UNIVERSIDADE FEDERAL DE ALAGOAS-UFAL
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

JEAN BARROS TEIXEIRA

IDENTIFICAÇÃO AUTOMÁTICA DA PRESENÇA SOCIAL EM DISCUSSÕES ONLINE ESCRITAS EM PORTUGUÊS

MACEIÓ
2020

Jean Barros Teixeira

IDENTIFICAÇÃO AUTOMÁTICA DA PRESENÇA SOCIAL EM DISCUSSÕES ONLINE ESCRITAS EM PORTUGUÊS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Programa de Pós-Graduação em Informática - PPGI da Universidade Federal de Alagoas - UFAL, Instituto de Computação.

Orientador: Prof. Dr. Evandro de B. Costa
Coorientador: Prof. Dr. Rafael Ferreira L. de Mello

Maceió
2020

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

T266i Teixeira, Jean Barros.
Identificação automática da presença social em discussões online escritas em português / Jean Barros Teixeira. – 2020.
101 f. : il., figs. e tabs. color.

Orientador: Evandro de B. Costa.
Coorientador: Rafael Ferreira L. de Mello.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2020.

Bibliografia: f. 92-101.

1. Codificação de mensagens. 2. Discussões online. 3. Presença social. 4. Comunidades de investigação. 5. Mineração de textos. I. Título.

CDU: 004.415.3

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PpgI
Instituto de Computação
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 I Telefone: (082) 3214-1401

Folha de Aprovação

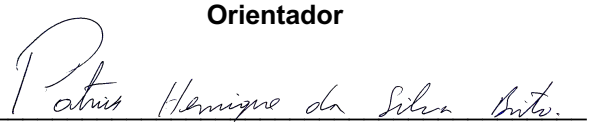
Jean Barros Teixeira

"IDENTIFICAÇÃO AUTOMÁTICA DA PRESENÇA SOCIAL EM DISCUSSÕES ONLINE ESCRITAS EM PORTUGUÊS"

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 28 de agosto de 2020.

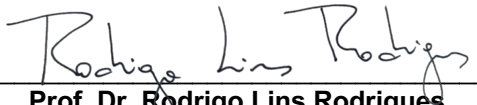
Banca Examinadora:

Prof. Dr. Evandro de Barros Costa
UFAL – Instituto de Computação
Orientador



Prof. Dr. Patrick Henrique da Silva Brito
UFAL - Curso de Ciência da Computação/Arapiraca
Examinador Interno

Prof. Dr. Rafael Ferreira Leite de Mello
UFRPE-Universidade Federal Rural de Pernambuco
Examinador Externo



Prof. Dr. Rodrigo Lins Rodrigues
UFRPE-Universidade Federal Rural de Pernambuco
Examinador Externo

Dedico este trabalho primeiramente a Deus pelas bençãos diárias e a Nossa Senhora por interceder durante toda a jornada. Aos meus pais por serem a minha base e a todos os familiares, namorada, amigos e professores que contribuíram de alguma forma.

AGRADECIMENTOS

A Deus e Nossa Senhora por me conduzirem pela jornada da vida e me abençoarem diariamente me dando forças para continuar sempre em frente.

Aos meus pais, José Ivo e Ana Lúcia, por serem os melhores pais e contribuírem diariamente na minha vida pessoal e profissional. Serei eternamente grato por todos os sacrifícios que fizeram e fazem por mim para que eu me tornasse o homem que sou hoje.

Ao meu tio Zé, por sempre investir na minha educação desde criança e acreditar no meu potencial. E a todos os familiares e amigos que me apoiaram durante toda essa fase.

A minha namorada, Carol, pelo companheirismo, amizade, paciência, compreensão, alegria e amor.

Aos meus orientadores Prof. Evandro e Prof. Rafael que confiaram no meu trabalho e me ajudaram a desenvolver minha pesquisa, contribuindo com pontos de melhoria e me ajudando a ser um pesquisador melhor.

Ao grupo de pesquisa AIbox.edu, da Universidade Federal Rural de Pernambuco, em especial o Máverick, que contribui na implementação da proposta.

Aos professores do PPGI - UFAL pelos ensinamentos durante o mestrado e a todos os professores que passaram pela minha vida e contribuíram para a minha formação acadêmica.

