW et al., 2016). As palavras que aparecem com frequência em um mesmo contexto adquirem vetores similares, sendo aprender essa noção considerado o processo de treinamento da rede.

O modelo de *embeddings* Word2Vec, proposto por MIKOLOV et al. (2013), é um dos mais conhecido para esse tipo de representação. A técnica utilizada para geração do modelo baseou-se no princípio do modelo de língua neural apresentado por COLLOBERT; WESTON (2008). Essa técnica consiste em modelos que permitem o treinamento de um *word embedding*, a partir de um grande volume de texto, para prever uma palavra baseado no seu contexto (*Continuous Bag-of-Words* - CBOW) ou, por meio de uma determinada palavra, prever o contexto no qual ela esta inserida (*Skip-gram*).

Em resumo, o propósito do *word embeddings* consiste em transformar palavras em números, de maneira que algoritmos como *deep learning* possam, então, ingerir e processar, para formular uma compreensão da linguagem natural.

Os vetores obtidos através dos modelos de *embeddings* permitem também representar a similaridade entre palavras. Considerando determinadas métrica de distância, palavras consideradas próximas no espaço vetorial tendem a possuir significados similares. Essa similaridade pode ser determinada com base em duas perspectivas. Na primeira, a palavra é considerada segundo sua representação em língua natural, ou seja como uma sequência de letras (caracteres). Nesta perspectiva, podem ser empregadas medidas de similaridade de *string* e de similaridade fonética. Na segunda perspectiva, a palavra é representada como um vetor numérico, ou *word embedding*. Nesse caso, medidas de distância em espaços vetoriais como a similaridade do cosseno (calcula

o cosseno do ângulo formado entre dois vetores – quanto maior seu valor, maior a similaridade) podem ser aplicadas (BERTAGLIA; NUNES, 2016).

2.2.3.4 Entidades Nomeadas

O termo Entidade Nomeada (EN) foi apresentado por GRISHMAN; SUNDHEIM (1996), buscando-se reconhecer unidades de informação, como nomes (pessoas, organizações e locais) e expressões numéricas (tempo, datas, moedas e expressões percentuais). A identificação das referências a estas entidades em textos foi caracterizada como uma subtarefa importante de Extração de Informações (EI) (aplicação de técnicas para a obtenção de informações específicas de um domínio a partir de conteúdos textuais em linguagem natural) e foi denominada de Reconhecimento de Entidades Nomeadas e Classificação (RENC) (NADEAU; SEKINE, 2007).

Desta forma, EN são termos que apresentam um ou mais designadores rígidos, num determinado texto. Os mais comuns tipos de entidades são substantivos próprios (nomes de pessoas, organizações, entidades locais), substantivos temporais (datas, tempo, dias, anos e meses) e entidades numéricas (medições, percentagens e valores monetários).

O Reconhecimento de Entidades Nomeadas (REN) é definida como uma tarefa cujo objetivo consiste em identificar as entidades nomeadas, bem como sua posterior classificação, atribuindo a essas entidades uma categoria semântica. Segundo AMARAL et al. (2014), é uma técnica amplamente utilizada em PLN e consiste da identificação de nomes de entidades-chave presentes na forma livre de dados textuais. Em um sistema de extração de entidades nomeadas, a entrada é um texto em sua forma livre, e sua saída um conjunto de textos anotados, ou seja, uma representação estruturada a partir da entrada de um texto não estruturado. Por exemplo, dado o texto entrada: “Valter estuda na UFRPE” – efetuando a extração das entidades nomeadas do texto exemplo, tem-se: [Valter] e [UFRPE], respectivamente, entidades cujas categorias são: Pessoa e Organização.

2.2.3.5 Balanceamento de Dados

Normalmente, os dados do mundo real são desbalanceados e estão contidos em diferentes domínios. O desbalanceamento de classes ocorre quando uma das classes (classe majoritária) contém muito mais exemplos do que outra (classe minoritária) na base de dados (GU et al., 2008). De acordo com BECKMANN (2010), no processo de classificação, para que a indução aprenda conhecimento novo representativo, os exemplos das classes precisam estar bem definidos e deve haver uma quantidade suficiente deles. A utilização de bases de dados com a distribuição desbalanceada entre as classes compromete significativamente o desempenho dos algoritmos de classificação.

Quando apresentado a um conjunto de dados desbalanceados, o algoritmo de classificação tende a classificar todos os dados como sendo da classe majoritária, que normalmente é a classe de menor interesse. Isso ocorre pelo fato dos algoritmos de aprendizagem de máquina

focarem na redução da taxa de erro geral, desconsiderando as diferenças entre os tipos de erro de classificação por julgá-las igualmente importantes (RÊGO, 2016). Assim, conforme SPILIOPOULOS; VOUROS; KARKALETSIS (2010), para que a classificação alcance maior eficiência, é importante que o conjunto de treinamento esteja balanceado em número, ou seja, as instâncias de treinamento de todas as classes devem ser numericamente iguais.

Dentre as técnicas existentes na literatura, destaca-se a *oversampling*. Essa consiste em replicar aleatoriamente exemplos pertencentes à classe minoritária para obter uma distribuição mais balanceada (PRATI et al., 2003). Neste trabalho, aplicou-se o método *oversampling* SMOTE (*Synthetic Minority Over-sampling TEchnique* (SMOTE)), que incrementa a classe minoritária criando novos exemplos sintéticos desta classe através da interpolação entre diversos exemplos da classe minoritária que se encontram próximos uns dos outros (CHAWLA et al., 2002). No método, tem-se que $|S| = |S_{min}| + |S_{maj}|$, onde S representa um determinado conjunto de treinamento com m instâncias, ou seja, $|S| = m$ e S_{min} e S_{maj}, representam os conjuntos das classe minoritária (classes positivas) e majoritária (classe negativas), respectivamente. No subconjunto S_{min}, para cada instância $x_i \in S_{min}$, considerem-se os k-vizinhos mais próximos, para um dado valor inteiro especificado k. Os k-vizinhos mais próximos são definidos como os k elementos de S_{min} cuja distância euclidiana entre si e a instância x_i apresentam o menor valor. Para gerar uma nova instância sintética, seleciona-se randomicamente um dos k-vizinhos mais próximos, subtrai-se a instância x_i de seu vizinho mais próximo, multiplica-se esta diferença por um número aleatório entre 0 e 1 e, adiciona-o ao valor da instância (x_i). Ou seja, $x_{new} = x_i + (y_i - x_i) * \delta$, onde x_i é uma instância da classe minoritária em consideração, y_i é um dos seus k-vizinhos mais próximos de x_i e δ é o número aleatório (HE; GARCIA, 2009).

2.2.3.6 Algoritmos de Classificação

Na literatura, há vários algoritmos para realizar a classificação automática de textos, tais como: *Naive Bayes*, Redes Neurais e *Support Vector Machine* (SVM). Para o cumprimento do objetivo proposto, este trabalho utilizou o algoritmo *Random Forest*.

O *Random Forest* (Floresta Aleatória) é um algoritmo desenvolvido por BREIMAN (2001) para classificação de dados. Esse algoritmo consiste em uma coleção de classificadores estruturados em árvores $\{h(X, \Theta_k), k = 1, ...\}$, onde $\{\Theta_k\}$ são vetores aleatórios identicamente distribuídos, e cada árvore realiza um voto unitário pela classe mais popular na entrada X. A proposta do algoritmo é gerar várias árvores de decisão construídas simultaneamente e considerando todas as variáveis selecionadas para a análise. Conjuntamente, essas árvores serão utilizadas para classificação de novos objetos através da agregação dos resultados. Para que isso ocorra, cada árvore irá realizar um voto indicando a sua escolha sobre a classe à que o objeto pertence. A decisão quanto à classificação é definida com o maior número de votos. A Figura 2.3 ilustra o funcionamento simplificado do algoritmo.

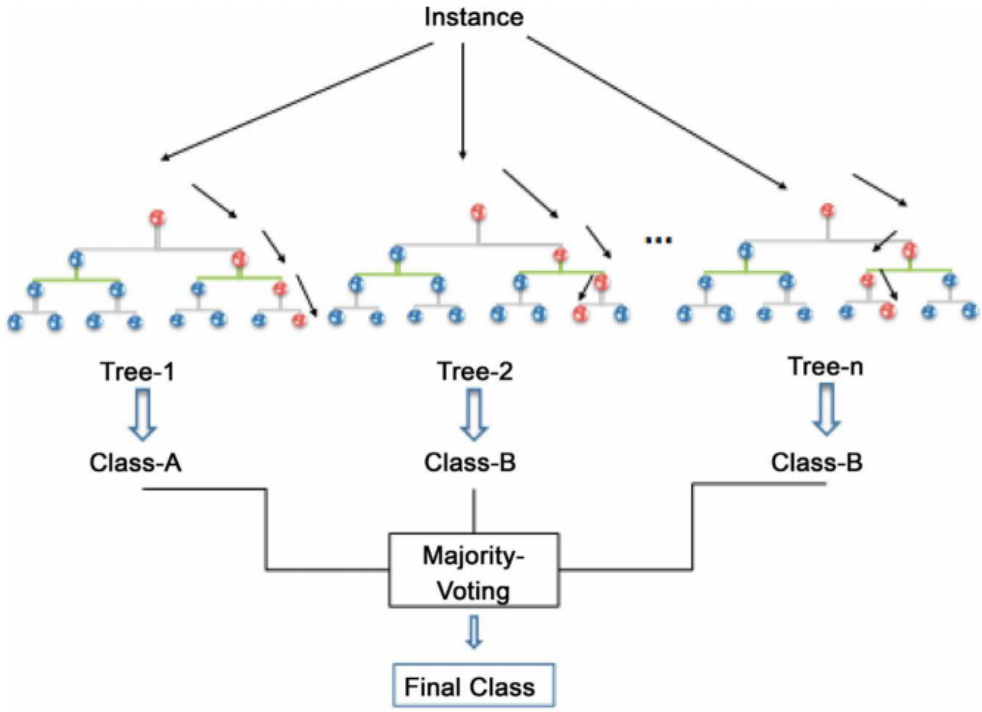

Figura 2.3: Funcionamento do algoritmo *Random Forest* (FU, 2017)

O algoritmo aplica o método *bagging* (*Bootstrap Aggregating*) para produzir amostras aleatórias de conjuntos de treinamento (amostras *bootstraps*) para cada árvore gerada. A amostragem *bootstrap* é uma técnica que faz amostragem com reposição: a partir do conjunto de treinamento inicial, são escolhidos aleatoriamente exemplos para um novo subconjunto de treinamento (EFRON, 1979). Assim, no método *bagging*, diferentes subconjuntos são aleatoriamente construídos, com reposição, a partir do conjunto original.

O uso de *bagging* traz benefícios como melhora no desempenho do algoritmo e a possibilidade de obter estimativas internas do erro de generalização do conjunto combinado de árvores, da força de uma árvore classificadora e correlação entre as árvores classificadoras, utilizando o método *out-of-bag*. No método, para um dado conjunto de treinamento C, a cada objeto $c_i = (x_i, y_i), \in C$, uma árvore classificadora T da floresta aleatória, usando um objeto C_i, é gerada realizando as médias dos votos somente das árvores classificadoras que não correspondem ao *bootstrap* de amostras que contenham C_i. Como vantagem, o *out-of-bag* fornece estimativas de erro que dispensam o uso de um conjunto de teste, que equivale à validação cruzada, entretanto realiza cálculos das estimativas internas durante o processo de treinamento (BREIMAN, 2001).

Outro ponto sobre o método *out-of-bag*, é que através dele estima-se a relevância de uma variável, a correlação entre classificadores, e mede-se a força de predição de cada variável. Após cada árvore ser gerada, as amostras *out-of-bag* são passadas para baixo da árvore, e a precisão de previsão é registrada. Em seguida, os valores para a j-ésima variável são aleatoriamente permutados nas amostras *out-of-bag* e a precisão é novamente calculada. Como resultado desta permuta tem-se a redução da precisão, que é calculada em média em todas as árvores e é utilizada como uma medida da relevância da variável j na *random forest* (FRIEDMAN;

HASTIE; TIBSHIRANI, 2001).

Segundo BREIMAN (2001), as florestas aleatórias podem ser aplicadas tanto para a tarefa de classificação quanto para a tarefa de regressão. Além disso, não são sensíveis a ruídos como árvores individuais, pois foram geradas por árvores construídas a partir de amostras com reposição, para a diminuição da correlação entre as árvores.

Neste contexto, o autor propôs duas medidas de importância de atributos, uma baseada na importância da permutação, também chamada decaimento da exatidão média (do inglês *Mean Decrease in Accuracy* (MDA)) e a outra na impureza de Gini, conhecida como diminuição média do índice de Gini (do inglês *Mean Decrease in Gini* (MDG)). Neste estudo aplicou-se o MDG. O MDG é proveniente do treinamento do classificador *Random Forest*. Em cada nó da árvore de decisão, a divisão ótima é buscada segundo a impureza de Gini, ou seja, uma medida de quão bem uma divisão potencial é capaz de separar as amostras de um nó particular. A medida é então calculada pela soma de todas as diminuições na impureza de Gini a cada divisão do nó, normalizada pelo número de árvores (BREIMAN, 2001).

A robustez do *Random Forest* quanto aos ruídos, *outliers* e ao problema de sobreajuste está relacionada com a utilização de metodologias que diminuem a variância do modelo utilizado (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). COUTEIRO (2010) afirma que o algoritmo tem melhor precisão que a maioria dos algoritmos atuais, mostrando-se eficaz para bancos de dados grandes. Além disso, as *Random Forest* geradas podem ser armazenadas para serem utilizadas futuro com outros dados. São construídos protótipos que fornecem informações a respeito da relação entre as variáveis e a classificação, e ainda proximidades entre pares de casos que podem ser aplicados nos casos de *clustering*.

2.2.4 Avaliação e interpretação dos resultados

Esta etapa, também chamada de Pós-processamento de dados, refere-se à verificação da eficiência da aplicação do classificador da etapa anterior. De acordo com SEBASTIANI (2002), a avaliação experimental de um classificador, de modo geral, mede sua eficácia, isto é, sua capacidade de tomar as decisões corretas de classificação. Segundo CAMILO; SILVA (2009), é uma etapa em que testes e validações, com o objetivo de obter a confiabilidade nos modelos, precisam ser executados (*cross validation*, *suplied test set*, *use training set*, *percentage split*). Além disso, indicadores que permitam a análise dos resultados devem ser obtidos (matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística kappa, acurácia, precisão, *F-measure*, dentre outros).

Nesta pesquisa, foram aplicadas a técnica de validação cruzada e as medidas: acurácia e kappa.

2.2.4.1 Validação Cruzadas

Validação cruzada (do inglês *Cross Validation*) refere-se a uma técnica estatística que particiona uma amostra de dados em subconjuntos de maneira que a análise é inicialmente executada em um único subconjunto, enquanto que os demais subconjuntos são mantidos para treino (KOHAVI, 1995). Segundo REZENDE (2003), o método funciona da seguinte forma: um conjunto de dados é aleatoriamente dividido em K subconjuntos mutuamente exclusivos (*folds*) cujo o tamanho é aproximadamente igual a n/K, onde n é o tamanho do conjunto de dados. Então, são realizados K experimentos, onde, em cada um, um subconjunto diferente é escolhido para o teste e os $K-1$ subconjuntos restantes são designados para o treinamento. A medida de eficiência é a média das medidas de eficiência calculadas para cada um dos subconjuntos.

O benefício deste método é que todos os dados do conjunto são utilizados tanto para treinamento quanto para teste. Dessa forma, a maneira como a divisão do conjunto foi feita influi menos no resultado final, qualquer dado será usado exatamente uma vez para teste e $k-1$ vezes para treinamento (PENG et al., 2004).

2.2.4.2 Acurácia

Uma da medidas mais comuns para avaliar um modelo de classificação (WITTEN; FRANK; MARK, 2011), reflete a taxa de acerto, ou seja, o número de classificações que o modelo inferiu corretamente. É uma medida relacionada com os percentuais de acertos e erros do modelo de predição da classe na classificação de novos exemplos e a mesma pode ser calculada a partir de uma estrutura denominada matriz de confusão.

A matriz de confusão (Tabela 2.4) apresenta o número de predições corretas e incorretas em cada classe, sendo que as linhas dessa matriz representam as classes verdadeiras e as colunas as classes preditas por um classificador.

	Predição Positiva	**Predição Negativa**
Classe Positiva	Verdadeiro Positivo (TP)	Falso Negativo (FN)
Classe Negativa	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Tabela 2.4: Matriz Confusão.

Onde, **Verdadeiro Positivo (TP)** representa o número de instâncias da classe positiva preditos corretamente, **Verdadeiro Negativo (TN)** representa o número de instâncias da classe negativa preditos corretamente, **Falso Positivo (FP)** representa o número de instâncias da classe negativa preditos como sendo da classe positiva e o **Falso Negativo (FN)** representa o número de instâncias da classe positiva preditos como sendo da classe negativa.

Assim, a acurácia, no contexto da mineração de textos, é a proporção dos documentos classificados corretamente (positivos e negativos) sobre o número total de documentos (HOTHO; NÜRNBERGER; PAASS, 2005) e é definida pela Fórmula 2.1 abaixo:

$$\textbf{Acurácia} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

2.2.4.3 Estatística Kappa

O coeficiente estatístico índice Kappa ou Estatística κ (COHEN, 1960), é uma medida de concordância em escalas nominais. Aplicado ao contexto de classificação em mineração de texto, o índice aponta o nível de concordância entre a classificação do modelo e a classificação de referência, isto é, o quão os dois estão de acordo quanto à classificação (CARLETTA, 1996).

O valor máximo 1 denota total concordância e os valores próximos, e até abaixo de 0, indicam nenhuma concordância, ou a concordância foi exatamente a prevista pelo acaso. LANDIS; KOCH (1977) associam valores de Kappa à qualidade da classificação de acordo com a Tabela 2.5 abaixo:

Valores de Kappa	Interpretação
<0	Sem acordo
0 – 0,19	Baixa concordância
0,20 – 0,39	Acordo justo
0,40 – 0,59	Concordância moderada
0,60 – 0,79	Acordo Substancial
0,80 – 1,00	Concordância quase perfeita

Tabela 2.5: Interpretação dos índices de Kappa (adaptado de LANDIS; KOCH (1977))

O índice Kappa (κ) é calculado de acordo com a Equação 2.2 abaixo

$$\kappa = \frac{P(A)-P(E)}{1-P(E)} \tag{2.2}$$

Em que P(A) é a proporção em que os avaliadores concordaram e P(E) a proporção esperada em que os avaliadores concordaram

2.3 Ferramentas

Para cumprir a tarefa de descobrir conhecimento útil em fontes textuais, a literatura disponibiliza diversas bibliotecas e ferramentas. A seguir, são apresentadas aquelas que foram utilizadas neste trabalho.

2.3.1 *Linguistic Inquiry and Word Count* (LIWC)

O LIWC é uma ferramenta desenvolvida por PENNEBAKER; FRANCIS; BOOTH (2001) com o objetivo de fornecer um método eficiente para estudos sobre fatores emocionais,

cognitivos, estruturais, entre outros, presentes em trechos de falas verbais e escritas de indivíduos. Essa possui como núcleo um dicionário de palavras que fornece informações sobre os fatores acima citados.

Originalmente, o LIWC foi projetado para descobrir experiências negativas relativas à vida de pacientes, para que possíveis melhorias na saúde pudessem ser previstas. Mais recentemente, sua aplicação foi expandida para o rastreio e uso da linguagem natural em fontes de texto abrangendo literatura clássica, narrativas pessoais, conferências de imprensa, e transcrições de conversas diárias.

A versão em inglês do dicionário LIWC é constituída por uma base de aproximadamente 6.540 palavras e cada uma delas está associada a uma ou mais categorias, dentre as 73 presentes no dicionário. As categorias incluem: i) **estatísticas comuns do texto**: número de palavras, palavras por sentença, palavras pertencentes ao dicionário, palavras únicas, palavras com mais de 6 caracteres; ii) **dimensão linguística**: contagens de pronomes, pronomes pessoais, negações, artigos, preposições, e números; iii) **processos psicológicos**: relacionados com reações a emoções (ansiedade, raiva, tristeza, por exemplo), mecanismos cognitivos, relacionados a sensações/percepções, visão, audição, toque, sociais, comunicativos, relacionados a amigos, família, humanos, dentre outros; iv) **relatividade**: contém referências a tempo, espaço, movimento, verbos no passado, presente e futuro; v) **assuntos pessoais**: traz referências a ocupação de trabalho, lazer, música, dinheiro, sexo, morte, religião, dentre outros; vi) **miscelânea**: captura palavras de ofensa e xingamento e particularidades da fala (quantidade de disfluências e preenchedores).

No dicionário, as palavras possuem rótulos que foram nomeados com códigos numéricos que representam as categorias as quais pertencem. Por exemplo, a palavra "dúvida" no dicionário pertence a 5 categorias: social, humano, cognitivo, intuição e tentativa. Estas são representadas no dicionário pelos códigos: 121, 124, 131, 132 e 135, respectivamente.

Em uma iniciativa do grupo de pesquisa do Centro Interinstitucional de Linguística da Computação (NILC) e do Instituto de Matemática e Ciências da Computação da Universidade de São Paulo, o dicionário do LIWC em português foi construído usando como base o léxico original *English LIWC Dictionary* e disponibilizado no site PortLex[1]. Essa versão possui 127.149 palavras, em que cada uma delas está assinalada a uma ou mais das 64 categorias disponíveis (BALAGE FILHO; PARDO; ALUÍSIO, 2013). Na seção 4.3.1, a Tabela 4.8 apresenta as 64 categorias, assim como uma breve descrição sobre cada uma delas.

Portanto, o LIWC é uma ferramenta de análise textual em que o documento é estruturado em categorias, atribuindo a cada palavra desse uma ou mais categorias correspondentes.

2.3.2 Coh-Metrix

O sistema Coh-Metrix, que significa *cohesion metrics*, é uma ferramenta para análise textual (em inglês). Desenvolvida por pesquisadores da Universidade de Memphis, nos Estados

[1] http://www.nilc.icmc.usp.br/portlex/index.php/en/liwc

Unidos (GRAESSER et al., 2004), a ferramenta calcula índices a coesão e a coerência textual considerando medidas léxicas, sintáticas, semânticas e referenciais com o intuito de indicar a adequação de um texto ao seu público-alvo (a "demanda cognitiva" e a legibilidade do texto). Além disso, possui a função de indicar dados para identificar problemas textuais de ordem estrutural (FINATTO, 2011; GRAESSER et al., 2004).

Para o desenvolvimento do sistema, os autores propõem uma distinção entre coesão e coerência em um texto. A coesão é uma característica do texto, cujas construções coesivas consistem nas palavras, expressões ou sentenças que norteiam o leitor no processo de estabelecimento mental de uma representação consistente do conteúdo do texto. A coerência é definida como uma característica da representação mental do conteúdo do texto que o leitor constrói no decorrer da leitura e é fortemente influenciada por sua bagagem cognitiva, ou seja, pelo conhecimento de mundo que possui, por suas habilidades de interpretação e raciocínio e pelos construtos coesivos do texto explícito (GRAESSER et al., 2004).

A ferramenta reúne a saída de diversas outras ferramentas de PLN e pode ser utilizada em vários cenários de análise e classificação de textos. Foram coletadas e avaliadas diversas métricas, que medem características do texto relacionadas a palavras, sentenças e à conexão entre sentenças (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

O Coh-Metrix 3.0 é a versão livre mais recente da ferramenta, com 108 índices que vão desde medidas simples, como contagem de parágrafos, sentenças e palavras até medidas mais complexas envolvendo algoritmos de resolução anafórica.

O Coh-Metrix foi adaptado para o português do Brasil pelo NILC (Núcleo Interinstitucional de Linguística Computacional) da Universidade de São Paulo (USP), como fruto do projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) que visava a construção de sistemas para promover o acesso a textos escritos em Português Brasileiro por analfabetos funcionais, pessoas com problemas cognitivos, e crianças e adultos em fase de aprendizado de leitura e escrita (CUNHA, 2015).

O Coh-Metrix-PT 2.0 [2], é a versão atual da ferramenta para o português, com 48 medidas implementadas de nível léxico, sintático em nível de sintagmas nominais, semântico, e discursivo (SCARTON; GASPERIN; ALUISIO, 2010). Na seção 4.3.2, a Tabela 4.9 exibe as categorias disponíveis, assim como uma síntese sobre cada uma delas.

Para cada texto submetido ao sistema, uma tabela é gerada com os resultados dos cálculos gerados pelas métricas existentes. Essa tabela apresenta aspectos como identificação do texto, contagens básicas, operadores lógicos, frequências, hiperônimos, pronomes, *tokens*, constituintes, conectivos, ambiguidades, correferência e anáforas. As Figuras 2.4 e 2.5 a seguir apresentam a submissão de um texto exemplo e o resultado fornecido pela ferramenta.

[2] http://143.107.183.175:22680/

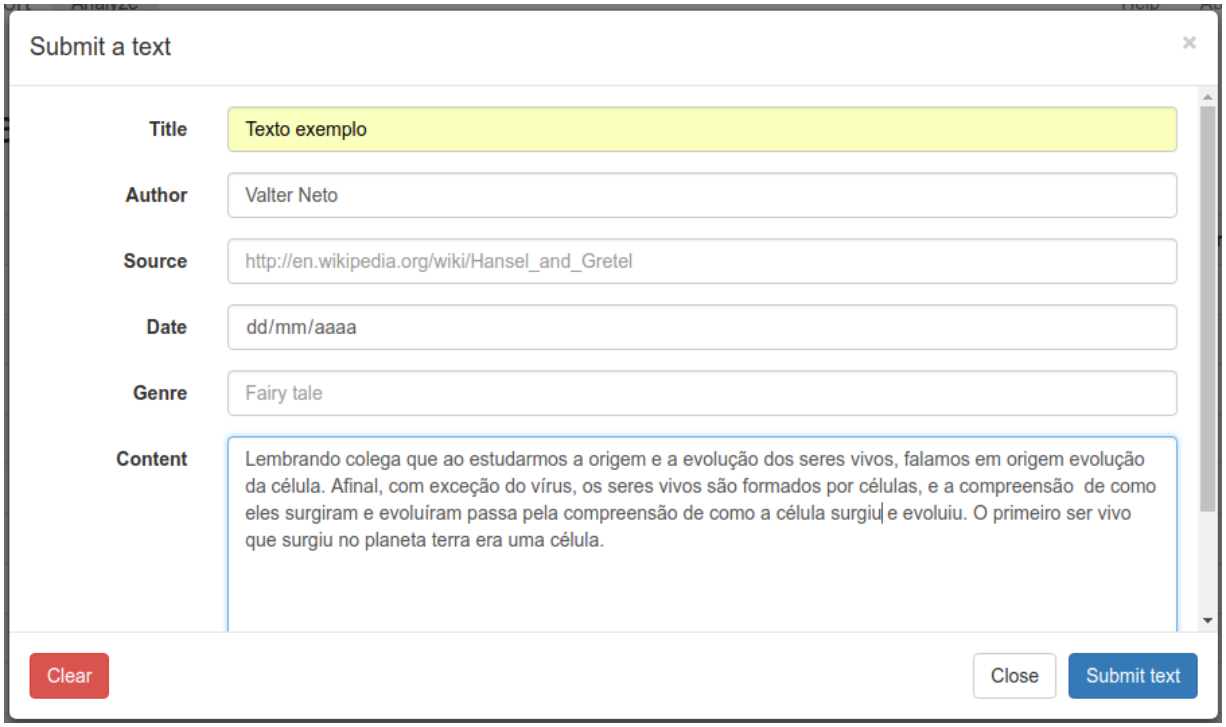

Figura 2.4: Submissão de um texto exemplo no Coh-Metrix-PT

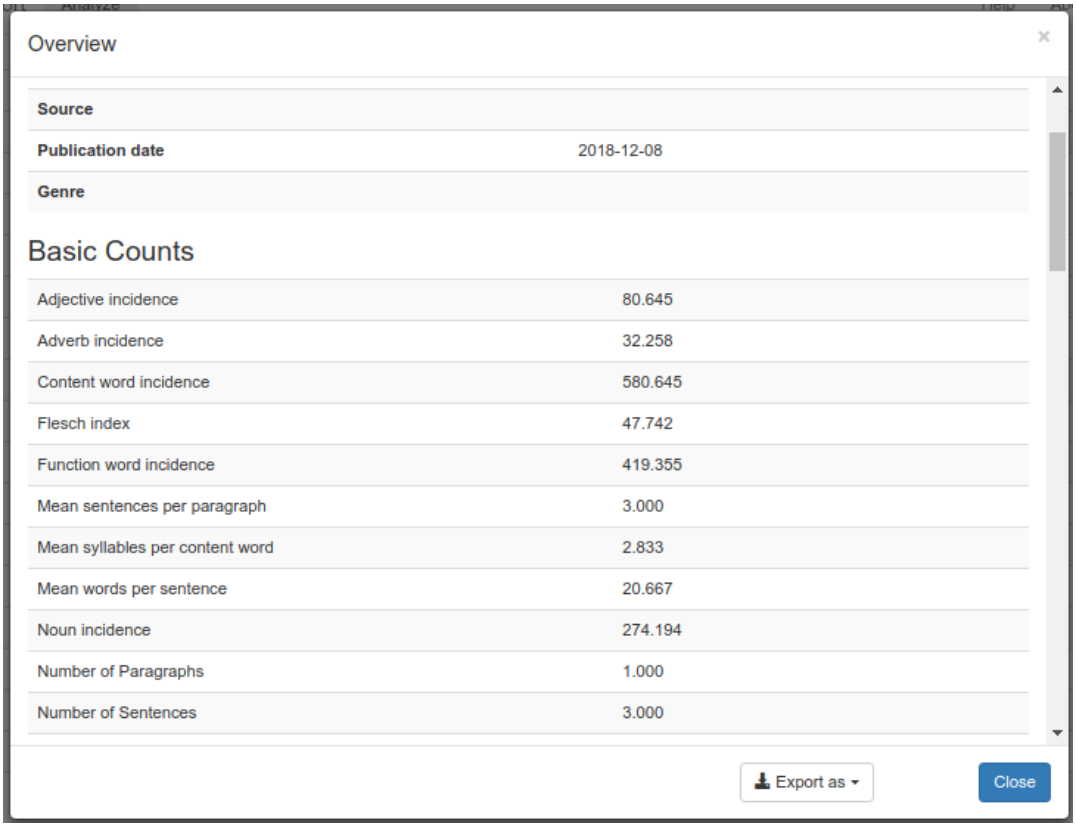

Figura 2.5: Saída do Coh-Metrix-PT para um texto exemplo

Nesta dissertação utilizou-se as medidas fornecidas pela ferramenta Coh-Metrix com o objetivo de extrair métricas com relevância no processo sociocognitivo.

2.3.3 Bibliotecas *Python*

Na computação científica, a linguagem de programação *Python* está se consolidando como uma das linguagens mais populares. Devido à sua natureza interativa, com estruturas de dados de alto-nível eficiente e ao seu amadurecido na desenvolvimento de bibliotecas científicas, é atualmente, uma escolha atraente para o desenvolvimento lógico e análise exploratória de dados (MILLMAN; AIVAZIS, 2011).

Para MERTZ (2003), *Python* por ser uma linguagem clara, expressiva e de propósito geral, mostrou-se muito versátil para o processamento de textos. CHUN (2006) reforça suas qualidades ao destacar seu caráter extensível e "plugável".

Assim, utilizaram-se neste trabalho alguns recursos disponíveis da linguagem para o desenvolvimento dos experimentos, conforme detalhados a seguir.

2.3.3.1 *spaCy library*

A principal função de um analisador sintático é determinar a estrutura (sintática) do texto de entrada na ferramenta. Nesta dissertação, utilizou-se o analisador sintático *spaCy* para complementar o processo de extração de características.

O *spaCy* é uma biblioteca desenvolvida em *Python* para realizar tarefas de Processamento da Linguagem Natural (PLN), mais focado para uso industrial (GAMALLO; GARCIA, 2017). Atualmente, está disponível para línguas como inglês, alemão, espanhol e português. Nesse último, possui suporte para as tarefas: *tokenization*, *part-of-speech tagging*, *dependency parsing* e *named entities*.

A escolha desta ferramenta, baseou-se nas funções acima citadas e pelo bom desempenho apresentado na pesquisa realizada por CHOI; TETREAULT; STENT (2015). Comparada a outros analisadores (*parsers*), os testes realizados mostraram que o analisador obteve uma acurácia (*overall accuracy*) bastante satisfatória. Em média, a ferramenta obteve uma acurácia de 90,53% de exatidão. Além disso, nos testes de velocidade (*overall speed*), a ferramenta obteve o melhor desempenho comparativo.

2.3.3.2 *scikit-learn*

O *Scikit-Learn* é uma biblioteca em *Python* com diversas implementações de algoritmos de aprendizagem de máquina para problemas supervisionados e não supervisionados de média escala e funções para extração de características, processamento de dados, e avaliação de modelos. A biblioteca aproveita o ambiente sofisticado oferecido pela linguagem *Python* para fornecer implementações de muitos algoritmos de aprendizado de máquina bem conhecidos como *Naïve Bayes*, SVM e *Random Forest* (PEDREGOSA et al., 2011).

Os autores destacam que a biblioteca difere-se das demais ferramentas de aprendizagem de máquina em *Python* por vários motivos: i) licença BSD (*open source*); ii) incorpora código

compilado para eficiência; iii) dependência mínima (apenas das bibliotecas *numpy* e *scipy*); e iv) concentra-se na programação imperativa.

Segundo HACKELING (2014), o *scikit-learn* é popular para pesquisa acadêmica porque possui uma API bem documentada, fácil de usar e versátil. Os desenvolvedores podem usar a biblioteca para experimentar algoritmos diferentes, alterando apenas algumas linhas do código.

Neste trabalho a biblioteca foi utilizada para obter a amostragem estratificada dos dados de treinamento e teste.

3

Trabalhos Relacionados

Em uma comunidade de investigação, a presença cognitiva é considerada um elemento essencial e básico para o sucesso da aprendizagem, e está fundamentada no pensamento crítico. Ela está associada à capacidade dos estudantes em construírem conhecimentos e saberes por meio da reflexão, debate e da comunicação entre os participantes (GARRISON; ANDERSON; ARCHER, 2001).

Assim, neste capítulo serão apresentados trabalhos que utilizaram a presença cognitiva e suas respectivas fases para a compreensão da construção do conhecimento. Esses serão discutidos em duas seções: trabalhos em que há classificação automática de postagens nas fases da presença cognitiva e aplicações da presença cognitiva na língua portuguesa.

3.1 Classificação Automática de Postagens nas Fases da Presença Cognitiva

Os estudos iniciais que analisaram a automatização da análise de conteúdo no contexto da presença cognitiva, basearam suas abordagens principalmente em contagem de frases e palavras, como MCKLIN (2004) e CORICH; HUNT; HUNT (2006).

Em seu trabalho, MCKLIN (2004) propôs um sistema usando redes neurais artificiais (*feedforward* e *backpropagation*) para classificar automaticamente 1.997 mensagens de fóruns de discussão. As *features* para a classificação foram extraídas das 182 diferentes categorias de palavras definidas através do dicionário léxico *General Inquirer* (STONE; DUNPHY; SMITH, 1966). Além disso, o autor utilizou um indicador binário para caracterizar se uma mensagem é uma resposta a outra mensagem (Eventos desencadeadores são mais propensos a serem indicadores de discussões), e definiu categorias personalizadas de palavras e frases, que foram consideradas indicativas das quatro fases da presença cognitiva (palavras indicativas, por exemplo). Assim, apesar dos problemas de generalização de algumas fases (Exploração e Integração) e da exclusão da fase Resolução pela rede neural artificial na classificação (nenhuma mensagem foi classificada na fase Resolução), o sistema alcançou 0,69 de Coeficiente de confiabilidade de Holsti e 0,31 de *Cohen's Kappa*.

A referência CORICH; HUNT; HUNT (2006) apresenta o sistema ACAT (*Automated Content Analysis Tool*), desenvolvido para automatizar a classificação de mensagens de fóruns de discussão para vários modelos. A ferramenta possui módulos que auxiliam no processo de codificação manual e permitem que os resultados da classificação manual sejam comparados com os resultados da classificação automática, que são realizados por meio de dicionários construídos com exemplos dos modelos. No experimento realizado com a presença cognitiva e suas fases, os autores submeteram ao sistema uma base de dados com 74 postagens, que foram divididas em 484 sentenças e salvas em uma tabela para serem classificadas manualmente por dois especialistas. Em seguida, o ACAT gerou um relatório com a classificação manual e outro com três classificações automáticas: i) sentenças sem pré-processamento; ii) sentenças submetidas a técnica de remoção de *stopwords*; iii) sentenças submetidas a técnica de *stemming*. Comparando as classificações automáticas com a manual, o sistema ACAT apresentou os seguintes resultados: i) 0,64; ii) 0,65; iii) 0,71 de coeficiente de confiabilidade de Holsti, respectivamente.

Estudos mais recentes relatam experiências com o uso de outras características (*features*) e classificadores diferentes. Dentre eles, destacam-se: KOVANOVIĆ et al. (2014),CHEN (2015), WATERS et al. (2015) e KOVANOVIĆ et al. (2016).

Na pesquisa de KOVANOVIĆ et al. (2014), os autores aplicaram técnicas de mineração de texto e classificação de texto para automatizar a classificação de postagens. O método proposto utilizou sete *features*: *N-grams* (*unigram*, *bigram*, *trigram*); *Part-of-Speech N-grams* (*bigrams*, *trigrams*); *Back-Off N-grams* (*bigrams*, *trigrams*); *Dependency Triplets*; *Back-Off Dependency Triplets*; *Named Entities*; *Thread Position Features*. Além disso, implementou o *Support Vector Machines* (SVM) como classificador e fez *folds cross-validation* para validação. Para a implementação do classificador e extração de features foram utilizadas as ferramentas de código aberto (*open source*) e as bibliotecas: *Stanford CoreNLP* (tokenização, *Part-of-Speech tagging*; *dependency parsin*); *Weka* e *LibSVM* (classificador); *Java Statistical Classes* (JSC) (teste McNemar para comparação de *features*). Desta forma, com um conjunto de dados composto por 1.747 mensagens, categorizadas manualmente por dois especialistas (*Cohen's Kappa* = 0,974) em 5 categorias (*Other*, *Triggering Event*, *Exploration*, *Integration*, *Resolution*), o método proposto alcançou 58,4% de acurácia e 0,41 de *Cohen's Kappa*.