Aos amigos dos Correios que me incentivam a continuar meus estudos e a todos os meus gestores que me apoiam desde a graduação.

Em especial a minha querida tia Celene, segunda mãe, que sempre cuidou de mim e torceu pelo meu sucesso. Tenho certeza que continua cuidando de mim ao lado de Deus.

Não vos esforceis pelas honras do mundo, mas honrai o SENHOR.

— São Francisco de Assis

RESUMO

Esta dissertação de mestrado apresenta um método que permite a identificação automática de mensagens provenientes de fóruns *online* de ensino a distância escritas em português brasileiro. Particularmente, analisa o problema da codificação de mensagens de discussão segundo as categorias de presença social, um importante construto do modelo de Comunidade de Investigação amplamente utilizado na aprendizagem *online*. Apesar de existirem técnicas de codificação para a presença social na língua inglesa, a literatura ainda é carente em métodos para as demais línguas, como o português. O método aqui proposto utiliza-se de um conjunto de características provenientes da frequência de palavras e 158 características extraídas de dois recursos, LIWC e Coh-Metrix, disponíveis para análise textual através de técnicas de Mineração de Texto, para criar um classificador para cada uma das três categorias da presença social (Afetiva, Interativa e Coesiva). Para isso foram utilizados três tipos de algoritmos, Random Forest, AdaBoost e XGBoost onde o melhor modelo desenvolvido utilizou o algoritmo XGBoost atingindo 85,68% de acurácia e índice Kappa (k) de 0,71, o que representa uma concordância substancial, e está bem acima do grau de puro acaso. Este trabalho também provê uma análise da natureza da presença social, observando as características de classificação que foram mais relevantes para distinguir as três categorias da presença e uma análise comparativa sobre as principais características identificadas nas fases da presença social em diferentes domínios.

Palavras-chave: Presença Social, Modelo de Comunidade de Investigação, Discussões *Online*, Classificação de Texto.

ABSTRACT

This M.Sc. dissertation presents a method that allows the automatic identification of messages from distance learning online forums written in Brazilian Portuguese. In particular, it analyzes the problem of coding discussion messages according to the categories of social presence, an important construct of the Community of Inquiry (CoI) model widely used in online learning. Although there are coding techniques for social presence in the English language, the literature is still lacking in methods for other languages, such as Portuguese. The method proposed here uses a set of characteristics derived from the frequency of words and 158 characteristics extracted from two resources, LIWC and Coh-Metrix, available for textual analysis using Text Mining techniques, to create a classifier for each one of the three categories of social presence (Affective, Interactive and Cohesive). For that, three types of algorithms were used, Random Forest, AdaBoost and XGBoost where the best model developed used the XGBoost algorithm reaching 85.68% accuracy and Kappa index (k) of 0.71, which represents a substantial agreement, and is well above the level of pure chance. This work also provides an analysis of the nature of social presence, observing the classification characteristics that were most relevant to distinguish the three categories of presence and a comparative analysis on the main characteristics identified in the phases of social presence in different domains.

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

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AVA	Ambiente Virtual de Aprendizagem
BAGGING	Bootstrap Aggregating
BOW	Bag of Words
CM	Coh-Metrix
COI	Community of Inquiry
EAD	Educaçã à Distância
ENA	Epistemic Network Analysis
IDF	Inverse Document Frequency
KDT	Knowledge Discovery from Text
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
MDA	Mean Decrease in Accuracy
MDG	Mean Decrease in Gini
MT	Mineração de Texto
NILC	Núcleo Interinstitucional de Linguística da Computação
NLTK	Natural Language Toolkit
NUMPY	Numerical Python
OOB	Out-of-Bag
PLN	Processamento de lInguagem Natural
PSS	Presença Social
QCA	Quantitative Content Analysis
RF	Random Forest
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
SPD	Social Presence Density calculation
TF	Term Frequency
USP	Universidade de São Paulo
VC	Cross Validation
XGBOOST	Extreme Gradient Boosting

LISTA DE SÍMBOLOS

ϕ	Phi
Θ	Theta
β	Beta
$\in$	Pertence
$\sum$	Somatório

SUMÁRIO

1 INTRODUÇÃO

Devido ao surgimento da modalidade de ensino a distância, o acesso a educação se tornou mais fácil e popular. Segundo (ABED, 2018), em 2017 houveram 7.773.828 milhões de alunos matriculados em cursos a distância e um total de 4.570 cursos totalmente a distância no Brasil. A EAD no Brasil está definida através do Decreto nº 5.622 de 19 de dezembro de 2005, que a descreve como uma modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorre com a utilização de meios e tecnologias de informação e comunicação, com estudantes e professores desenvolvendo práticas educativas em lugares ou tempos diversos (BRASIL, 2005).