Em seu estudo, CHEN (2015) apresentou duas abordagens de classificadores: o *Four-category classifier* (com um modelo para as quatro fases) e o *Binary classifier* (com quatro modelos, um para cada fase). Nelas, foram utilizados os algoritmos *Multinomial Naïve Bayes* (MNB) e SVM para os testes de classificação e 3 *folds cross-validation* na validação. Para os experimentos o autor codificou manualmente 738 postagens de um fórum de discussões nas quatro fases da presença cognitiva e as utilizou para construir sua base de treinamento. Além disso, utilizou *N-grams features* (*unigram*, *bigram*, *trigram*) e métodos de vetorização (*count*, *boolean*, *term frequency*) para comparar os resultados. O desenvolvimento e avaliação dos modelos foram realizados em *Python*. Assim, a primeira abordagem apresentou melhor resultado (acurácia de 0,575) usando o modelo MNB (*Trigram-Boolean*), e a segunda, o modelo

binário SVM obteve o melhor desempenho na classificação de diferentes categorias (acurácias de 0,910, 0,988 e 0,712 nas fases Evento desencadeador, Integração e Resolução, respectivamente) utilizando todas postagens no *dataset* de treinamento.

WATERS et al. (2015) apresentam um modelo para classificação automática de postagens nas fases da presença cognitiva utilizando *Conditional Random Fields* (CRFs). Para esse, os autores implementam o *Linear-Chain Conditional Random Field* (LCCRF), utilizando Java e a biblioteca *Mallet*. Além disso, definiram como características (*features*): *word unigrams*, *Entity Count*, *First Post and Last Post*, *Comment Depth*, *Post Similarity*, *Word and Sentence counts*, *Number of Replies*. Para os testes, foi utilizado o *dataset* criado por KOVANOVIĆ et al. (2014) (70, 20, 10% para treinamento, teste e validação, respectivamente). Assim, aplicando o método *remerging* usando o voto majoritário nas postagens encadeadas, o classificador alcançou 64,2% de acurácia e 0,482 de *Cohen's Kappa*.

KOVANOVIĆ et al. (2016) propõem usar 205 características extraídas em *Coh-Metrix*, *LIWC*, *Latent Semantic Analysis* (LSA), Entidades Nomeadas e contexto de discussão. Como classificador, implementam o *Random Forest* e *10-fold cross validation*, para validação. Para os experimentos foi utilizado o conjunto de dados de KOVANOVIĆ et al. (2014), balanceado através do algoritmo de *SMOTE*. Assim, com 75% dos dados para treinamento e 25% para teste, o classificador atingiu a acurácia de 70,3% e 0,63 de *Cohen's Kappa* e permitiu também a análise da influência das diferentes características nos resultados da classificação final, como presença de pontos de interrogação e pronomes de primeira pessoa do singular eram indicativos dos níveis mais baixos de presença cognitiva.

Os trabalhos acima citados avaliam mensagens de fóruns escritas na língua inglesa. Na literatura, constatou-se que não existe nenhuma publicação que desenvolveu alguma metodologia para analisar automaticamente a presença cognitiva em textos escritos em português, nem em qualquer outra língua.

Como o foco deste estudo é examinar o uso da análise de textos para avaliar as mensagens de discussão *online* em português quanto a presença cognitiva, estudos que abordaram a presença nos cursos *online* em português também foram examinados. As pesquisas selecionadas serão abordadas na seção a seguir.

3.2 Aplicações para Língua Portuguesa

Existem vários estudos que aplicam os conceitos da presença cognitiva na língua portuguesa. Para esta pesquisa serão apresentados somente aqueles que nortearam sua metodologia baseada nas fases que compõem a presença ou que desenvolveram propostas diferenciadas para identificação da presença cognitiva.

GOMES; PESSOA (2012b) analisaram a experiência de aprendizagem na área da supervisão pedagógica, utilizando as presenças social e cognitiva, a partir da criação de um grupo de estudo no *Facebook*, de nove sujeitos através das publicações nas seções "Mural" e "Discussões"

do grupo criado. A metodologia de análise para categorização das mensagens combinou as fases e indicadores das presenças. Em particular, foi acrescentada a fase Comum(idade) à presença cognitiva para classificação das mensagens geradas em "Discussões". Assim, foram analisadas 27 mensagens do "Mural" e 70 mensagens das "Discussões", separadamente. Através das análises, foram definidos os indicadores de cada uma das categorias (social e cognitiva) e de cada seção (Mural e Discussões), e em seguida, foi realizada a classificação das mensagens segundo esses indicadores.

LAGARTO; MARQUES (2017) utilizaram as presenças cognitiva e pedagógica para mostrar como os participantes de uma conferência *online* organizam e manifestam as suas aprendizagens num chat. Inicialmente, o autor analisou todas as mensagens para extrair apenas aquelas que demonstraram alguma relação de sentido e/ou contexto com as presenças em foco no estudo. Em seguida, para as mensagens selecionadas (467), foi definido um modelo de análise de investigação em que a presença cognitiva possuía duas categorias (conhecimento implícito e conhecimento explícito) e indicadores temáticos em cada uma delas, como concordância simples com conteúdo, citação associada a conhecimento e questão sobre conteúdo.

Com o intuito de demonstrar o uso de recursos da Tecnologia da informação e comunicação (TIC) para desenvolver competências e autonomia de professores de língua portuguesa, ao mesmo tempo que se habituam a trabalhar colaborativamente, SÁ; MACÁRIO (2014) propõem um processo de ensino/aprendizagem centrado em projetos e com foco no aluno, através de um fórum de discussão estruturado nas fases da presença cognitiva. O fórum de discussão possuía três partes (três temas de discussão) e em cada tema ou subtema compreendia quatro tópicos, referentes as quatro fases da presença cognitiva. Com esta estrutura, foi possível coletar *posts* e documentos produzidos e anexados à plataforma pelos estudantes que possibilitaram a identificação do trabalho colaborativo desenvolvido e avaliação da presença cognitiva.

MACÁRIO; SÁ; MOREIRA (2014) apresentam uma experiência com foco na construção do conhecimento de forma colaborativa utilizando como recurso um fórum de discussão *online* estruturado segundo as fases da presença cognitiva. O fórum foi dividido em três partes (temas de discussão) em que cada tema ou subtema possuía quatro tópicos, que correspondia as quatro fases da presença cognitiva. Dessa forma, sete alunos, divididos em dois grupos, participaram do fórum de discussões desenhado para que trabalhassem colaborativamente, norteado pelas fases da presença cognitiva, de forma a construírem conhecimento sobre o tema Ortografia.

ROZENFELD (2014) propôs um modelo de três fases (evento disparador, elaboração e resolução) com base no conceito de pensamento crítico e na presença cognitiva, para verificar as contribuições de fóruns *online* para a manifestação do pensamento crítico de futuros professores. A autora utilizou ainda, a Linguística Sistêmico-Funcional (HALLIDAY, 1994) e os tipos de movimentos conversacionais (moves) propostos por (EGGINS; SLADE, 1997), para identificação de elementos linguísticos da língua portuguesa que forneçam mais dados para a classificação das mensagens. Assim, o modelo era iniciado sempre do evento disparador, seguindo para a fase de resolução ou de elaboração ou para a intersecção das duas. Além disso, existiam o mundo interno

(reflexivo, particular) e o externo (compartilhado, social), equivalentes à fase de Exploração do pensamento cognitivo, presentes em todos os momentos, uma vez que as ideias são exploradas colaborativamente e os estudantes vão construindo sentidos individualmente e na interação com o grupo para as proposições.

Em seu trabalho, ARAUJO (2014) propôs o Fórum Socrático Cognitivo (FSC), um modelo de fórum que reuniu as fases de desenvolvimento da presença cognitiva, as categorias de perguntas socráticas, propostas por (PAUL, 1993) e palavras-chaves de maior relevância para o processo de explicitação de novos conceitos e conhecimentos, segundo a Teoria do Conceito de (DAHLBERG, 1978). Para a pesquisa foram construídos dois conjuntos de dados (*datasets*): o *dataset message* (3.400 mensagens postadas por 544 alunos) e o *dataset reference* (palavras-chaves e conceitos importantes de 3 textos científicos) e, definiu-se uma variável resposta (notas de alunos) e 5 covariáveis (conceitos, presença cognitiva, tópicos, riqueza vocabular e sala) para a avaliação dos alunos (instrumento de avaliação das mensagens ou grid de avaliação). Assim, foram realizadas: (i) uma Análise textual, a exploração e navegação em textos para obter uma ordenação (*rank*) de palavras com maior frequência, com o software *SPHINX Léxica v5*®; (ii) uma Análise estatística (regressão linear e testes ANOVA) para avaliar as diferenças entre as variáveis; e (iii) uma Análise e construção de redes complexas (com técnicas de pré-processamento e *bag-of-words*) para visualizar o grau de relacionamento entre os autores das mensagens, para estabelecer as principais relações entre o desempenho vocabular dos alunos e os conceitos tratados no corpus-referência.

A Tabela 3.1 apresenta um comparativo dos trabalhos relacionados com a abordagem proposta nesta dissertação. Para isso, contém os seguintes campos: Trabalhos - referência dos estudos; Presenças - quais as presenças do modelo de investigação que são analisadas; Bases de Dados - origem do conjunto de dados utilizado; Fases Presença Cognitiva (PC) - se utilizam as fases da presença cognitiva que foram estudadas, em caso de uso; Teorias - quais as teorias complementares que contribuíram para a pesquisa (MT), em caso de uso; Ferramentas e Técnicas de MT - quais as principais ferramentas e técnicas de mineração de texto que foram aplicadas, em caso de uso; Identificação PC - como foi realizada a identificação da presença cognitiva; e Língua - qual a língua analisada.

Trabalhos	Presenças	Bases de Dados	Fases PC	Teorias	Ferramentas e Técnicas de MT	Identificação PC	Língua
MCKLIN (2004)	cognitiva	Fórum de discussões	Sim	Não	Redes neurais artificiais; dicionário léxico General Inquirer	Automático	Inglês
CORICH; HUNT; HUNT (2006)	cognitiva	Fórum de discussões	Sim	Não	Remoção de *stopwords*; *Stemming*	Automático	Inglês
KOVANOVIĆ et al. (2014)	cognitiva	Fórum de discussões	Sim	Não	*N-grams*; *Part-of-Speech* e *Back-Off N-grams*; *Dependency Triplets*; *Named Entities*; *Thread Position Features*;validação cruzada; SVM	Automático	Inglês
CHEN (2015)	cognitiva	Fórum de discussões	Sim	Não	*N-grams features*; métodos de vetorização; validação cruzada; MNB; SVM	Automático	Inglês
WATERS et al. (2015)	cognitiva	Fórum de discussões	Sim	Não	*Unigrams*, *Entity Count*, *First and Last Post*, *Comment Depth*, *Post Similarity*, *Word and Sentence counts*, *Number of Replies*; CRFs;	Automático	Inglês
KOVANOVIĆ et al. (2016)	cognitiva	Fórum de discussões	Sim	Não	Coh-Metrix; LIWC; LSA; Entidades Nomeadas; *Discussion context features*; SMOTE; validação cruzada; *Random Forest*	Automático	Inglês
GOMES; PESSOA (2012b)	cognitiva e social	Rede Social (*Facebook*)	Sim	Não	Não	Manual	Português
LAGARTO; MARQUES (2017)	cognitiva e pedagógica	*Chat*	Não	Conhecimento Implícito e Explícito	Não	Manual	Português
SÁ; MACÁRIO (2014)	cognitiva	Fórum de discussões e documentos	Sim	Não	Não	Manual	Português
MACÁRIO; SÁ; MOREIRA (2014)	cognitiva	Fórum de discussões e documentos	Sim	Não	Não	Manual	Português
ROZENFELD (2014)	cognitiva	Fórum de discussões	Sim	Linguística Sistêmico-Funcional e movimentos conversacionais	Não	Manual	Português
ARAUJO (2014)	cognitiva	Fórum de discussões	Sim	Método de Questionamento Socrático e Teoria do Conceito	SPHINX Léxica; remoção de *stopwords*; *stemming*; *bag-of-words*	Manual	Português
Proposta da Dissertação	cognitiva	Fórum de discussões	Sim	Não	Coh-Metrix; LIWC; *Word Embedding*; Entidades Nomeadas; contexto de discussão; SMOTE; validação cruzada; *Random Forest*	Automático	Português

Tabela 3.1: Tabela com a comparação entre os trabalhos

A partir das informações apresentadas na Tabela 3.1 acima, é possível identificar dois diferenciais da abordagem proposta nesta pesquisa frente aos estudos relacionados. O primeiro está relacionado com a utilização de várias ferramentas e técnicas de mineração de texto com bons resultados na literatura. O estudo realizado por ARAUJO (2014) também utiliza técnicas de MT, entretanto aplica somente técnicas de pré-processamento (remoção de *stopwords* e *stemming*) e *bag-of-words*. O outro diferencial está no quesito identificação da presença cognitiva, em que se propõe a automatização da classificação das postagens segundo as fases da presença, assim como da detecção da PC através de mensagens de textos geradas pelos discentes. Como descrito, os estudos analisados no contexto da língua portuguesa realizaram esses processos manualmente, tornando-os trabalhosos e difíceis de serem realizados em um contexto maior. Além disso, em geral, as mensagens desses estudos foram classificadas com o objetivo de avaliar a aprendizagem dos alunos, através da presença cognitiva, considerando apenas a quantidade de mensagens em cada fase.

Sobre os pontos acima é importante ressaltar que, apesar de estarem presentes nos estudos analisados no contexto da língua inglesa, automatizar a análise da presença cognitiva para língua portuguesa não é uma tarefa trivial, devido a escassez de ferramentas e métodos disponíveis para língua, comparado a outras línguas, como o inglês. Assim, como diferencial utilizou-se recursos criados para o português destacados na literatura pelos bons resultados, o que viabiliza o modelo proposto e o torna mais representativo para o contexto em análise. Além disso, um corpus foi constituído com mensagens que foram extraídas de cursos de domínios diferentes, o que torna a proposta desenvolvida mais generalizada.

Além disso, um dos objetivos desse trabalho é realizar uma análise das características relacionadas a presença cognitiva em duas turmas de cursos EaD de contextos diferentes. Não se encontrou na literatura referências que realizassem esse tipo de análise propriamente, mas existem estudos abordando possíveis diferenças na forma como os estudantes do sexo masculino e feminino se comunicam em um fórum de discussões *online*. Este é um dos aspectos analisados nesta dissertação por estar ligado ao desenvolvimento da presença cognitiva.

Desta forma, este trabalho de dissertação apresentará um método para automatizar a identificação da presença cognitiva, utilizando recursos da Mineração de Texto e analisará aspectos da presença em dois contextos distintos, conforme será detalhado nas próximas seções.

4

Metodologia Adotada

Este capítulo descreve a metodologia utilizada para automatizar a análise da presença cognitiva em mensagens de fóruns de discussão assíncronas escritas em português, e está organizado em cinco seções: a seção 4.1 apresenta as etapas realizadas; a seção 4.2 descreve o corpus utilizado para realização da pesquisa; a seção 4.3 apresenta as características extraídas para a composição do modelo proposto; a seção 4.4 descreve o balanceamento realizado no *corpus*; e por fim, a seção 4.5 traz a seleção e avaliação do modelo proposto.

4.1 Etapas da Metodologia

A Figura 4.1 abaixo apresenta as etapas realizadas para a construção do modelo proposto.

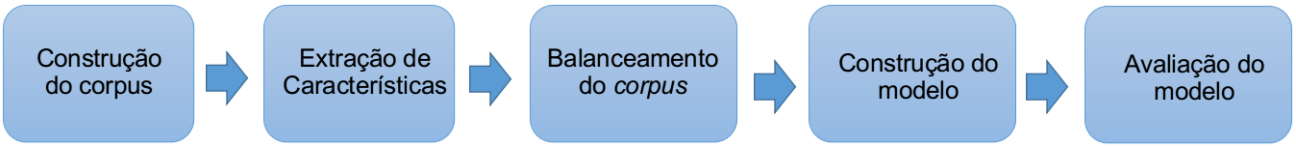

Figura 4.1: Etapas da Metodologia

- **Construção do *corpus*** - processo de coleta das mensagens oriundas de fóruns online, bem como a construção do *corpus* e sua anotação, segundo as fases da presença cognitiva;
- **Extração de Características** - submissão das mensagens do *corpus* nos recursos utilizados na pesquisa para extração das características para construção do modelo;
- **Balanceamento do *corpus*** - balanceamento do *corpus* para treinamento e avaliação do modelo;
- **Construção do modelo** - seleção, treinamento e otimização do classificador;
- **Avaliação do modelo** - avaliação do modelo proposto.

As tarefas realizadas em cada etapa são detalhas nas seções a seguir.

4.2 Descrição do *Corpus*

O procedimento para seleção dos dados de análise seguiu as recomendações de (ROURKE et al., 2001b) e (GARRISON; ANDERSON; ARCHER, 2001). Assim, para esta pesquisa foi utilizado um *corpus* construído a partir da combinação de duas bases de dados compostas de mensagens de fóruns de discussão *online* de dois cursos de ensino superior.

A primeira base (BaseBio) foi gerada com as mensagens trocadas em fóruns de discussão de um curso de graduação em Biologia, oferecido totalmente *online* por uma universidade pública brasileira. Foram extraídas 1.500 mensagens de discussão produzidas por 215 alunos (64 homens e 151 mulheres) durante quatro semanas de curso, conforme descrito na Tabela 4.1 a seguir.

Semanas	**Temas**	**Mensagem (%)**
1	Uso do microscópio	511 (34,06%)
2	Teoria Celular	400 (26,66%)
3	Genética	314 (20,93%)
4	DNA e Clonagem	275 (18,35%)
	Total	1.500 (100%)

Tabela 4.1: Tópicos do curso por semana (BaseBio)

A segunda base de dados (BaseTec) foi gerada através de mensagens retiradas de fóruns de discussão de um curso de graduação em Tecnologia ofertado também de forma totalmente *online* pela mesma universidade. Foram extraídas 734 mensagens de discussão produzidas por 216 alunos (169 homens e 57 mulheres) durante três semanas de curso, como detalhado na Tabela 4.2 adiante.

Semanas	**Temas**	**Mensagem (%)**
1	Evolução das TIC's	278 (37,87%)
2	TIC's na Educação	230 (31,34%)
3	Tecnologias e Aprendizagem	226 (30,79%)
	Total	734 (100%)

Tabela 4.2: Tópicos do curso por semana (BaseTec)

Em relação a quantidade média de mensagens e palavras por mensagem, na primeira base (BaseBio), cada aluno produziu 6,97 mensagens, com cerca de 89 palavras cada uma. Na segunda (BaseTec), foram geradas 3,39 mensagens por aluno, contendo cerca de 113 palavras em cada uma.

A Tabela 4.3 abaixo mostra uma síntese das principais informações relacionadas a alunos e mensagens nos fóruns.

	BaseBio	**BaseTec**
Número de Alunos (Homens)	64	169
Número de Alunos (Mulheres)	151	57
Quantidade total de mensagens	1.500	734
Quantidade média de mensagens por aluno	6,97	3,39
Quantidade média de palavras por mensagem	89	113

Tabela 4.3: Resumo dos dados das bases

Quanto ao objetivo dos fóruns, em ambas as bases, o foco foi promover discussões sobre um tema proposto pelo professor, em que a participação representava 20% da nota final do curso. As discussões foram principalmente do tipo perguntas e respostas, ou seja, os fóruns eram iniciados com uma pergunta pelo professor e os alunos deveriam responder deixando suas contribuições, que consistiram em respostas a pergunta inicial e/ou novas perguntas relacionadas ao tema abordado.

A união das bases gerou um *corpus* contendo 2.234 mensagens, que foi anotado por dois codificadores independentes, ambos conhecedores dos conceitos referentes a presença cognitiva e com conhecimento sobre o processo de análise de conteúdo, que analisaram as mensagens em separado, segundo as quatro fases da presença cognitiva (Evento desencadeador, Exploração, Integração e Resolução).

Para o processo de codificação escolheu-se a mensagem como unidade de análise, conforme indicado por (GARRISON; ANDERSON; ARCHER, 2001). Assim, os codificadores analisaram cada mensagem do corpus, guiados pelos indicadores e/ou descritores de cada uma das quatro fases da PC, com intuito de identificar características que definissem suas classes. Ou seja, cada mensagem era categorizada em apenas uma fase (ROURKE et al., 2001b).

No caso de uma mensagem refletir múltiplas fases, foram utilizadas as duas heurísticas sugeridas por (GARRISON; ANDERSON; ARCHER, 2001) para a categorização das mensagens: *code down* (ou seja, para a fase mais baixa), se não estiver claro qual fase é refletida; e *code up* (isto é, para a fase mais alta), se estiverem presentes evidências claras de múltiplas fases.

Para as mensagens que, após a análise, não exibiram os indicadores de nenhuma fase de presença cognitiva foi criada a categoria “Outros”.

A Tabela 4.4 abaixo apresenta mensagens retiradas do corpus e suas respectivas classes, para exemplificar o resultado da anotação. Os códigos 0, 1, 2, 3 e 4, referem-se as categorias: Outros, Evento Desencadeador, Exploração, Integração e Resolução, respectivamente.

Mensagem	Categoria
Ok ok !	0
Muito bem! Além do estudo de seres unicelulares onde mais podemos encontrar a aplicação dos "poderes"do microscópio?	1
Como o olho humano possui a resolução de 0.2 mm, com o microscópio foi possível estudar os seres unicelulares e pluricelulares invisíveis a olho nu que incluem bactérias e protistas que são responsáveis pela maioria das doenças hoje conhecidas.	2
Lembrando colega que ao estudarmos a origem e a evolução dos seres vivos, falamos em origem evolução da célula. Afinal, com exceção do vírus, os seres vivos são formados por células, e a compreensão de como eles surgiram e evoluíram passa pela compreensão de como a célula surgiu e evoluiu. O primeiro ser vivo que surgiu no planeta terra era uma célula.	3
Gostei desse tema, porque pude perceber o quanto evoluímos com essa criação, graças a ele podemos nos dar o luxo de termos essa variedade de tecnologias a nossa disposição, não digo somente na nossa área mas em um contexto geral, vejamos o processador de um simples *smartphone* quantas coisa você carrega em seu bolso ,calculadora, aplicativos de mensagens, Maquina fotográfica etc. tudo isso graças a invenção do microscópio,processadores mais robustos e minúsculos, Nos ajudarão a evoluir em vários aspectos. os primeiros processadores já contavam com milhares de transistores. Os mais evoluídos passaram para os milhões. E os atuais chegam a bilhões, Imagine isso ser possível sem os microscópios.	4

Tabela 4.4: Exemplo de mensagens anotadas

Ao final do processo de codificação, calculou-se o grau de concordância obtido entre os avaliadores por meio do índice Kappa. Assim, obteve-se no *corpus* uma concordância percentual de 89,17% e $\kappa = 0{,}82$, um indicador excelente. Os desacordos (242 no total) foram resolvidos por um terceiro codificador (seguindo os critérios supracitados, decidiu entre as escolhas dos dois codificadores).

As Tabelas 4.5, 4.6 e 4.7 mostram a distribuição das quatro fases da presença cognitiva e a categoria "Outros", nos conjuntos de dados BaseBio, BaseTec e *corpus*, respectivamente.

ID	Fase	Mensagens (%)
0	Outros	196 (13,07%)
1	Evento Desencadeador	235 (15,67%)
2	Exploração	871 (58,07%)
3	Integração	154 (10,27%)
4	Resolução	44 (2,92%)
	Total	1.500 (100%)

Tabela 4.5: Presença Cognitiva (BaseBio)

ID	Fase	Mensagens (%)
0	Outros	17 (2,32%)
1	Evento Desencadeador	33 (4,50%)
2	Exploração	462 (62,94%)
3	Integração	84 (11,44%)
4	Resolução	138 (18,80%)
	Total	734 (100%)

Tabela 4.6: Presença Cognitiva (BaseTec)

ID	Fase	Mensagens (%)
0	Outros	213 (9,53%)
1	Evento Desencadeador	268 (12,00%)
2	Exploração	1.333 (59,67%)
3	Integração	238 (10.65%)
4	Resolução	182 (8,15%)
	Total	2.234 (100%)

Tabela 4.7: Presença Cognitiva (*corpus*)

As mais frequentes foram as mensagens de Exploração, correspondendo por mais de 59,67% dos dados, enquanto as menos frequentes foram as mensagens de Resolução, representando apenas 8,15% dos dados. A diferença substancial entre as frequências das fases de presença cognitiva era esperada (GARRISON; ANDERSON; ARCHER, 2010) e também relatada em estudos anteriores do modelo CoI (KOVANOVIĆ et al., 2014, 2016).

Na literatura é possível encontrar várias explicações para o padrão encontrado na distribuição de mensagens entre as fases (AKYOL et al., 2009). Neste caso particular, o fórum apresentou características de discussões do tipo perguntas e respostas. Assim, parece razoável que os alunos realizem mais perguntas (Evento desencadeador). Além disso, como as discussões foram planejadas para ocorrerem entre a primeira e a última semana dos cursos, os alunos não costumam passar para a fase de Resolução logo no início do curso.

4.3 Extração de Características

Este trabalho seguiu a mesma metodologia apresentada por KOVANOVIĆ et al. (2016), na qual as características tradicionais de classificação de texto (por exemplo, N-gramas, *Part-of-Speech*, *dependency triplets*) não foram adotadas. Em primeiro lugar, pelo fato dessas características gerarem várias novas características, até mesmo para bases de dados pequenas, aumentando as chances de *overfitting* (quando o modelo criado está tão bem ajustado aos dados, que não consegue generalizar bem para novas amostras) dos dados de treinamento.

Em segundo lugar, essas características são muito “dependentes do conjunto de dados”, já que os dados definem o espaço de classificação. Assim, é difícil definir antecipadamente um conjunto fixo de características para classificação, uma vez que a escolha particular de palavras nos documentos de treinamento definirá quais delas serão usadas para classificação. Por fim, o fato dos N-gramas e outras características mais simples de mineração de texto não serem baseadas em nenhuma teoria existente de cognição humana relacionada à comunidade de investigação, pode levar a modelos que são difíceis de entender seu significado teórico. Dessa forma, concentrou-se na extração de características fortemente orientadas pela teoria e baseadas em estudos empíricos. Assim, considerando que os recursos e ferramentas para análise de texto em português disponíveis são limitados, e os experimentos realizados por KOVANOVIĆ et al. (2016), 127 características foram coletadas para os experimentos. Os detalhes de cada uma são apresentados a seguir.

4.3.1 Características LIWC

A ferramenta LIWC (seção 2.3.1) possibilita a extração de um grande número de contagens de palavras que são indicativas de diferentes processos psicológicos (por exemplo, afetivo, cognitivo, social, perceptual). Estudos destacam os bons resultados obtidos com a utilização da ferramenta em contextos relevantes para esta pesquisa, como TAUSCZIK; PENNEBAKER (2010), para a identificação de traços de personalidades; MACHADO et al. (2015), na busca dos sentimentos associados às palavras; PAIM; CAMATI; ENEMBRECK (2016), para extração de caraterísticas linguísticas empregadas no texto; e WEN; YANG; ROSÉ (2014) Wen et al. (2014) que realizou análise linguística aplicadas às postagens de fóruns para encontrar a relação entre os textos escritos, a motivação e o envolvimento cognitivo.

Desta forma, foram utilizadas neste estudo as 64 categorias disponibilizadas pela ferramenta na versão em português brasileiro (BALAGE FILHO; PARDO; ALUÍSIO, 2013) como características, incluindo aquelas que alcançaram os melhores resultados para o classificador de presença cognitiva no estudo realizado por KOVANOVIĆ et al. (2016). Além disso, considerou-se relevante para pesquisa a inclusão de quatro novas características, oriundas das categorias do léxico LIWC de língua inglesa (PENNEBAKER; FRANCIS; BOOTH, 2001). Essas são independentes de idiomas, pelo fato serem apenas contagem de características textuais. Assim, as características são formadas pelo total da quantidade de: palavras, termos encontrados no léxico, palavras com mais de seis letras e palavras por frase.

A Tabela 4.8 a seguir detalha as 68 categorias, em suas respectivas dimensões, utilizadas como características.

Nº	**Características**	**Descrição**
Linguística		
1	Total de palavras	Número total de palavras
2	Termos Dicionário	Número total de termos encontrados no léxico

3	Palavras > 6 letras	Número de palavras com mais de seis letras
4	Palavras por frase	Número de Palavras por frase
5	Função de palavras	Total de palavras com função
6	Total de pronomes	Número total de pronomes
7	Pronome pessoal	Número de pronomes pessoais
8	1ª pessoa do singular	Número de pronomes em primeira pessoa do singular
9	1ª pessoa do plural	Número de pronomes em primeira pessoa do plural
10	2ª pessoa	Número de pronomes em segunda pessoa
11	3ª pessoa do singular	Número de pronomes em terceira pessoa do singular
12	3ª pessoa do plural	Número de pronomes em terceira pessoa do plural
13	Pronomes impessoais	Número de pronomes impessoais
14	Artigos	Número de artigos
15	Verbos comuns	Número de verbos
16	Verbos auxiliares	Número de verbos auxiliares
17	Passado	Número de verbos no passado
18	Presente	Número de verbos no presente
19	Futuro	Número de verbos no futuro
20	Advérbios	Número de advérbios
21	Preposição	Número de preposições
22	Conjunções	Número de conjunções
23	Negação	Número de palavras que expressam negação
24	Quantificadores	Número de quantificadores
25	Números	Total de palavras numéricas
26	Palavrões	Número de palavrões
Psicológica		
27	Social	Número de palavras relacionadas a interação entre as pessoas (ex.: companheiro, conversa)
28	Família	Número de palavras que fazem referência a membros da família (ex.: filho, tia)
29	Amigos	Número de palavras que fazem referência a amizade (ex.: amigo, vizinho)
30	Humano	Número de palavras que fazem referência a seres humanos (ex.: adulto, garoto)
31	Afetivo	Número de palavras que expressam sentimentos (ex.: feliz, chorão)
32	Emoções positivas	Número de palavras que expressam emoções positivas (ex.: feliz, bondade)

33	Emoções negativas	Número de palavras que expressam emoções negativas (ex.: feio, desagradável)
34	Ansiedade	Número de palavras que expressam ansiedade (ex.: medo, neurose)
35	Raiva	Número de palavras que expressam raiva (ex.: matar, aborrecido)
36	Tristeza	Número de palavras que expressam tristeza (ex.: triste, chorando)
37	Cognitivos	Número de palavras que fazem referência a características cognitivos (ex.: porque, sabedoria)
38	Intuição	Número de palavras que expressam intuição (ex.: penso, conheço)
39	Causa	Número de palavras que expressam causa (ex.: executado, efeito)
40	Discordância	Número de palavras que expressam discordância (ex.: expectativa, deveria)
41	Tentativa	Número de palavras que expressam tentativa (ex.: talvez, adivinhe)
42	Certeza	Número de palavras que expressam certeza (ex.: sempre, nunca)
43	Inibição	Número de palavras que expressam inibição (ex.: restringir, bloquear)
44	Inclusivo	Número de palavras que expressam inclusão (ex.: com, e)
45	Exclusivo	Número de palavras que expressam exclusão (ex.: mas, excluir)
46	Perceptivo	Número de palavras que expressam percepção (ex.: observando, ouvindo)
47	Ver	Número de palavras que fazem referência a visão (ex.: ver, visto)
48	Ouvir	Número de palavras que fazem referência a audição (ex.: ouça, ouvindo)
49	Sentir	Número de palavras que fazem referência ao sentir (ex.: sente, toque)
50	Biológico	Número de palavras que fazem referência a características biológicas (ex.: sangue, dor)

51	Corpo	Número de palavras que fazem referência ao corpo (ex.: bochecha, mãos)
52	Saúde	Número de palavras que fazem referência à saúde (ex.: gripe, pílula)
53	Sexual	Número de palavras que fazem referência ao sexo (ex.: tesão, amor)
54	Ingestão	Número de palavras que fazem referência a ingestão (ex.: prato, pizzas)
55	Relatividade	Número de palavras que fazem referência ao relativo (ex.: dobre, pare)
56	Movimento	Número de palavras que fazem referência a movimento (ex.: carro, vai)
57	Espaço	Número de palavras que fazem referência a espaço (ex.: abaixo, fina)
58	Tempo	Número de palavras que fazem referência ao tempo (ex.: fim, temporada)
59	Trabalho	Número de palavras que fazem referência a trabalho (ex.: trabalho, abandono)
60	Conquista	Número de palavras que fazem referência a conquistas (ex.: ganhe, herói)
61	Lazer	Número de palavras que fazem referência a lazer (ex.: bate-papo, filme)
62	Casa	Número de palavras que fazem referência ao lar (ex.: família, cozinha)
63	Dinheiro	Número de palavras que fazem referência a dinheiro (ex.: dinheiro, auditoria)
64	Religião	Número de palavras que fazem referência a religião (ex.: altar, igreja)
65	Morte	Número de palavras que fazem referência a morte (ex.: enterre, mate)
Falada		
66	Concordância	Número de palavras que fazem referência a concordância (ex.: concordo, OK)
67	Sem fluência	Número de palavras do tipo onomatopeia (figura de linguagem que permite o uso de vocábulos para representar som) (ex.: er, hm)

68	Enchimento	Número de palavras que fazem referência a enchimento (ex.: blá, sacou)

Tabela 4.8: Características LIWC

4.3.2 Características Coh-Metrix

O Coh-Metrix (seção 2.3.2) é uma ferramenta que calcula medidas e índices de coesão e coerência de textos num amplo contexto de medidas. A versão em português do Coh-Metrix, conhecida como Coh-Metrix-Port, tem 48 medidas diferentes (enquanto a versão inglesa tem 108). Entretanto, é importante mencionar que todas as características da versão em português obtiveram bons resultados na classificação da presença cognitiva para o inglês. Além disso, estudos como TOLEDO et al. (2014), CUNHA et al. (2013), (CROSSLEY, 2013) destacaram o uso das métricas da ferramenta como características para o desenvolvimento de classificadores.

Ao conjunto de medidas disponibilizado pela ferramenta para o português brasileiro, foram acrescentadas quatro novas categorias, inspiradas na versão de língua inglesa, destacadas por KOVANOVIĆ et al. (2016) pelos bons resultados obtidos. Assim, foram incluídas: *givenness*, *LDVOCD*, *LDContW* e *Tokens*.

A Tabela 4.9 mostra as medidas utilizadas como características para classificação e uma breve descrição de cada uma.

Nº	**Características**	**Descrição**
Contagens Básicas		
1	Índice Flesch	Índice Flesch (medida de complexidade do texto associada à sua inteligibilidade para diferentes tipos de leitores).
2	Número de Palavras	Número de palavras do texto.
3	Número de Sentenças	Número de sentenças de um texto.
4	Número de Parágrafos	Número de parágrafos de um texto. Parágrafos são apenas onde há quebra de linha (não identações).
5	Palavras por Sentenças	Número de palavras dividido pelo número de sentenças.
6	Sentenças por Parágrafos	Número de sentenças dividido pelo número de parágrafos.
7	Sílabas por Palavras de Conteúdo	Número médio de sílabas por palavras de conteúdo (substantivos,verbos, adjetivos e advérbios).
8	Incidência de Verbos	Incidência de verbos em um texto.
9	Incidência de Substantivos	Incidência de substantivos em um texto.
10	Incidência de Adjetivos	Incidência de adjetivos em um texto.
11	Incidência de Advérbios	Incidência de advérbios em um texto.
12	Incidência de Pronomes	Incidência de pronomes em um texto.

13	Incidência de Palavras de Conteúdo	Incidência de Palavras de Conteúdo (substantivos, adjetivos, advérbios e verbos).
14	Incidência de Funcionais	Incidência de Palavras Funcionais (artigos, preposições, pronomes, conjunções e interjeições).
	Operadores Lógicos	
15	Incidência de Operadores Lógicos.	Incidência de operadores lógicos em um texto. São considerados como operadores lógicos: e, ou, se, negações e um número de condições.
16	Incidência de *E*	Incidência do operador lógico *E* em um texto.
17	Incidência de *OU*	Incidência do operador lógico *OU* em um texto.
18	Incidência de *SE*	Incidência do operador lógico *SE* em um texto.
19	Incidência de Negações	Incidência de Negações. São considerados como negações: não, nem, nenhum, nenhuma, nada, nunca e jamais.
	Frequências	
20	Frequências	Média de todas as frequências das palavras de conteúdo encontradas no texto. O valor da frequência das palavras é retirado da lista de frequências do corpus Banco do Português.
21	Mínimo Frequências	Identifica-se a menor frequência dentre todas as palavras de conteúdo em cada sentença. Depois, calcula-se uma média de todas as frequências mínimas. A palavra com a menor frequência é a mais rara da sentença.
	Hiperônimos	
22	Hiperônimos de verbos	Hiperônimos de verbos. (Hiperônimos é uma palavra que pertence ao mesmo campo semântico de outra mas com o sentido mais abrangente. Ex. Animais é hiperônimo de cachorro e cavalo).
	Pronomes, Tipos e *Token*	
23	Incidência de Pronomes Pessoais	Incidência de pronomes pessoais em um texto. São considerados como pronomes pessoais: eu, tu, ele/ela, nós, nós, eles/elas, você e vocês.
24	Pronomes por Sintagmas	Média do número de pronomes que aparecem em um texto pelo número de sintagmas. (O temo sintagma é comumente empregado para se referir às partes da sentença).

25	*Type/Token*	Número de palavras únicas dividido pelo número de *tokens* dessas palavras. Cada palavra única é um tipo. Cada instância desta palavra é um *token*.
	Constituintes	
26	Incidência de Sintagmas	Incidência de sintagmas nominais por 1000 palavras. (Sintagmas nominais é quando o núcleo do sintagma é um nome. Ex. Os alunos gostaram do fórum - Sintagma1: Os alunos e Sintagma2: do tema).
27	Modificadores por Sintagmas	Média do número de modificadores por sintagmas nominais, adjetivos, advérbios e artigos, que participam de um sintagma.
28	Palavras antes de verbos principais	Média de palavras antes de verbos principais na cláusula principal da sentença.
	Conectivos	
29	Incidência de Conectivos	Incidência de todos os conectivos que aparecem em um texto.
30	Conectivos Aditivos Positivos	Incidência de conectivos classificados como aditivos positivos.
31	Conectivos Aditivos Negativos	Incidência de conectivos classificados como aditivos negativos.
32	Conectivos Temporais Positivos	Incidência de conectivos classificados como causais positivos.
33	Conectivos Temporais Negativos	Incidência de conectivos classificados como causais negativos.
34	Conectivos Causais Positivos	Incidência de conectivos classificados como lógicos positivos.
35	Conectivos Causais Negativos	Incidência de conectivos classificados como causais negativos.
36	Conectivos Lógicos Positivos	Incidência de conectivos classificados como lógicos positivos.
37	Conectivos Lógicos Negativos	Incidência de conectivos classificados como lógicos negativos.
	Ambiguidades	
38	Verbos	Ambiguidade de Verbos.
39	Substantivos	Ambiguidade de Substantivos.
40	Adjetivos	Ambiguidade de Adjetivos.
41	Advérbios	Ambiguidade de Advérbios.