Diante desse contexto o Ambiente Virtual de Aprendizagem (AVA) é utilizado como um mecanismo para facilitar a interação entre professores/tutores e estudantes, além disso é uma ferramenta que vem ganhando destaque nos últimos anos (MCGILL; KLOBAS, 2009). Os AVAs apresentam diversos recursos que proporcionam essa interação, um dos mais populares é o fórum de discussão (SOARES et al., 2016). Cerca de 87,2% das instituições que ofertam cursos totalmente a distância utilizam o fórum como um recurso didático para apoiar seus alunos durante o processo de ensino-aprendizagem (ABED, 2018).

Os fóruns de discussão são ferramentas assíncronas que proporcionam uma enorme interatividade entre alunos e professores (GRAçAS; GOMES, 2011), a ferramenta permite: postagem de dúvidas, cometários sobre o conteúdo da disciplina, postagens de materiais extras entre outros. Uma pesquisa em aprendizagem online e educação a distância aponta que o envolvimento em fóruns de discussão assíncronos acarreta em uma melhora dos resultados acadêmicos (SUHANG et al., 2014). Portanto, analisar como se dá a interação entre os alunos e professores, torna-se bastante relevante para o processo pedagógico, visto que o aprendizado concreto no EAD ocorre quando os participantes envolvidos nesse processo conseguem formar uma comunidade de investigação (GARRISON et al., 1999).

O modelo da Comunidade de Investigação (*Community of Inquiry (CoI)*) é um dos modelos pedagógicos mais populares desenvolvido para apoiar os professores/tutores no desenvolvimento de experiências de aprendizagem *online* moderna, além de ser um dos mais pesquisados e validados no domínio da EAD (GARRISON et al., 1999). O modelo estabelece três elementos, conhecidos como presenças, que modelam o aprendizado online dos alunos: presença social, presença de ensino e presença cognitiva. Dentre elas, a presença social demonstra ser um elemento importante para o sucesso da experiência educacional (GARRISON et al.,

1999) e também é considerada relevante na observação da forma com que os alunos se lançam nas interações e em sua manutenção nos cursos a distância (PALLOFF; PRATT, 2004). Na literatura, também encontramos outros modelos que medem a colaboração em discussões online assíncronas como o citado em (MURPHY, 2004), onde é apresentado um modelo de colaboração, conceituado como uma série de processos ou estágios que vão da interação para a colaboração. Ainda assim o modelo CoI destaca-se devido ao grande número de pesquisas que o utilizam.

Um desafio na área de informática na educação é identificar essas presenças automaticamente. Diante disto, este trabalho de dissertação apresenta uma proposta para facilitar o processo de identificação da presença social no contexto da língua portuguesa, visto que os métodos utilizados são bastante trabalhosos, pois são realizados de forma manual.

1.1 PROBLEMÁTICA

A Análise de Conteúdo Quantitativo, do inglês *Quantitative Content Analysis (QCA)*, é apontada como um método largamente adotado no contexto das três presenças de CoI (STRIJBOS et al., 2006) com o objetivo de medir/avaliar os processos relacionados a construção do conhecimento nas discussões online e fornecer suposições válidas e confiáveis a partir da análise de dados textuais (BAUER, 2007). O modelo CoI especifica três modelos de codificação QCA, cada presença possui seu respectivo modelo, que podem ser aplicados na análise de mensagens de discussão online.

Um grande problema para aplicar o método QCA no contexto da presença social em mensagens de fóruns de discussão online, é que devido ao enorme número de interações é gerado um imenso volume de dados dificultando a análise exclusivamente manual (AZEVEDO, 2011) por parte do professor/tutor. Além de tudo, para o professor é um grande desafio desenvolver aulas, que englobem comunidade e colaboração, que possam ajudar os alunos a alcançarem um aprendizado profundo e significativo em meio a esse grande volume de dados (LIPMAN, 2003); (RAMSDEN, 2003).