Correferência		
42	Sobreposição de argumentos adjacentes	Sobreposição de argumentos em sentenças adjacentes.
43	Sobreposição de argumentos	Sobreposição de argumentos em todos os pares de sentenças.
44	Sobreposição de radicais de palavras adjacentes	Sobreposição de argumentos em sentenças adjacentes.
45	Sobreposição de radicais de palavras	Sobreposição de radicais de palavras em todos os pares de sentenças.
46	Sobreposição de palavras de conteúdo	Sobreposição de palavras de conteúdo em sentenças adjacentes.
Anáforas		
47	Referência anafórica adjacente	Referência anafórica em sentenças adjacentes. (Anáfora consiste na repetição de uma ou mais palavras no início de orações, períodos ou versos sucessivos).
48	Referência anafórica	Referência anafórica em até cinco sentenças anteriores.
Acréscimos		
49	*Givenness*	Média da similaridade entre cada sentença e todo o texto que a precede.
50	LDVOCD	Diversidade lexical, VOCD.
51	LDContW	Diversidade lexical, palavras de conteúdo.
52	*Token*	Número de *tokens*.

Tabela 4.9: Características Coh-Metrix

Assim, com o intuito de coletar nos discursos diferentes indicadores que tenham relevância no processo sociocognitivo, utilizaram-se as medidas fornecidas pela ferramenta Coh-Metrix, que podem ser facilmente extraídas e usadas para automação dos esquemas de codificação do modelo da CoI.

4.3.3 Características de Contexto da Discussão

Para incorporar mais informação de contexto ao conjunto de características deste trabalho, foram incluídas as características de contexto propostas por WATERS et al. (2015) e usadas por KOVANOVIĆ et al. (2016). No total foram adicionadas cinco, que são apresentadas na Tabela 4.10.

Nº	**Características**	**Descrição**
1	Número de respostas	Uma variável inteira indicando o número de respostas que uma determinada mensagem recebeu.

2	Profundidade da Mensagem	Uma variável inteira mostrando posição de uma mensagem dentro de uma árvore de discussão.
3	Similaridade de Cosseno para a mensagem anterior	A ideia é obter o quanto a mensagem atual se baseia nas informações anteriormente apresentadas.
4	Similaridade de Cosseno para a mensagem seguinte	A ideia é obter o quanto a mensagem atual se baseia nas informações seguintes.
5	Indicadores de início/fim	Utiliza um indicador (0/1) que mostra se uma mensagem é a primeira/última da discussão.

Tabela 4.10: Características de Contexto da Discussão

A relevância das características acima descritas para o problema em estudo está relacionada com à natureza processual do modelo CoI (GARRISON; ANDERSON; ARCHER, 2010), em que a presença cognitiva dos alunos é vista como sendo desenvolvida ao longo do tempo através do discurso e da reflexão. Para atingir níveis mais altos de presença cognitiva, os estudantes precisam: i) construir conhecimento no ambiente compartilhado por meio da troca de uma determinada quantidade de mensagens de discussão, ou ii) construir conhecimento de forma reflexiva em seu próprio ambiente privado de aprendizagem.

Outro aspecto refere-se à visão sócio-construtivista da aprendizagem no modelo CoI (SWAN; GARRISON; RICHARDSON, 2009), em que a aprendizagem colaborativa consiste em uma estratégia pedagógica que envolve atividade intelectual conjunta desenvolvida pelos aprendizes. Estes trabalham de forma colaborativa para construir sentido ou desenvolver compreensão, solucionando problemas e criando produtos em conjunto (GARRISON; CLEVELAND-INNES; FUNG, 2010). Neste sentido, as diferentes fases da presença cognitiva tendem a mudar ao longo do tempo. Desta forma, esperava-se que as mensagens do tipo Evento desencadeador e Exploração fossem mais frequentes nos estágios iniciais das discussões, enquanto as mensagens de Integração e Resolução seriam mais comuns nos estágios posteriores, uma vez que o aprendizagem e o seu desenvolvimento são produtos da interação social.

4.3.4 Similaridade *Word embedding*

Em seu trabalho, KOVANOVIĆ et al. (2016) fez um paralelo entre as fases cognitivas e as informações apresentadas nas diversas etapas do processo de aprendizagem. Em resumo, a fase Evento desencadeador introduz um tópico, enquanto a fase de Exploração introduz novas ideias e respostas. A fase de Integração continua falando sobre as mesmas ideias (construindo o significado a partir das ideias previamente apresentadas), e a Resolução conclui a discussão apresentando as diretrizes explícitas para a aplicação do conhecimento construído (PARK, 2009).

Devido aos motivos listados acima, considerou-se benéfico ter uma característica que

possa identificar se o contexto de cada mensagem muda ao longo do tempo em uma discussão. A principal diferença relacionada ao trabalho original de KOVANOVIĆ et al. (2016) é que o estudo atual adotou *word embeddings* (seção 2.2.3.3) para representar similaridade de palavras em vez de LSA. *Word embedding* refere-se a representação de palavras em um espaço vetorial, de maneira a gerar propriedades específicas. Palavras similares estão relativamente próximas no espaço vetorial e acabam por demonstrar algumas particularidades (MIKOLOV et al., 2013). De acordo com (KUSNER et al., 2015), o método faz uso de algoritmos de redes neurais para traduzir palavras em vetores numéricos com base em suas ocorrências em um texto.

Na literatura é possível encontrar estudos recentes que apontam para popularidade e os bons resultados obtidos com a utilização destes modelos para representações vetoriais de palavras. NAILI; CHAIBI; GHEZALA (2017) relataram em seu estudo que, na maioria dos casos, o modelo de *embeddings* Word2Vec mostrou ser mais eficientes que o LSA. BARONI; DINU; KRUSZEWSKI (2014) mostraram a eficiência do Word2Vec quando comparado com os métodos tradicionais, como *Pointwise Mutual Information* (PMI). Da mesma forma, LEVY; GOLDBERG; DAGAN (2015) afirmam que o Word2Vec supera o GloVe em muitas tarefas, como a similaridade de palavras.

LEVY; GOLDBERG (2014) destacam que os modelos baseados em *word embeddings* são fáceis de trabalhar, pois proporcionam um cálculo eficiente da similaridade entre as palavras através de operações de matrizes de baixa dimensão. Além disso, essa representação colabora para que algoritmos de aprendizado de máquina atinjam melhores resultados em tarefas de PLN, através do agrupamento de palavras similares (MIKOLOV et al., 2013).

Neste trabalho, utilizou-se a sentença como unidade de medida para criar uma única característica que representasse a similaridade média da sentença (ou seja, coerência) dentro de uma mensagem. Para a realização desta tarefa foram aplicados os algoritmos de *word embeddings* e o conjunto de dados treinados disponíveis na ferramenta *spaCy*.

4.3.5 Número de Entidades Nomeadas

Trabalhos como MU et al. (2012) e KOVANOVIĆ et al. (2014) sugerem que o número de entidades nomeadas (por exemplo, objetos nomeados como pessoas, organizações e localizações geográficas) (seção 2.2.3.4) seria diferente para cada uma das fases da presença cognitiva.

Assim, espera-se, por exemplo, que as mensagens de Exploração, em que os alunos, após compreender a natureza do problema, são levados a explorar fontes para obter informações mais relevantes, como novos conceitos e opiniões, tenham mais entidades nomeadas do que mensagens de integração e resolução. Para executar a tarefa de extrair o número de entidades nomeadas que foram mencionadas na mensagem, com o intuito de verificar a quantidade existente em cada fase e sua a relevância para distingui-las, foi utilizada a biblioteca *spaCy*.

A Tabela 4.11 abaixo mostra uma síntese das principais informações relacionadas aos recursos utilizados para extração de características.

Nº	Recurso	Descrição	Nº de Características
1	Dicionário léxico LIWC	Contagens de palavras que são indicativas de diferentes processos psicológicos.	68
2	Coh-Metrix	Calcula medidas e índices de coesão e coerência de textos.	52
3	Contexto da Discussão	Características de contexto.	5
4	Similaridade *Word embedding*	Similaridade de palavras	1
5	Entidades Nomeadas	Objetos nomeados como nomes de pessoas, organizações, entidades locais, medições, tempo.	1
		Total	127

Tabela 4.11: Resumo dos recursos

4.4 Balanceamento do *Corpus*

Nesta etapa foram utilizados os recursos da biblioteca *scikit-learn*, que possibilita a utilização de aprendizado de máquina em projetos, sejam eles acadêmicos ou comerciais (VAROQUAUX et al., 2015).

A primeira etapa da análise de dados realizada foi a divisão dos dados em dois conjuntos: treinamento e teste, com 75% dos dados reservados para treinamento e os outros 25% para testes, como frequentemente ocorre no aprendizado de máquina (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Essa divisão tem como objetivo separar os dados que o algoritmo de classificação usará para treinar dos dados para teste. Esta etapa foi realizada para evitar que as figuras de desempenho do modelo sejam superestimadas, fato que pode ocorrer se a acurácia do modelo for calculada com os mesmos dados utilizados como parâmetro para o modelo. Ou seja, usando os mesmos dados para treinamento e teste pode-se obter resultados muito otimistas e não reais.

A respeito da proporção, é importante mencionar que as amostras separadas relativas às categorias de codificação (ou seja, Evento desencadeador, Exploração, Integração, Resolução e Outros) foram criadas para preservar sua distribuição em subconjuntos de treinamento e de teste. Dessa forma, do conjunto de dados total, 1.689 instâncias foram utilizadas para compor a base de dados de treinamento e 545 instâncias para a base de teste, conforme descrito na Tabela 4.12 a seguir.

	Base					
Fase	Treinamento		Teste		Total	
Outros	178	(10,54%)	35	(6,42%)	213	(9,53%)
Evento desencadeador	217	(12,85%)	51	(9,36%)	268	(12%)
Exploração	1.004	(59,44%)	329	(60,37%)	1.333	(59,67%)
Integração	152	(9%)	86	(15,78%)	238	(10,65%)
Resolução	138	(8,17%)	44	(8,07%)	182	(8,15%
Total	1.689	(100%)	545	(100%)	2.234	(100%)

Tabela 4.12: Distribuição das categorias de codificação nas bases de teste e treinamento

Para obter a amostragem estratificada dos dados de treinamento e teste utilizou-se a biblioteca *scikit-learn*, que fornece ferramentas de aprendizado de máquina eficientes e bem estabelecidas em várias áreas científicas (BUITINCK et al., 2013).

Após o particionamento de *corpus*, identificou-se o problema do desequilíbrio de classes, como mostrado na Tabela 4.12. Esse problema pode levar a efeitos negativos nos resultados das análises de classificação (ROSÉ et al., 2008), uma vez que todas as cinco classes possuem a mesma relevância para esta pesquisa (quatro fases da presença cognitiva e as outras mensagens), pois representam fases diferentes nos ciclos de aprendizado dos alunos. Em razão dos bons resultados obtidos no trabalho desenvolvido por KOVANOVIĆ et al. (2016), adotou-se o algoritmo SMOTE (CHAWLA et al., 2002), que cria sinteticamente instâncias de classes adicionais como uma combinação linear de instâncias existentes.

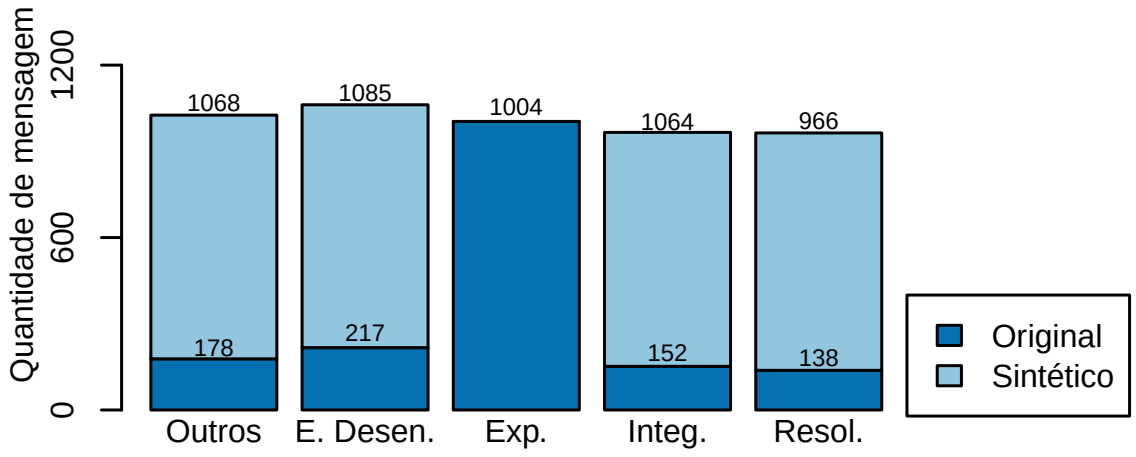

Figura 4.2: Balanceamento de classe com SMOTE

A Figura 4.2 acima apresenta o resultado final da aplicação do algoritmo SMOTE no conjunto de treinamento do *corpus*. O tamanho dos códigos das categorias foram aumentados de 5 a 7 vezes.

4.5 Seleção e Avaliação do Modelo

O conjunto de treinamento submetido a um algoritmo de classificação é responsável por criar o modelo de classificação que categorizará novas instâncias. A escolha desse algoritmo é uma etapa importante na classificação de texto, que demanda esforço e compreensão para definir qual o melhor para um dado problema.

Existem diversos algoritmos de aprendizado de máquina para construir modelos supervisionados. O estudo realizado por FERNÁNDEZ-DELGADO et al. (2014) apresenta uma análise comparativa realizada com 179 algoritmos de classificação de propósito geral em 121 conjuntos de dados diferentes. Os resultados mostraram que os algoritmos *random forest* e SVM com kernel gaussiano apresentaram os melhores desempenhos. Este trabalho utilizou *random forest*, pois, além de seu excelente desempenho, desejava-se avaliar até que ponto cada característica contribuiu para o classificador e identificar aquelas que mais influenciaram na classificação das fases da presença cognitiva - algoritmo tipo caixa branca [1] (BREIMAN, 2001).

Dentre as medidas que permitem a avaliação da importância das características da classificação, a mais utilizada é o *Mean Decrease Gini* (MDG), que explica a separabilidade de uma determinada característica em relação às categorias (BREIMAN, 2001).

Para o desenvolvimento do modelo preditivo utilizaram-se os métodos disponíveis do pacote R, considerando que a linguagem R é muito usada em ambientes de aprendizagem de máquina e, que seus métodos são consolidados e robustos para aplicações no contexto (LANTZ, 2013). No classificador foi utilizado o método *randomForest* do pacote R (LIAW; WIENER et al., 2002). Foram estabelecidos dois parâmetros para serem utilizados no classificador *random forest*: (i) *ntree*: o número de árvores geradas pelo algoritmo; e (ii) *mtry*: o número de características aleatórias selecionadas por cada árvore. Aqui, valores diferentes para cada parâmetro foram avaliados sobre os dados de treinamento usando validação cruzada de 10 execuções, com o intuito de otimizar os parâmetros. Em ambos os casos, os valores que maximizam o desempenho final foram selecionados.

[1] Os algoritmos de caixa preta processam informações sem revelar o impacto exato das entradas, produzindo, assim, modelos difíceis de entender. Algoritmos de caixa branca fornecem informações claras sobre o impacto da entrada. Um algoritmo de caixa cinza se refere a um algoritmo que cria soluções que potencialmente poderiam ser compreendidas, mas requerem processamento adicional para que as informações sejam compreensíveis ou rastreáveis (LJUNG, 2001).

5

Resultados

Neste capítulo são apresentados os resultados obtidos com os experimentos realizados e as discussões. Serão exibidas as 20 características mais relevantes para o classificador cada domínio que compõe o corpus (Biologia e Tecnologia) considerando o *corpus* utilizado, assim como o desempenho alcançado pelo mesmo, para realizar as discussões sobre os dados obtidos e suas particularidades.

5.1 Modelo de treinamento e avaliação

Para esta etapa utilizou-se a biblioteca *Caret* (*Classification and Regression Training*) do pacote R (KUHN et al., 2017). A biblioteca contém uma série de funções para serem aplicadas no pré-processamento, treinamento e validação de algoritmos de aprendizado de máquina, além de possuir um sistema de encapsulamento que possibilita o uso de bibliotecas de técnicas de aprendizado de máquina já consolidadas na literatura.

Como foi dito anteriormente, foram selecionados dois parâmetros para serem utilizados no classificador *random forest*: *ntree* e *mtry*. A primeira parte do experimento foi dedicada a encontrar bons valores para eles.

Na melhor das hipóteses, o classificador proposto alcançou um desempenho com acurácia de 0,96 (desvio padrão de 0,01) e κ de 0,95 (desvio padrão de 0,01). Estes resultados foram alcançados com seis características por árvore de decisão no conjunto de dados de treinamento (*mtry* = 6).

mtry	Acurácia	Kappa (κ)
6	0,96 (0,01)	0,95 (0,01)
10	0,95 (0,01)	0,94 (0,01)
15	0,95 (0,01)	0,94 (0,01)
19	0,95 (0,01)	0,93 (0,02)
24	0,94 (0,01)	0,93 (0,01)
28	0,94 (0,01)	0,93 (0,01)
33	0,94 (0,01)	0,93 (0,02)
37	0,94 (0,01)	0,92 (0,02)
42	0,94 (0,01)	0,92 (0,02)
46	0,93 (0,01)	0,92 (0,02)
51	0,93 (0,01)	0,92 (0,02)
55	0,93 (0,02)	0,92 (0,02)
60	0,93 (0,02)	0,92 (0,02)
64	0,93 (0,02)	0,91 (0,02)
69	0,93 (0,02)	0,91 (0,02)
73	0,93 (0,02)	0,91 (0,02)
78	0,92 (0,02)	0,90 (0,02)
82	0,92 (0,02)	0,90 (0,02)
87	0,92 (0,02)	0,90 (0,02)
Min	0,92	0,90
Max	0,96	0,95
Max-Min	0,04	0,05
Média	0,94	0,92

Tabela 5.1: Resumo da otimização dos parâmetros

Entre o modelo de melhor e pior desempenho houve a melhoria de 0,04 e 0,05 (Tabela 5.1) para acurácia da classificação e κ, respectivamente. Essa melhoria demonstra a importância da otimização dos parâmetros no desempenho final.

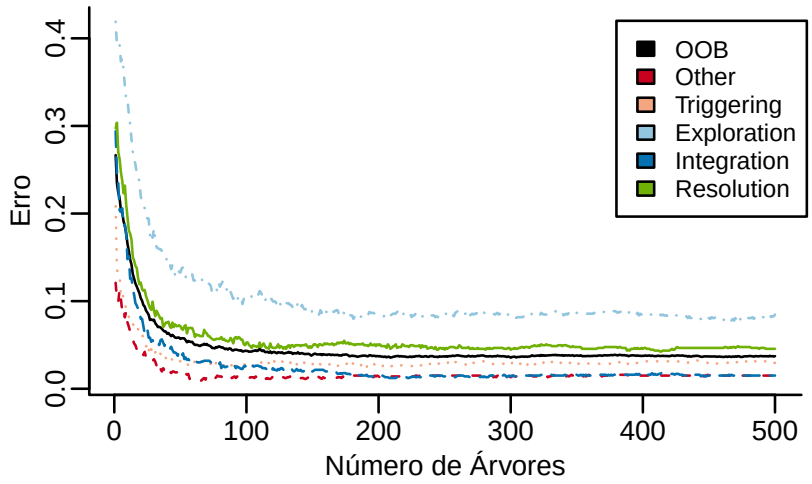

Figura 5.1: Melhor desempenho do classificador *random forest*

A Figura 5.1 apresenta o desempenho do modelo *random forest* utilizando o valor ótimo de *mtry* (*mtry* = 6) no conjunto de treinamento. Nela, existem três resultados essenciais a serem analisados: (i) o número selecionado de árvores (*ntree* = 500) é suficiente para garantir um bom desempenho do classificador, uma vez que se estabilizou com um pouco menos de 100 árvores de decisão; (ii) a média de taxa de erro *out-of-bag* (OOB) alcançou o resultado em 0,2, sugerindo que menos de 20% dos pontos de dados podem ter sido classificados incorretamente; (iii) a maior taxa de erro foi observada para a Exploração; esse resultado era esperado, pois essa categoria não foi reamostrada.

As Tabelas 5.2 e 5.3 mostram a matriz de confusão para os dados de teste, os 25% que ficaram como o *holdout* (Tabela 4.12), antes e depois, da aplicação do algoritmo SMOTE, respectivamente.

Real	Classificado					
	Outros	Evento Desencadeador	Exploração	Integração	Resolução	Taxa de Erro
Outros	25	3	6	1	0	0,29
Evento Desencadeador	1	45	5	0	0	0,12
Exploração	5	1	308	8	7	0,06
Integração	3	4	48	30	1	0,65
Resolução	0	0	34	1	9	0,80

Tabela 5.2: Matriz confusão dados de teste sem a aplicação do SMOTE

Real	Classificado					
	Outros	Evento Desencadeador	Exploração	Integração	Resolução	Taxa de Erro
Outros	22	7	1	5	0	0,37
Evento Desencadeador	4	29	14	4	0	0,43
Exploração	5	6	307	3	8	0,07
Integração	1	8	43	34	0	0,60
Resolução	0	0	33	2	9	0,80

Tabela 5.3: Matriz confusão dados de teste com a aplicação do SMOTE

Ambas as tabelas mostram resultados semelhantes, em que a taxa de erro da fase Exploração é a mais baixa, logo depois, Outros e Evento desencadeador. As tabelas também mostram que as classificações nas fases Integração e Resolução foram, na maioria, erradas. Isso provavelmente aconteceu porque no conjunto de dados de teste (Tabela 4.12) essas fases possuíam as menores quantidades de instâncias, tornando difícil para o classificador aprender efetivamente como reconhecer as mensagens das fases. Finalmente, é importante notar que o modelo *random forest* proposto obteve a acurácia de 0,76 na classificação (95% de confiança com intervalo entre 0,72 e 0,80) e κ de 0,55 no conjunto de teste, o que é considerado uma concordância moderada, um nível acima do puro acaso (LANDIS; KOCH, 1977).

5.2 Características importantes

Este estudo também analisou as contribuições das diferentes características para o desempenho final do classificador. A Figura 5.2 apresenta as pontuações MDGs para todas as características de classificação.

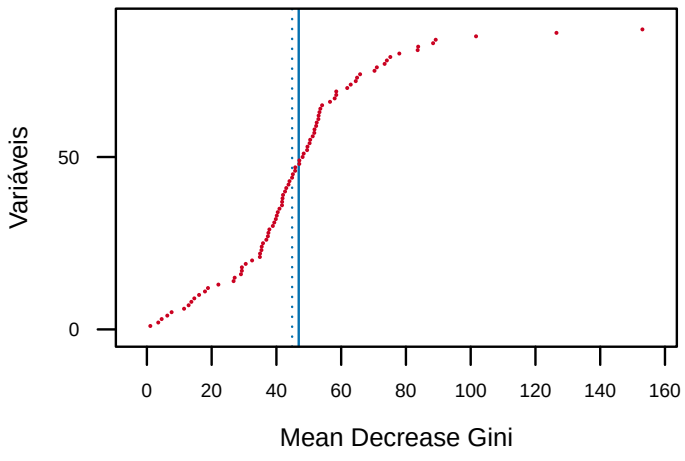

Figura 5.2: Importância da característica pela média do índice de Gini (MDG)

É possível identificar que 50% das características atingiram a pontuação MDG abaixo da mediana (44,83 - linha pontilhada azul) e 57% obtiveram a pontuação MDG menor que a média (46,89 - linha continua azul). Por outro lado, algumas características alcançaram pontuações MDG elevadas, atingindo 153,03 para a melhor característica.

Nos experimentos, das 127 características avaliadas, 55 apresentaram pontuações MDG acima do valor médio. Entretanto, este estudo vai investigar mais profundamente as 20 características com maior MDG, inclusive comparando os resultados para as diferentes bases de dados que compõe o *corpus* construído. A Tabela 5.4 mostra uma análise detalhada das vinte características mais relevantes do corpus. Além disso, as Tabelas 5.5 e 5.6 apresentam as vinte principais características das bases coletadas (BaseBio e BaseTec). Em todas elas são exibidas as pontuações MDG e os valores médios em cada categoria definida (ou seja, Outros e as fases da presença cognitiva).

Para a identificação das origens das características analisadas, utilizaram-se os prefixos *liwc*, *cm*, *message*, *wemb* e *ner*. Esses referem-se, respectivamente, ao LIWC, Coh-Metrix, Contexto da discussão, *word embedding* e entidades nomeadas. Percebe-se que a característica mais relevante foi *liwc.QMark* (número de pontos de interrogação em uma mensagem), que está diretamente relacionado à fase Evento desencadeador. Além disso, o número de palavras, o número de palavras com mais de seis letras, a média de palavras por sentenças e o comprimento médio da sentença, apresentam uma tendência similar, com valores mais altos associados às fases Exploração e Resolução, seguido pela Integração.

Das características do Coh-Metrix analisadas é possível tirar várias conclusões. Primeiro a *Givenness* (ou seja, a média da similaridade entre cada sentença e o texto que a precede) teve a maior associação com os níveis mais altos de presença cognitiva. Em segundo, a característica referente ao número de sentenças teve maior pontuação na fase Resolução. E finalmente, as características relativas ao número de palavras únicas pelo total de *tokens*, mínimo entre as frequências de palavras de conteúdo e diversidade lexical (palavras de conteúdo), atingiram os valores mais altos para Outros e Evento desencadeador.

Quanto às características extraídas do LIWC, basearam-se principalmente em valores quantitativos (número de artigos, preposições, conjunções, verbos e pronomes). Em geral, alcançaram os maiores valores em Exploração e Resolução. As exceções foram as relacionadas a quantidade de pronomes que tiveram maior associação com Evento desencadeador e Integração.

Importante notar que, a característica associada às entidades nomeadas apresentaram os valores mais altos para Integração e Resolução e não para a fase de Exploração, como esperado.

Finalmente, as características de contexto da discussão: posição da mensagem dentro da discussão e similaridade com a próxima sentença, indicaram maior associação com Evento desencadeador e Resolução, respectivamente. Esses resultados estão diretamente relacionados com o tipo das discussões (perguntas e respostas com um grande número de intervenções do professor), em que a maioria das mensagens de encadeamento foram postadas pelo professor para incentivar o engajamento dos alunos, contribuindo para o avanço nas demais fases da presença.

Nº	Variável	Descrição	MDG	Outros	Fases da Presença Cognitiva			
					Evento Desencadeador	Exploração	Integração	Resolução
1	liwc.QMark	Nº de pontos de interrogação	153,03	0,07 (0,34)	2,15 (3.90)	0,12 (0,68)	0,43 (1,97)	0,11 (0,56)
2	cm.AveSen	Comprimento médio da sentença	126,49	6,63 (4,54)	8,44 (4,17)	27,68 (16,54)	25,02 (16,80)	31,85 (14,96)
3	message.depth	Posição da mensagem dentro da discussão	101,65	2,61 (1,21)	2,53 (1,17)	1,43 (0,93)	2,46 (1,20)	1,46 (1,11)
4	liwc.6Word	Nº de palavras com mais de seis letras	89,20	3,26 (2,81)	10,17 (16,89)	37,03 (32,20)	23,70 (29,78)	58,76 (30,83)
5	liwc.Art	Nº de artigos	88,39	0,93 (1,35)	3,27 (5,15)	12,39 (11,20)	7,46 (9,01)	18,82 (10,21)
6	cm.NumWord	Nº de palavras	83,78	12,24 (11,53)	33,93 (51,68)	114,77 (97,65)	77,76 (88,08)	175.65 (91,18)
7	cm.WPerSen	Média de palavras por sentença	83,62	10,46 (13,43)	12,43 (9,81)	30,06 (19,66)	26,67 (18,65)	33,63 (19,80)
8	liwc.Prep	Nº de preposições	77,96	1,30 (1,88)	3,69 (6,55)	15,63 (14,02)	10,03 (11,76)	24,83 (14,04)
9	liwc.Conj	Nº de conjunções	75,18	0,40 (0,79)	1,88 (3,45)	6,35 (5,95)	4,84 (6,33)	10,16 (6,16)
10	liwc.words	Nº de palavras	74,11	18,89 (31,00)	30,13 (54,12)	121,97 (103,11)	80,54 (72,30)	113,99 (74,63)
11	cm.Giveness	Média da simil. entre cada sent. e o texto que a precede	73,42	0,48 (0,18)	0,55 (0,19)	0,73 (0,12)	0,74 (0,13)	0,74 (0,10)
12	liwc.Verb	Nº de verbos	71,02	1,46 (1,78)	4,69 (7,18)	16,92 (14,52)	11,64 (12,18)	25,95 (13,56)
13	cm.TTR	Nº de palavras únicas por total de *tokens*	70,29	0,93 (0,13)	0,93 (0,09)	0,75 (0,11)	0,81 (0,11)	0,75 (0,11)
14	cm.MCWord	Mín. entre as frequências de palavras de conteúdo	65,85	47,89 (11,86)	19,00 (65,09)	5,54 (5,86)	3,81 (8,98)	2,18 (5,12)
15	liwc.ppron	Nº de pronomes pessoais	64,91	1,72 (1,17)	4,16 (7,05)	2,85 (5,21)	4,58 (10,97)	3,43 (5,16)
16	liwc.pron	Nº total de pronomes	64,50	1,72 (1,17)	4,16 (7,05)	2,85 (5,21)	4,58 (10,97)	3,43 (5,16)
17	message.SNext	Similaridade com a próxima sentença	62,95	0,57 (0,19)	0,63 (0,15)	0,76 (0,17)	0,73 (0,15)	0,81 (0,13)
18	cm.LDContW	Diversidade lexical, palavras de conteúdo	61,87	0,68 (0,17)	0,56 (0,13)	0,42 (0,07)	0,45 (0,08)	0,40 (0,06)
19	ner.Cont	Nº de entidades nomeadas	58,48	1,46 (0,98)	3,44 (5,83)	3,08 (4,55)	4,28 (9,56)	3,63 (4,03)
20	cm.Sent	Nº de sentenças	58,48	1,99 (1,02)	3,96 (5,04)	5,55 (6,01)	4,17 (5,84)	6,79 (5,08)

Tabela 5.4: Vinte características mais importantes para distinguir entre as fases de presença cognitiva e seus valores nas diferentes fases da presença (*corpus*)

					Fases da Presença Cognitiva			
Nº	Variável	Descrição	MDG	Outros	Evento Desencadeador	Exploração	Integração	Resolução
1	liwc.QMark	Nº de pontos de interrogação	154,66	0,07 (0,36)	1,17 (0,80)	0,17 (0,81)	0,08 (0,37)	0,20 (0,82)
2	cm.AveSen	Comprimento médio da sentença	99,69	5,98 (3,48)	8,15 (4,17)	25,3 (15,4)	23,5 (14,0)	25,7 (10,8)
3	message.depth	Posição da mensagem dentro da discussão	98,12	2,65 (1,22)	2,58 (1,17)	1,52 (1,01)	2,56 (1,17)	2,18 (1,63)
4	liwc.6Word	Nº de palavras com mais de seis letras	86,96	3,04 (2,65)	5,17 (6,63)	40,4 (36,4)	17,6 (17,4)	39,9 (28,8)
5	cm.WPerSen	Média de palavras por sentença	69,23	8,31 (9,53)	10,5 (6,29)	27,3 (17,9)	25,8 (16,6)	27,2 (11,0)
6	liwc.Art	Nº de artigos	68,58	0,86 (1,31)	1,81 (2,19)	13,6 (12,8)	5,64 (5,85)	12,7 (10,2)
7	cm.Tokens	Nº de *tokens*	56,39	11,5 (10,98)	19,1 (20,9)	127 (111)	60,0 (51,6)	131 (95,8)
8	cm.Giveness	Média da simil. entre cada sent. e o texto que a precede	53,24	0,47 (0,18)	0,54 (0,19)	0,73 (0,13)	0,73 (0,13)	0,69 (0,11)
9	liwc.Prep	Nº de preposições	52,33	1,16 (1,78)	1,92 (2,74)	17,1 (15,8)	7,32 (6,42)	17,9 (12,4)
10	cm.NumWord	Nº de palavras	51,67	12,0 (12,89)	19,1 (20,5)	126 (111)	59,8 (52,2)	131 (95,3)
11	liwc.Conj	Nº de conjunções	49,89	0,35 (0,75)	0,97 (1,36)	6,80 (6,73)	3,56 (3,85)	7,43 (5,57)
12	cm.MCWord	Mín. entre as frequências de palavras de conteúdo	44,58	520 (123)	213 (69,2)	69,3 (72,3)	39,1 (10,3)	38,4 (9,15)
13	liwc.Verb	Nº de verbos	43,66	1,38 (1,69)	2,79 (3,44)	18,3 (16,4)	9,27 (7,89)	19,3 (14,6)
14	cm.LDVOCD	Diversidade lexical, VOCD	42,79	0,69 (0,17)	0,59 (0,12)	0,42 (0,07)	0,46 (0,08)	0,43 (0,07)
15	cm.LDContW	Diversidade lexical, palavras de conteúdo	39,62	0,65 (0,15)	0,57 (0,11)	0,40 (0,06)	0,43 (0,05)	0,43 (0,07)
16	cm.TTR	Nº de palavras únicas por total de *tokens*	38,79	0,95 (0,13)	0,95 (0,07)	0,74 (0,11)	0,84 (0,10)	0,76 (0,12)
17	liwc.Quant	Nº de quantificadores	38,67	0,56 (0,90)	0,43 (0,81)	3,13 (3,13)	1,85 (1,95)	3,95 (3,57)
18	cm.ContWord	Frequência de palavras de conteúdo	37,29	730 (186)	629 (109)	587 (57,5)	599 (65.0)	590 (46.4)
19	liwc.3Pron	Nº de pronomes em terceira pessoa do singular	36,33	0,04 (0,19)	0,13 (0,41)	1,55 (2,13)	0,67 (1,02)	1,68 (1,91)
20	cm.PronNP	Média de pronomes por frase nominal	35,81	0,02 (0,07)	0,04 (0,11)	0,02 (0,05)	0,05 (0,07)	0,01 (0,02)

Tabela 5.5: Vinte características mais importantes para distinguir entre as fases de presença cognitiva e seus valores nas diferentes fases da presença (BaseBio)

					Fases da Presença Cognitiva			
					Evento			
Nº	Variável	Descrição	MDG	Outros	Desencadeador	Exploração	Integração	Resolução
1	liwc.Art	Nº de artigos	46,49	1,82 (1,51)	13,67 (7,67)	10,11 (6,81)	10,79 (12,31)	20,77 (9,46)
2	liwc.6Word	Nº de palavras com mais de seis letras	44,13	5,76 (3,40)	45,73 (23,90)	30,68 (20,65)	34,92 (42,18)	64,78 (29,06)
3	liwc.pron	Nº total de pronomes	42,13	2,06 (1,60)	19,21 (11,20)	1,08 (2,18)	6,39 (15,61)	3,00 (4,49)
4	cm.NumWord	Nº de palavras	40,58	21,18 (14,15)	139,21 (77,90)	91,99 (58,27)	110,40 (124,78)	189,82 (85,26)
5	liwc.Verb	Nº de verbos	39,22	2,47 (2,40)	18,24 (11,29)	14,28 (9,55)	15,98 (16,73)	28,07 (12,51)
6	liwc.ppron	Nº de pronomes pessoais	39,08	2,06 (1,60)	19,21 (11,20)	1,08 (2,18)	6,39 (15,61)	3,00 (4,49)
7	wemb.similarity	Similaridade entre sentenças	37,43	0,16 (0,21)	0,39 (0,06)	0,48 (0,26)	0,39 (0,23)	0,61 (0,13)
8	ner.Cont	Nº de entidades nomeadas	36,76	2,18 (1,38)	16,48 (8,27)	1,50 (2,21)	6,11 (14,18)	3,43 (3,79)
9	cm.AveSen	Comprimento médio da sentença	36,35	14,18 (7,75)	10,55 (3,55)	32,13 (17,74)	27,73 (20,79)	33,81 (15,60)
10	liwc.QMark	Nº de pontos de interrogação	36,14	0,00 (0,00)	9,12 (8,06)	0,04 (0,26)	1,08 (3,18)	0,08 (0,44)
11	liwc.Conj	Nº de conjunções	33,52	0,94 (0,97)	8,39 (5,98)	5,50 (3,97)	7,18 (8,86)	11,04 (6,10)
12	cm.Sent	Nº de sentenças	33,01	1,76 (1,09)	14,30 (7,73)	3,61 (2,70)	5,74 (8,69)	6,85 (4,54)
13	cm.TTR	Nº de palavras únicas por total de *tokens*	32,02	0,84 (0,07)	0,56 (0,08)	0,69 (0,09)	0,70 (0,13)	0,59 (0,06)
14	liwc.Prep	Nº de preposições	31,47	2,82 (2,35)	16,27 (10,78)	12,78 (9,13)	14,99 (16,74)	27,04 (13,84)
15	message.depth	Posição da mensagem dentro da discussão	28,98	2,12 (0,99)	2,15 (1,15)	1,26 (0,71)	2,27 (1,25)	1,23 (0,76)
16	cm.SDWordSy	Desvio padrão da contagem de sílabas de palavras	24,69	1,13 (0,15)	1,26 (0,27)	1,30 (0,22)	1,24 (0,24)	1,33 (0,21)
17	liwc.PresTense	Nº de verbos no presente	22,67	1,24 (1,15)	7,67 (4,74)	5,38 (3,87)	6,80 (7,33)	10,00 (5,57)
18	liwc.Quant	Nº de quantificadores	21,32	0,65 (0,70)	1,21 (1,08)	2,29 (2,07)	2,54 (2,27)	4,67 (3,22)
19	liwc.IndPron	Nº de pronomes indefinidos	21,12	0,65 (0,61)	1,58 (1,23)	1,69 (1,71)	2,15 (2,38)	3,59 (2,75)
20	message.SNext	Similaridade com a próxima sentença	20,48	0,60 (0,18)	0,76 (0,13)	0,79 (0,14)	0,76 (0,10)	0,82 (0,12)

Tabela 5.6: Vinte características mais importantes para distinguir entre as fases de presença cognitiva e seus valores nas diferentes fases da presença (BaseTec)

5.3 Discussões

Esta seção apresenta as discussões dos resultados da aplicação do método proposto para automatizar a análise da presença cognitiva. Além disso, mostra uma análise comparativa sobre as características da presença cognitiva em dois contextos diferentes.

5.3.1 Análise dos resultados no *corpus*

Os resultados obtidos com a classificação automática da presença cognitiva no conjunto de dados de teste mostraram que as características baseadas no LIWC e no Coh-Metrix, somadas às características referentes ao contexto das discussões e entidades nomeadas, são eficazes para classificar mensagens de fóruns em português. O valor de κ obtido foi 0,55, o que representa uma taxa de concordância de nível moderado (LANDIS; KOCH, 1977).

A avaliação da classificação mostrou ainda que a otimização do parâmetro *mtry* (isto é, o número de características utilizados em cada árvore da floresta) melhorou o resultado final em 0,05, no valor de κ e 0,04, na acurácia da classificação (Tabela 5.1). Embora não se tenha encontrado nenhum outro trabalho relacionado em português que tenha realizado uma análise semelhante para comparação, é importante mencionar que a abordagem apresentada nesta dissertação alcançou resultados de acurácia melhores que os classificadores de presença cognitiva desenvolvidos para o inglês (KOVANOVIĆ et al., 2014; WATERS et al., 2015; KOVANOVIĆ et al., 2016).