Outro problema enfrentado por professores e tutores é identificar de forma eficiente, em meio a muitos alunos, quais são aqueles menos participativos. Com pouca participação do aluno, os professores não podem avaliar corretamente seus níveis de compreensão e muitas vezes fazem suposições incorretas (RICHARDS; VELASQUEZ, 2014). Além disso, a interação limitada causa dificuldades para os professores diferenciarem o ritmo e a instrução que se adaptam aos diferentes níveis de progresso dos alunos (GOODWIN; MILLER, 2013).

Segundo (YU; ZHENG, 2017), através de técnicas de mineração de textos, a dificuldade para avaliar as mensagens dos fóruns pode ser reduzida, e ainda contribui para a diminuição das inconsistências nas avaliações manuais. Nesta perspectiva, (GAŠEVIĆ et al., 2017) mostram excelentes resultados da utilização dos métodos de análise automática de texto para análise de aprendizagem, sinalizando para o potencias destes em tornar a avaliação das presenças do CoI mais simples e menos custosa.

Na literatura, encontramos poucos trabalhos que visam automatizar a identificação da presença social através de técnicas de mineração de texto ((TANIGUCHI et al., 2019); (FERREIRA et al., 2020)) mas seus métodos são voltados para cursos em inglês, limitando sua utilização para o contexto da língua inglesa. De modo igual, a disponibilidade de ferramentas que viabilizem a análise de texto para outros idiomas além do inglês, português por exemplo, é bastante limitada, consequentemente prejudicam de forma significativa a precisão dos modelos desenvolvidos para os demais idiomas.

Sendo assim, se os professores possuíssem uma ferramenta para realizar a análise automática da presença social em textos de fóruns de discussão online escritos em português, poderiam de forma fácil e menos trabalhosa utilizar o modelo CoI para direcionar intervenções e afetar os resultados de aprendizagem dos alunos (KOVANOVIC et al., 2014a).

1.2 QUESTÕES DE PESQUISA

Diante da problemática descrita nesse trabalho, é apresentada as seguintes questões de pesquisa:

Questão de Pesquisa 1 (QP01): *Um método que utiliza técnicas de mineração de texto, aplicado em fóruns educacionais escritos em português, classifica automaticamente as categorias da presença social com uma concordância (k) moderada a quase perfeita?*

Além da questão citada acima, pretende-se também disponibilizar informações sobre quais as características que são mais relevantes para classificar cada uma das três categorias. Para isso, utilizamos alguns parâmetros aplicados por (KOVANOVIĆ et al., 2016), (NETO et al., 2018) e (FERREIRA et al., 2020) com essa mesma finalidade. Então nossa segunda questão de pesquisa é:

Questão de Pesquisa 2 (QP02): *Quais características, dentre as extraídas dos recursos LIWC, Coh-Metrix e Bag of Words, tem maior relevância preditiva para os classificadores em cada categoria da presença social?*

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Dessa forma o objetivo principal deste trabalho é desenvolver um método para a identificação automática da presença social em discussões assíncronas escritas em português utilizando técnicas de mineração de texto e avaliar os resultados de sua aplicação.

1.3.2 Objetivos Específicos

Para alcançar o objetivo principal apontado, os seguintes objetivos específicos devem ser atingidos:

- Identificar os métodos de análise da presença social em discussões *online* existentes na literatura;
- Verificar quais as técnicas de mineração de texto que podem auxiliar na identificação da presença social;
- Criar modelos para identificar cada uma das três categorias da presença social em mensagens escritas em português provenientes de fóruns educacionais *online*;
- Avaliar o modelo utilizando uma base de dados extraídas de discussões online em português;
- Realizar uma análise comparativa das características da presença social em duas turmas de cursos EaD de contextos diferentes.

1.4 RELEVÂNCIA

A Comunidade de Investigação (CoI) surgiu nas últimas duas décadas como o modelo mais amplamente citado tanto para o desenvolvimento de cursos quanto para o ensino de pesquisa em educação online (BOZKURT et al., 2015). Esse modelo tem sido usado para desenvolver muitos cursos e programas online, e também utilizado como modelo conceitual para centenas de teses e demais pesquisas (MOISEY et al., 2016). Por exemplo, sistemas de realidade imersiva (MCKERLICH; ANDERSON, 2007), tecnologias síncronas (FAYRAM, 2017), wikis (LAMBERT; FISHER, 2009) e MOOCs (FRAU-MEIGS; BOSSU, 2017); (HOLSTEIN; COHEN, 2016).

Um dos grandes problemas apontados neste trabalho é a dificuldade do professor analisar de forma manual as postagens realizadas nos fóruns educacionais devido ao grande volume de dados (AZEVEDO, 2011), mais precisamente identificar a dimensão presença social de uma comunidade de investigação. Existem alguns trabalhos na literatura que abordam esse tema, um deles apresenta um estudo entre as relações causais entre as três presenças mas utiliza-se de um questionário submetido aos alunos com grupos de questões referentes a cada dimensão, ou seja, utiliza-se de um mecanismo manual para identificar a existência de cada uma das presenças (GARRISON et al., 2010).

Já em (ROLIM et al., 2019 no prelo) é apresentada uma abordagem para estudar as relações entre a presença social dos alunos e os tópicos a partir das mensagens, nesse caso para a codificação das mensagens para uma das três categorias de presença social foram utilizados dois codificadores para realizar esse trabalho de forma manual. Em contrapartida, um estudo mais recente realizado por (FERREIRA et al., 2020), propõe uma abordagem para classificar automaticamente as mensagens de discussão online escritas em inglês, de acordo com as categorias da presença social.

Percebe-se que esses trabalhos, em sua maioria, realizam a codificação da presença social de forma manual, além disso o foco tem sido cursos em inglês, limitando sua utilização apenas para o contexto da língua inglesa. A presença social mede a capacidade de humanizar os relacionamentos entre os participantes de uma discussão. Ela se concentra nas interações sociais e tenta modelar o clima social dentro de um grupo de alunos (ou seja, coesão, afetividade e comunicação aberta) (GARRISON et al., 1999). Segundo (GARRISON et al., 2001), ela torna-se necessária para o desenvolvimento da presença cognitiva, apoiando no trabalho de facilitar o processo de pensamento critico, visto que os participantes de uma comunidade precisam estar socialmente envolvidos e emocionalmente motivados para interagir no contexto virtual. Dessa forma a presença social demonstra ser um elemento relevante para o sucesso da experiência educacional.

Diante da relevância do tema abordado, a solução a ser desenvolvida nesse trabalho vislumbra avaliar um modelo para classificação automática da presença social em discussões assíncronas online em ambientes virtuais de aprendizagem no contexto da língua portuguesa, uma vez que a maioria dos métodos existentes são realizados de forma manual e voltados para a língua inglesa.

1.5 ORGANIZAÇÃO DO TRABALHO

Este trabalho está segmentado do seguinte modo: no Capítulo 2 é apresentada a fundamentação teórica, contendo os conceitos essenciais para o entendimento do trabalho realizado. No Capítulo 3 serão expostos os principais trabalhos relacionados a esta dissertação, dividindo-os entre os aplicados para o inglês e português.

O Capítulo 4 especifica a metodologia utilizada para o desenvolvimento do método proposto neste trabalho, com o objetivo de automatizar a análise da presença social utilizando Mineração de Texto. No Capítulo 5 serão exibidos os resultados obtidos com a aplicação do método em um *corpus* formado por mensagens de fóruns de discussão *online* escritos em português e as discussões desses resultados.

Por fim, o Capítulo 6 apresenta as considerações finais, contribuições da pesquisa, artigo submetido, as limitações da pesquisa e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos elementares para o entendimento desta dissertação, dividido em três seções. A primeira aborda a Comunidade de Investigação (CoI) e as três dimensões que a constitui. A segunda seção apresenta os principais fundamentos aplicados neste trabalho relacionados a mineração de textos. Finalmente, na terceira seção são expostas as ferramentas que foram utilizadas para o desenvolvimento da pesquisa.

2.1 O MODELO DE COMUNIDADE DE INVESTIGAÇÃO (COI)

CoI é um modelo conceitual de análise que tem como foco o processo social de construção conjunta e colaborativa do conhecimento em ambientes de comunicação assíncrona baseada em texto como proposto em (GARRISON et al., 1999). Ele surgiu através de um estudo que investigava os efeitos na qualidade do processo de aprendizagem e os seus resultados, por meio de ensino baseado na comunicação mediada por computador, em textos oriundos de ambientes virtuais de aprendizagem de nível superior.