Este estudo realizou uma análise detalhada das características utilizadas. Primeiro, o modelo foi treinado com 127 características e não utilizou vetores gerados pela técnica de *bag-of-words* como características. Assim, as chances de *overfitting* dos dados de treinamento diminuem significativamente. Em segundo lugar, os resultados indicaram que um pequeno subconjunto de características possuía indicadores altamente preditivos das diferentes fases da presença cognitiva (Figura 5.2).

É importante destacar que os indicadores de classificação mais relevantes (Tabela 5.4) foram alinhados com a teoria da presença cognitiva (GARRISON; ANDERSON; ARCHER, 2001). Os níveis mais altos de presença cognitiva foram relacionados a mensagens que: (i) são longas, com mais palavras e sentenças; (ii) são complexas, com palavras complexas (palavras maiores que 6 letras) e sentenças mais longas; (iii) são compostas por mais verbos, preposições, artigos e conjunções; (iv) possuem menor diversidade léxica; (v) têm maior quantidade de dados sobre as informações (*givenness*); (vi) possuem maior similaridade entre as sentenças; (vii) incluem mais entidades nomeadas; (viii) utilizam menos pontos de interrogação.

As conclusões tiradas acima são consistentes com as reportadas em estudos anteriores. Por exemplo, 50% das 20 principais características encontradas no presente estudo condizem com as encontradas por (KOVANOVIĆ et al., 2016). Pesquisas futuras são necessárias para entender melhor as razões das diferenças nas contribuições das características em diferentes

estudos.

Destaca-se ainda a categoria Outros. Nos resultados, percebeu-se que as mensagens dessa categoria são caracterizadas pela alta diversidade léxica em relação a outras mensagens (como visto nas características de diversidade léxica e TTR) e tendem a ser mais informais (com menos palavras e sentenças). Apresentaram um número menor de entidades nomeadas e *givenness* (consequência relacionada ao fato serem mensagens que geralmente não estão relacionadas ao tópico do curso). Essas mensagens também tendem a adotar uma linguagem mais simples, conforme indicado pela pontuação baixa em palavras com mais de seis letras, por exemplo. Além disso, estão propensas a ocorrer mais frequentemente perto do final da discussão, como apontado pelo seu alto valor para a característica relacionada a posição da mensagem dentro da discussão, o que já era esperado, pois muitas discussões normalmente terminam com os alunos agradecendo uns aos outros por suas contribuições.

5.3.2 Análise comparativa dos resultados das bases de dados de domínios diferentes

Os resultados das Tabelas 5.5 e 5.6, referentes as bases de dados em Biologia e Tecnologia, respectivamente, mostraram que entre as duas coincidiram 55% das características, em especial as extraídas das ferramentas LIWC e Coh-Metrix. Em comparação com as características do *corpus* completo (Tabela 5.4), em BaseBio houve uma coincidência de 70% e, em BaseTec, 75%, destacando também as obtidas nas ferramentas supracitadas.

Quanto as diferenças, um aspecto importante está relacionado à fase Evento desencadeador. Na BaseTec, as mensagens dessa fase são: (i) longas, com mais palavras e sentenças; (ii) complexas, com palavras complexas (palavras maiores que 6 letras); (iii) compostas por mais verbos, artigos, preposições, conjunções; (iv) apresentadas com menor diversidade léxica. Em BaseBio, elas apresentaram características contrárias: (i) são curtas, com menos palavras e sentenças; (ii) são simples, com palavras simples (palavras maiores que seis letras) e sentenças mais curtas; (iii) possuem menos verbos, preposições, artigos e conjunções; (iv) têm maior diversidade léxica.

Conforme relatado anteriormente, as mensagens dessa fase são, em sua maioria, postadas pelo professor para motivar a participação dos alunos nos fóruns. Assim, a análise descrita acima, indica uma atuação mais constante por parte do professor (presença de ensino) no curso de Tecnologia, quando comparada a atuação do professor no curso de Biologia. Esse fato reflete diretamente no avanço nas demais fases da presença cognitiva, conforme apontado por GARRISON; ANDERSON; ARCHER (2001), quando cita a importância da presença de ensino para o desenvolvimento da presença cognitiva e verificado nos altos valores obtidos nas fases posteriores da presença cognitiva, principalmente na Resolução.

A análise aponta ainda para mensagens mais estruturadas, com perguntas bem definidas e contextualizadas (como pode ser visto nas características pontos de interrogação e entidades

nomeadas) em BaseTec. Em (GARRISON; ARBAUGH, 2007), os autores afirmam que tarefas formuladas adequadamente no início das discussões influenciam nas respostas dos alunos, facilitando o avanço nas fases da presença cognitiva (Tabela 5.6).

A participação ativa do professor nas discussões é um fator importante para o desenvolvimento da presença social em ambientes online de aprendizagem, conforme apontado por (GARRISON; ARBAUGH, 2007) e demonstrado em estudos anteriores (AN; SHIN; LIM, 2009; CHO; KIM, 2013; HEW; CHEUNG; NG, 2010). Na BaseTec é possível perceber uma expressiva concordância entre as discussões e coesão de grupo (indicadores da presença de social) nas fases posteriores da presença cognitiva (como visto nas características de similaridade entre sentenças e de pronomes). De acordo com GARRISON; ANDERSON; ARCHER (2001), a presença de ensino integrada à presença social fornecem suporte para a presença cognitiva, proporcionam um ambiente onde a ela pode ser desenvolvida, como visto nos resultados dos níveis mais altos.

Outro aspecto relevante sobre as diferenças entre os contextos refere-se aos níveis mais altos da presença cognitiva. Em especial, identificou-se um contraste entre as características das mensagens da fase Exploração, em que predomina a crescente divergência de ideias e opiniões, questionamentos, *brainstorming*, e trocas de informações. Assim, na BaseBio foram encontradas mensagens: (i) longas, com mais palavras e sentenças; (ii) complexas, com palavras complexas e sentenças longas; (iii) compostas por mais verbos, artigos, preposições, conjunções. Em BaseTec, a fase reuniu mensagens: (i) curtas, com menos palavras e sentenças; (ii) simples, com palavras simples e sentenças curtas; (iii) compostas por menos verbos, artigos, preposições, conjunções. Nesse sentido, é possível fazer uma associação entre os dados supra citados e o gênero dos estudantes, conforme apontado por (GARRISON; ARBAUGH, 2007). No curso de Biologia, a maioria dos alunos são do sexo feminino (70%), diferente do curso de Tecnologia, em que predominaram alunos do sexo masculino (78%). Assim, uma provável explicação para o contraste de características entre as mensagens da fase Exploratória das bases de dados seja a forma como os estudantes se comunicam no ambiente online, que apresenta diferenças dependendo do sexo. Segundo (ROVAI, 2001), as mulheres são mais interativas do que os homens, pois buscam construir um senso de comunidade online, promovendo mais participações e interações nos fóruns, como verificado em BaseBio.

O resultado na fase Exploração acima mostrou-se, entretanto, contrário ao encontrado no estudo de (BARRETT; LALLY, 1999). Nele, comparado às mulheres, os homens postaram mais mensagens e com uma quantidade maior de palavras. Essa diferença aponta para outra provável explicação, relacionada à área do curso. Em geral, no curso da área de exatas, os estudantes do escrevem menos que nos cursos das áreas de saúde e humanas e sociais. Nessa perspectiva, percebeu-se ainda uma diferença expressiva entre as fases Resolução dos dois contextos. Em Tecnologia, a fase apresenta-se mais representativa que em Biologia. Características como *liwc.Art*, *liwc.6Word* e *cm.NumWord*, as primeiras destacadas na Tabela 5.6, possuem os maiores valores nesta fase. Além disso, como visto na Tabela 4.6, a quantidade de mensagens classificadas como Resolução é maior, comparada a BaseBio (Tabela 4.5). A justificativa pode estar também

correlacionada a forma diferente de comunicação dos estudantes do sexo masculino e feminino, como apontado por BARRETT; LALLY (1999) ao verificar que os homens postaram mensagens mais focadas na discussão proposta. Contudo, sobre este ponto, seria necessário realizar um estudo mais amplo, analisando, além desses aspectos, outros que estejam relacionados, por exemplo, tipo das discussões, influência das demais presenças, para conclusões mais precisas sobre este tópico.

É importante destacar ainda as características diferentes entre as duas bases. A BaseBio contou com a presença de mais características relacionadas a diversidade léxica, em especial, *cm.LDVOCD* e *cm.LDContW*, presentes somente nessa base. A BaseTec apresentou uma variedade maior de características. Ou seja, a base apresentou características de todos os recursos utilizados neste estudo, diferente da BaseBio. Em destaque as características *wemb.similarity* e *ner.Cont*, provenientes dos recursos *word embedding* e entidade nomeada, respectivamente, existentes apenas nessa base.

Por fim, em comparação com o trabalho de (KOVANOVIĆ et al., 2016), houve uma semelhança entre as características de 45% e 50%, em BaseBio e BaseTec, respectivamente. (GARRISON; ARBAUGH, 2007) alertam para a inclusão de outras variáveis (por exemplo, curso e assunto) na análise das relações entre os elementos do modelo CoI. As mensagens que compõem a base de dados do estudo de (KOVANOVIĆ et al., 2016) foram extraídas de um curso da área de Tecnologia, assim como BaseTec. Desta forma, estando em um mesmo contexto, é possível perceber maior semelhança entre as características (como, *wemb.similarity*, *ner.Cont* e *message.SNext*) e o desenvolvimento da presença cognitiva.

6

Considerações Finais

Esta dissertação de mestrado teve como objetivo principal propor um método para automatizar a análise da presença cognitiva em mensagens, provenientes de fóruns de discussões, escritas em português como habitualmente falado no Brasil, por meio de técnicas de Mineração de Texto. São três as suas contribuições principais. Primeiro, um novo classificador foi desenvolvido para codificar as mensagens dos alunos, em relação as fases da presença cognitiva, escritas em português. A abordagem desenvolvida obteve 76% de acurácia e κ de 0,55, o que é considerado concordância moderada, acima do nível de puro acaso (LANDIS; KOCH, 1977). Esse resultado mostra o potencial de fornecer um sistema automatizado para codificação da presença cognitiva em português.

A segunda contribuição desta dissertação é a análise detalhada da relevância das características propostas, baseadas principalmente no Coh-Metrix e LIWC. Em tal contexto, os experimentos realizados mostraram que mensagens longas e complexas, compostas por mais verbos, preposições, artigos e conjunções, com maior disponibilidade de informações, de entidades nomeadas e similaridade entre as sentenças, estavam relacionadas a níveis mais altos de presença cognitiva. A maior diversidade léxica e um maior número de pontos de interrogação foram associados a níveis mais baixos de presença cognitiva. Tais conclusões confirmam os resultados do trabalho realizado por KOVANOVIĆ et al. (2016).

O terceiro resultado obtido é o estudo sobre as principais diferenças relacionadas à presença cognitiva em dois contextos diferentes: Biologia (BaseBio) e Tecnologia (BaseTec). Os resultados no contexto do ensino tecnológico (BaseTec) apresentaram fortes indícios de presença de ensino, através de um professor mais atuante, e mensagens mais estruturadas, com perguntas bem definidas e contextualizadas, que refletiram no desenvolvimento da presença cognitiva e no alcance de níveis mais altos (em especial, fase Resolução). Já no caso estudado do ensino de Biologia (BaseBio), as características evidenciadas na fase Exploração trouxeram questões relacionadas à influência das diferentes formas de comunicação entre os gêneros masculino e feminino, e da área do curso em análise, na presença cognitiva.

A análise das diferenças entre as características das bases mostrou ainda que, a identificação da presença cognitiva no contexto Tecnologia envolveu características de todos os

recursos aplicados. O contexto Biologia utilizou menos recursos, entretanto, houve destaque de características relacionadas a diversidade léxica.

As conclusões apresentadas neste trabalho confirmam e complementam os resultados de estudos anteriores (por exemplo, KOVANOVIĆ et al. (2016)), contribuindo de forma relevante para o processo de detecção e análise da presença cognitiva dos alunos em fóruns de discussões.

Deste modo, através do modelo proposto neste trabalho para automatizar a codificação da presença cognitiva para a língua portuguesa, espera-se tornar o processo de análise mais fácil, contribuindo para o gerenciamento do conhecimento e das habilidades cognitivas e críticas dos alunos em fóruns de discussões online. Além disso, os resultados aqui mostrados o potencial, por exemplo, em desenvolver um módulo incorporado aos AVAs que possibilite um acompanhamento continuado dos alunos durante o curso online e a aplicação de estratégias que afetem os resultados de aprendizagem dos mesmos.

6.1 Artigos submetidos/aceitos

Em termos de divulgação do trabalho, artigos foram submetidos para conferências e periódicos qualificados na área de Ciência da Computação, conforme listado a seguir:

- *Automated Analysis of Cognitive Presence in Online Discussions Written in Portuguese. Thirteenth European Conference on Technology Enhanced Learning (EC-TEL 2018), Leeds (UK), 2018.* (MENDONCA NETO et al., 2018).
- *Towards Automated Cognitive Presence: A case in online discussions written in Portuguese* (submetido).

6.2 Limitações da pesquisa

Dentre as limitações da abordagem apresentada aqui, destacam-se duas relacionadas ao conjunto de dados. Primeiro, os dados coletados eram de dois domínios de estudo (Biologia e Tecnologia) com discussões desenhadas com um propósito pedagógico particular (ou seja, discussão do tipo perguntas e respostas) de cursos em uma universidade de língua portuguesa. Assim, o estudo pode não ser inteiramente representativo das diferentes interações que podem levar a diferentes mensagens de presença cognitiva. E em segundo lugar, o tamanho do conjunto de dados utilizados e as categorias desequilibradas, embora coerentes com os encontrados na literatura, podem afetar o desempenho do classificador.

Outros pontos a serem destacados referem-se a cobertura do dicionário léxico usado e a análise puramente estatística. Primeiro, embora o LIWC seja abrangente e com bons resultados, a utilização de um dicionário com uma quantidade maior de palavras, ou uma extensão deste usando outras técnicas de mineração de texto ou combinando outros dicionários, poderia melhorar os resultados para o modelo proposto. E segundo, mesmo consistente e coerente com as teorias

do modelo CoI, a análise realizada neste trabalho não aborda aspectos linguísticos (semânticos, por exemplo) que poderia complementada e ampliar as discussões sobre o desenvolvimento cognitivo dos alunos.

6.3 Trabalhos futuros

Como trabalhos futuros pretende-se:

- Testar a generalização do classificador em outros contextos educacionais (por exemplo, *blended learning* comparado com totalmente *online* e MOOC; ensino de graduação vs. pós-graduação);
- Aplicar a abordagem utilizando bases de dados de diferentes domínios;
- Incluir outros recursos e técnicas disponíveis para a língua portuguesa (por exemplo, outros dicionários) e outros tipos de análise (sintática, semântica, por exemplo);
- Verificar a eficácia das características propostas nesta pesquisa para outras línguas (como por exemplo, o espanhol).

Referências

ABAWAJY, J. Analysis of asynchronous online discussion forums for collaborative learning. **International Journal of Education and Learning**, [S.l.], v.1, n.2, p.11–21, 2012.

ABED. Censo EAD.BR 2017 - Relatório analítico da aprendizagem a distância no Brasil. **IBPEX**, [S.l.], 2018.

AKYOL, Z. et al. A response to the review of the community of inquiry framework. **International Journal of E-Learning & Distance Education**, [S.l.], v.23, n.2, p.123–136, 2009.

AMARAL, D. et al. Comparative analysis of Portuguese named entities recognition tools. In: NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2014). **Proceedings...** [S.l.: s.n.], 2014. p.2554–2558.

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. et al. Assessing teaching presence in a computer conferencing context. **Journal of Asynchronous Learning Networks**, [S.l.], v.5, n.2, 2001.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, [S.l.], v.5, n.2, 2006.

ARAUJO, E. M. d. **Avaliação do pensamento crítico e da presença cognitiva em fórum de discussão online**. 2014. Tese (Doutorado em Engenharia de Produção) — Universidade de São Paulo.

ARAUJO, E. M.; OLIVEIRA NETO, J. D. de. Avaliação do pensamento crítico e da presença cognitiva em fórum de discussão online utilizando a análise estatística textual. In: INTERNATIONAL CONFERENCE ON ENGINEERING AND COMPUTER EDUCATION. **Proceedings...** [S.l.: s.n.], 2013. v.8, p.113–117.

BALAGE FILHO, P. P.; PARDO, T. A. S.; ALUÍSIO, S. M. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 9. **Proceedings...** [S.l.: s.n.], 2013.

BARONI, M.; DINU, G.; KRUSZEWSKI, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (VOLUME 1: LONG PAPERS), 52. **Proceedings...** [S.l.: s.n.], 2014. v.1, p.238–247.

BARRETT, E.; LALLY, V. Gender differences in an on-line learning environment. **Journal of Computer Assisted Learning**, [S.l.], v.15, n.1, p.48–60, 1999.

BAUER, M. W. Content Analysis. An Introduction to its Methodology–By Klaus Krippendorff From Words to Numbers. Narrative, Data and Social Science–By Roberto Franzosi. **The British Journal of Sociology**, [S.l.], v.58, n.2, p.329–331, 2007.

BECKMANN, M. **Algoritmos genéticos como estratégia de pré-processamento em conjuntos de dados desbalanceados.** 2010. Dissertação (Mestrado em Engenharia Civil) — Universidade Federal do Rio de Janeiro.

BENGIO, Y. et al. A neural probabilistic language model. **Journal of machine learning research**, [S.l.], v.3, n.Feb, p.1137–1155, 2003.

BERTAGLIA, T. F. C.; NUNES, M. d. G. V. Exploring Word Embeddings for Unsupervised Textual User-Generated Content Normalization. In: WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings...** [S.l.: s.n.], 2016. p.112–120.

BREIMAN, L. Random forests. **Machine learning**, [S.l.], v.45, n.1, p.5–32, 2001.

BROWN, P. F. et al. Class-based n-gram models of natural language. **Computational linguistics**, [S.l.], v.18, n.4, p.467–479, 1992.

BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. **Workshop Languages for Data Mining and Machine Learning - (ECML/PKDD)**, [S.l.], 2013.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, [S.l.], p.1–29, 2009.

CARLETTA, J. Assessing agreement on classification tasks: the kappa statistic. **Computational linguistics**, [S.l.], v.22, n.2, p.249–254, 1996.

CHAWLA, N. V. et al. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, [S.l.], v.16, p.321–357, 2002.

CHEN, Y. Automated Content Analysis of Students' Cognitive Presence In Asynchronous Online Discussion. In: EDMEDIA: WORLD CONFERENCE ON EDUCATIONAL MEDIA AND TECHNOLOGY. **Anais...** [S.l.: s.n.], 2015. p.38–43.

CHO, M.-H.; KIM, B. J. Students' self-regulation for interaction with others in online learning environments. **The Internet and Higher Education**, [S.l.], v.17, p.69–75, 2013.

CHOI, J. D.; TETREAULT, J.; STENT, A. It depends: dependency parser comparison using a web-based evaluation tool. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS AND THE 7TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (VOLUME 1: LONG PAPERS), 53. **Proceedings...** [S.l.: s.n.], 2015. v.1, p.387–396.

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

CHUN, W. **Core python programming**. [S.l.]: Prentice Hall, 2006.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, [S.l.], v.20, n.1, p.37–46, 1960.

COLLOBERT, R. et al. Natural language processing (almost) from scratch. **Journal of Machine Learning Research**, [S.l.], v.12, n.Aug, p.2493–2537, 2011.

COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: MACHINE LEARNING, 25. **Proceedings...** [S.l.: s.n.], 2008. p.160–167.

CONDE, M. Á. et al. An evolving Learning Management System for new educational environments using 2.0 tools. **Interactive Learning Environments**, [S.l.], v.22, n.2, p.188–204, 2012.

CORICH, S.; HUNT, K.; HUNT, L. Computerised content analysis for measuring critical thinking within discussion forums. **Journal of e-learning and knowledge society**, [S.l.], v.2, n.1, p.1–14, 2006.

COUTEIRO, T. A. F. **Métodos estatísticos relacionais para previsão de resultados médicos.** 2010. Dissertação (Mestrado em Engenharia Informática e Computação) — Faculdade de Engenharia da Universidade do Porto.

CROSSLEY, S. A. Advancing research in second language writing through computational tools and machine learning techniques: a research agenda. **Language Teaching**, [S.l.], v.46, n.2, p.256–271, 2013.

CRUZ, C. S. A. M. **A educação para o desenvolvimento sustentável na formação de professores**: a web 2.0 e as interações numa comunidade de prática online. 2013Tese (doutorado). Universidade de Aveiro.

CUNHA, A. et al. Classificação Automática de Discurso Descritivo Escrito de Adultos Sadios: referência para a avaliação da linguagem de lesados cerebrais. **Encontro Nacional de Inteligência Artificial e Computacional, ENIAC**, [S.l.], v.2013, 2013.

CUNHA, A. L. V. d. **Coh-Metrix-Dementia**: análise automática de distúrbios de linguagem nas demências utilizando processamento de línguas naturais. 2015. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo.

DAHLBERG, I. Teoria do conceito. **Ciência da informação**, [S.l.], v.7, n.2, 1978.

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.

EFRON, B. Bootstrap methods: another look at the jackknife. **The Annals of Statistics**, [S.l.], v.7, n.1, p.1–26, 1979.

EGGINS, S.; SLADE, D. Analysing casual conversation. **London: Continuum**, [S.l.], 1997.

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

FERNÁNDEZ-DELGADO, M. et al. Do we need hundreds of classifiers to solve real world classification problems? **The Journal of Machine Learning Research**, [S.l.], v.15, n.1, p.3133–3181, 2014.

FINATTO, M. J. B. Complexidade textual em artigos científicos: contribuições para o estudo do texto científico em português. **Organon. Porto Alegre**, [S.l.], v.25, n.50, p.67–100, 2011.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics New York, NY, USA:, 2001. v.1, n.10.

FU, Y. Combination of random forests and neural networks in social lending. **Journal of Financial Risk Management**, [S.l.], v.6, n.04, p.418, 2017.

GAMALLO, P.; GARCIA, M. LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. **Linguamática**, [S.l.], v.9, n.1, p.19–28, 2017.

GARRISON, D. R.; ANDERSON, T. E-learning in the 21st century: a framework for research and practice. **New York, Routledge**, [S.l.], 2003.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. Critical inquiry in a text-based environment: computer conferencing in higher education. **The internet and higher education**, [S.l.], v.2, n.2-3, p.87–105, 2000.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. Critical thinking, cognitive presence, and computer conferencing in distance education. **American Journal of Distance Education**, [S.l.], v.15, n.1, p.7–23, 2001.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. The first decade of the community of inquiry framework: a retrospective. **The internet and higher education**, [S.l.], v.13, n.1-2, p.5–9, 2010.

GARRISON, D. R.; ARBAUGH, J. B. Researching the community of inquiry framework: review, issues, and future directions. **The Internet and Higher Education**, [S.l.], v.10, n.3, p.157–172, 2007.

GARRISON, D. R.; CLEVELAND-INNES, M.; FUNG, T. S. Exploring causal relationships among teaching, cognitive and social presence: student perceptions of the community of inquiry framework. **The internet and higher education**, [S.l.], v.13, n.1-2, p.31–36, 2010.

GAŠEVIĆ, D.; KOVANOVIĆ, V.; JOKSIMOVIĆ, S. Piecing the learning analytics puzzle: a consolidated model of a field of research and practice. **Learning: Research and Practice**, [S.l.], v.3, n.1, p.63–78, 2017.

GOMES, C. M.; PESSOA, T. A presença pedagógica num ambiente online criado na rede social Facebook. **Educação, Formação & Tecnologias**, [S.l.], v.5, n.2, p.60–70, 2012.

GOMES, C. M.; PESSOA, T. Um Ambiente Online de Supervisão Pedagógica Criado na Rede Social Facebook–a Presença Social e a Presença Cognitiva. In: II CONGRESSO INTERNACIONAL TIC E EDUCAÇÃO. LISBOA: INSTITUTO DE EDUCAÇÃO DA UNIVERSIDADE DE LISBOA. **Anais...** [S.l.: s.n.], 2012. p.2532–2550.

GOODFELLOW, I. et al. **Deep learning**. [S.l.]: MIT press Cambridge, 2016. v.1.

GRAESSER, A. C. et al. Coh-Metrix: analysis of text on cohesion and language. **Behavior research methods, instruments, & computers**, [S.l.], v.36, n.2, p.193–202, 2004.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-Metrix: providing multilevel analyses of text characteristics. **Educational researcher**, [S.l.], v.40, n.5, p.223–234, 2011.

GRISHMAN, R.; SUNDHEIM, B. Message understanding conference-6: a brief history. In: COLING 1996 VOLUME 1: THE 16TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. **Anais...** [S.l.: s.n.], 1996. v.1.

GU, Q. et al. Data mining on imbalanced data sets. In: ADVANCED COMPUTER THEORY AND ENGINEERING, 2008. ICACTE'08. INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2008. p.1020–1024.

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

HACKELING, G. Mastering machine learning with scikit-learn. **Packt Publishing Ltd**, [S.l.], 2014.

HALLIDAY, M. A. K. An introduction to functional grammar. **London: Edward Arnold**, [S.l.], 1994.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on knowledge and data engineering**, [S.l.], v.21, n.9, p.1263–1284, 2009.

HEW, K. F.; CHEUNG, W. S. Attracting student participation in asynchronous online discussions: a case study of peer facilitation. **Computers & Education**, [S.l.], v.51, n.3, p.1111–1124, 2008.

HEW, K. F.; CHEUNG, W. S.; NG, C. S. L. Student contribution in asynchronous online discussion: a review of the research and empirical exploration. **Instructional science**, [S.l.], v.38, n.6, p.571–606, 2010.

HINTON, G. E. et al. **Distributed representations, Parallel distributed processing**: explorations in the microstructure of cognition. [S.l.]: MIT Press, Cambridge, MA, 1986.

HOTHO, A.; NÜRNBERGER, A.; PAASS, G. A brief survey of text mining. In: LDV FORUM. **Anais...** [S.l.: s.n.], 2005. v.20, n.1, p.19–62.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI. **Anais...** [S.l.: s.n.], 1995. v.14, n.2, p.1137–1145.

KOVANOVIĆ, V. et al. Automated Cognitive Presence Detection in Online Discussion Transcripts. In: INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS AND KNOWLEDGE, 4. **Proceedings...** [S.l.: s.n.], 2014.

KOVANOVIĆ, V. et al. Towards automated content analysis of discussion transcripts: a cognitive presence case. In: OF THE SIXTH INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS & KNOWLEDGE. **Proceedings...** [S.l.: s.n.], 2016. p.15–24.

KOVANOVIĆ, V. et al. Content Analytics: the definition, scope, and an overview of published research. **Handbook of Learning Analytics and Educational Data Mining**, [S.l.], p.77–92, 2017.

KOVANOVIĆ, V.; GAŠEVIĆ, D.; HATALA, M. Learning analytics for communities of inquiry. **Journal of Learning Analytics**, [S.l.], v.1, n.3, p.195–198, 2014.

KUHN, M. et al. Caret: classification and regression training. 2016. **R package version**, [S.l.], v.4, 2017.

KUSNER, M. et al. From word embeddings to document distances. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING. **Proceedings...** [S.l.: s.n.], 2015. p.957–966.

LAGARTO, J.; MARQUES, H. Conferências Online – um espaço de aprendizagem significativa. In: ATAS DA X CONFERÊNCIA INTERNACIONAL DE TIC NA EDUCAÇÃO - CHALLENGES 2017. **Anais...** [S.l.: s.n.], 2017. p.209–226.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, [S.l.], p.159–174, 1977.

LANTZ, B. **Machine learning with R**. [S.l.]: Packt Publishing Ltd, 2013.

LEE, H. D. **Seleção e construção de features relevantes para o aprendizado de máquina.** 2000. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo.

LEVY, O.; GOLDBERG, Y. Dependency-based word embeddings. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (VOLUME 2: SHORT PAPERS), 52. **Proceedings...** [S.l.: s.n.], 2014. v.2, p.302–308.

LEVY, O.; GOLDBERG, Y.; DAGAN, I. Improving distributional similarity with lessons learned from word embeddings. **Transactions of the Association for Computational Linguistics**, [S.l.], v.3, p.211–225, 2015.

LIAW, A.; WIENER, M. et al. Classification and regression by randomForest. **R news**, [S.l.], v.2, n.3, p.18–22, 2002.

LISBÔA, E. S.; COUTINHO, C. P. Instrumentos para avaliação das aprendizagens em fóruns de discussão online: um contributo teórico e prático/tools for evaluation of learning in online discussion forums: a practical and theoretical contribution. **Revista EducaOnline**, [S.l.], v.6, n.3, p.86–104, 2012.

LJUNG, L. Black-box models from input-output measurements. In: INSTRUMENTATION AND MEASUREMENT TECHNOLOGY CONFERENCE, 2001. IMTC 2001. PROCEEDINGS OF THE 18TH IEEE. **Anais...** [S.l.: s.n.], 2001. v.1, p.138–146.

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.

MACÁRIO, M. J.; SÁ, C. M.; MOREIRA, A. Trabalho colaborativo em fóruns de discussão online: lugares de encontro na formação inicial de professores. **Investigar em Educação**, [S.l.], v.2, n.2, 2014.

MACHADO, A. A. et al. Personalitatem Lexicon: um léxico em português brasileiro para mineração de traços de personalidade em textos. In: CONFERÊNCIA LATINO-AMERICANA DE OBJETOS E TECNOLOGIAS DE APRENDIZAGEM, X; CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, IV; SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, XXVI. **Anais...** [S.l.: s.n.], 2015.

MARNEFFE, M.-C. et al. Generating typed dependency parses from phrase structure parses. In: LREC. **Proceedings...** [S.l.: s.n.], 2006. v.6, n.2006, p.449–454.

MCGILL, T. J.; KLOBAS, J. E. A task–technology fit view of learning management system impact. **Computers & Education**, [S.l.], v.52, n.2, p.496–508, 2009.

MCKLIN, T. E. **Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network**. 2004PhD thesis.United States:Georgia State University, College of Education, Atlanta, GA.

MENDES, P. N. et al. DBpedia spotlight: shedding light on the web of documents. In: OF THE 7TH INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS. **Proceedings...** [S.l.: s.n.], 2011. p.1–8.

MENDONCA NETO, V. et al. Automated Analysis of Cognitive Presence in Online Discussions Written in Portuguese. In: EUROPEAN CONFERENCE ON TECHNOLOGY ENHANCED LEARNING. **Anais...** [S.l.: s.n.], 2018. p.245–261.

MERTZ, D. **Text Processing with Python**. [S.l.]: Addison-Wesley Longman Publishing Co., Inc., 2003.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS. **Proceedings...** [S.l.: s.n.], 2013.

MILLMAN, K. J.; AIVAZIS, M. Python for scientists and engineers. **Computing in Science & Engineering**, [S.l.], v.13, n.2, p.9–12, 2011.

MORAN, J. M. Modelos e avaliação do ensino superior a distância no Brasil The models and the evaluation of higher distance education in Brazil. **ETD: Educação Temática Digital**, [S.l.], v.10, n.2, p.54–70, 2009.

MU, J. et al. The ACODEA framework: developing segmentation and classification schemes for fully automatic analysis of online discussions. **International journal of computer-supported collaborative learning**, [S.l.], v.7, n.2, p.285–305, 2012.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, [S.l.], v.30, n.1, p.3–26, 2007.

NAILI, M.; CHAIBI, A. H.; GHEZALA, H. H. B. Comparative study of word embedding methods in topic segmentation. **Procedia Computer Science**, [S.l.], v.112, p.340–349, 2017.

NILSSON, N. J. **Introduction to machine learning**: an early draft of a proposed textbook. [S.l.]: USA; Stanford University, 1996.

PAIM, A.; CAMATI, R.; ENEMBRECK, F. Inferência de personalidade a partir de textos em português utilizando léxico linguıstico e aprendizagem de máquina. **Anais do XIII Encontro Nacional de Inteligência Artificial e Computacional**, [S.l.], p.481–492, 2016.

PARK, C. L. Replicating the use of a cognitive presence measurement tool. **Journal of Interactive Online Learning**, [S.l.], v.8, n.2, p.140–155, 2009.

PAUL, R. Critical thinking: what every person needs to survive in a rapidly changing world. **Rohnert Park: C.A.: Centre for Critical Thinking and Moral Critique**, [S.l.], 1993.

PEDREGOSA, F. et al. Scikit-learn: machine learning in python. **Journal of machine learning research**, [S.l.], v.12, n.Oct, p.2825–2830, 2011.

PENG, Y. et al. Cross-validation and ensemble analyses on multiple-criteria linear programming classification for credit cardholder behavior. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE. **Anais...** [S.l.: s.n.], 2004. p.931–939.

PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. Linguistic inquiry and word count: liwc 2001. **Mahway: Lawrence Erlbaum Associates**, [S.l.], v.71, n.2001, p.2001, 2001.

PRATI, R. C. et al. Uma experiência no balanceamento artificial de conjuntos de dados para aprendizado com classes desbalanceadas utilizando análise ROC. In: WORKSHOP ON ADVANCES & TRENDS IN AI FOR PROBLEM SOLVING. **Proceedings...** [S.l.: s.n.], 2003. v.1, p.28–33.

RAMASUBRAMANIAN, C.; RAMYA, R. Effective pre-processing activities in text mining using improved porter's stemming algorithm. **International Journal of Advanced Research in Computer and Communication Engineering**, [S.l.], v.2, n.12, p.4536–4538, 2013.

RÊGO, A. S. d. C. **Aprendizado automático de relações semânticas entre tags de folksonomias.** 2016. Tese (Doutorado em Ciência da Computação) — Universidade Federal de Campina Grande.

REZENDE, S. O. **Sistemas inteligentes**: fundamentos e aplicações. [S.l.]: Editora Manole Ltda, 2003.

ROSÉ, C. et al. Analyzing collaborative learning processes automatically: exploiting the advances of computational linguistics in computer-supported collaborative learning. **International journal of computer-supported collaborative learning**, [S.l.], v.3, n.3, p.237–271, 2008.

ROSSI, R. G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. 2015. Tese (Doutorado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo.

ROURKE, L. et al. Assessing social presence in asynchronous text-based computer conferencing. **Journal of distance education**, [S.l.], v.14, n.2, p.51–70, 2001.

ROURKE, L. et al. Methodological issues in the content analysis of computer conference transcripts. **International journal of artificial intelligence in education (IJAIED)**, [S.l.], v.12, p.8–22, 2001.

ROVAI, A. P. Building classroom community at a distance: a case study. **Educational Technology Research and Development**, [S.l.], v.49, n.4, p.33, 2001.

ROZENFELD, C. C. d. F. Fóruns online na formação crítico-reflexiva de professores de línguas estrangeiras: uma representação do pensamento crítico em fases na/pela linguagem. **ALFA: Revista de Linguística**, [S.l.], v.58, n.1, 2014.

SÁ, C. M.; MACÁRIO, M. J. TIC e desenvolvimento de competências em trabalho colaborativo na formação em didática de línguas. **Indagatio Didactica**, [S.l.], v.6, n.1, 2014.

SCARTON, C.; GASPERIN, C.; ALUISIO, S. Revisiting the readability assessment of texts in Portuguese. In: IBERO-AMERICAN CONFERENCE ON ARTIFICIAL INTELLIGENCE. **Anais...** [S.l.: s.n.], 2010. p.306–315.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, [S.l.], v.34, n.1, p.1–47, 2002.

SHAH, F. P.; PATEL, V. A review on feature selection and feature extraction for text classification. In: WIRELESS COMMUNICATIONS, SIGNAL PROCESSING AND NETWORKING (WISPNET), INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2016. p.2264–2268.

SPILIOPOULOS, V.; VOUROS, G. A.; KARKALETSIS, V. On the discovery of subsumption relations for the alignment of ontologies. **Web Semantics: Science, Services and Agents on the World Wide Web**, [S.l.], v.8, n.1, p.69–88, 2010.

STONE, P. J.; DUNPHY, D. C.; SMITH, M. S. The general inquirer: a computer approach to content analysis. **MIT press**, [S.l.], 1966.

STRIJBOS, J.-W. Assessment of (computer-supported) collaborative learning. **IEEE transactions on learning technologies**, [S.l.], v.4, n.1, p.59–73, 2011.

STRIJBOS, J.-W. et al. Content analysis: what are they talking about? **Computers & Education**, [S.l.], v.46, n.1, p.29–48, 2006.

SWAN, K.; GARRISON, D.; RICHARDSON, J. C. A constructivist approach to online learning: the community of inquiry framework. In: **Information technology and constructivism in higher education**: progressive learning frameworks. [S.l.]: IGI Global, 2009. p.43–57.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: liwc and computerized text analysis methods. **Journal of language and social psychology**, [S.l.], v.29, n.1, p.24–54, 2010.

TOLEDO, C. M. et al. Automatic classification of written descriptions by healthy adults: an overview of the application of natural language processing and machine learning techniques to clinical discourse analysis. **Dementia & Neuropsychologia**, [S.l.], v.8, n.3, p.227–235, 2014.

TURIAN, J.; RATINOV, L.; BENGIO, Y. Word representations: a simple and general method for semi-supervised learning. In: OF THE 48TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings...** [S.l.: s.n.], 2010. p.384–394.

TURNER, V. et al. The digital universe of opportunities: rich data and the increasing value of the internet of things. **IDC Analyze the Future**, [S.l.], v.16, 2014.

VAROQUAUX, G. et al. Scikit-learn: machine learning without learning the machinery. **GetMobile: Mobile Computing and Communications**, [S.l.], v.19, n.1, p.29–33, 2015.

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.

VIEIRA, R.; LOPES, L. PROCESSAMENTO DE LINGUAGEM NATURAL E O TRATAMENTO COMPUTACIONAL DE LINGUAGENS CIENTÍFICAS. **EM CORPORA**, [S.l.], p.183, 2010.

WATERS, Z. et al. Structure matters: adoption of structured classification approach in the context of cognitive presence classification. In: ASIA INFORMATION RETRIEVAL SYMPOSIUM. **Anais...** [S.l.: s.n.], 2015. p.227–238.

WEISS, S. M.; KULIKOWSKI, C. A. **Computer systems that learn**: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. [S.l.]: Morgan Kaufmann Publishers Inc., 1991.

WEN, M.; YANG, D.; ROSÉ, C. P. Linguistic Reflections of Student Engagement in Massive Open Online Courses. In: ICWSM. **Anais...** [S.l.: s.n.], 2014.

WITTEN, I. H.; FRANK, E.; MARK, A. Hall. 2011. **Data mining: Practical machine learning tools and techniques**, [S.l.], v.3, 2011.