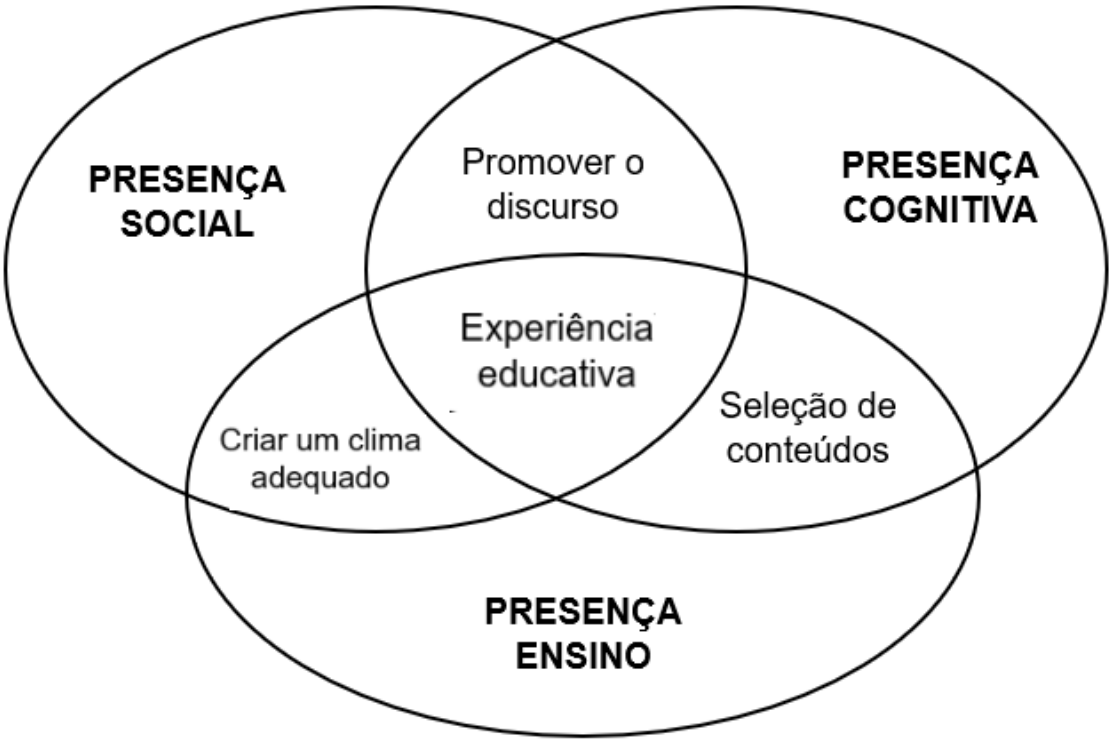
Segundo (GARRISON et al., 2010) o CoI é a estrutura teórica mais utilizada e pesquisada que descreve as importantes características das interações sociais na educação online e mista. Para os autores, a educação virtual permite grandes níveis de interação entre professores e alunos, principais integrantes do processo educacional, e que, ao se integrar em um trabalho conjunto, possibilitam desenvolver capacidades cognitivas mais complexas, produzindo uma aprendizagem mais rica, organizada e com aprofundamento contínuo.

Esse modelo é bastante utilizado em diversas áreas, por exemplo, em (SHIELDS, 2003) argumenta que o CoI tem muito a oferecer na área de administração pública, em que ele é uma posição ideal a partir da qual os administradores públicos podem examinar efetivamente como abordam os problemas, consideram os dados e se comunicam. Já em (GOOS, 2004), considera que a partir do CoI pode direcionar quais ações específicas um professor pode realizar para criar uma cultura de investigação em uma sala de aula de matemática do ensino médio.

O CoI explica comportamentos de alunos e professores com o objetivo de descrever como a experiência educacional pode ser mais eficaz, fornecendo três indicadores/dimensões de seus relacionamentos, conhecidos como presenças (GARRISON et al., 1999) e que além de fundamentais para que ocorra uma prática educacional de sucesso possuem influência entre elas, as três dimensões são: a presença social, presença de ensino e presença cognitiva, conforme ilustrado na Figura 1.

Em seguida, cada dimensão será detalhada, juntamente com suas fases e indicadores utilizados para detalhá-las e identificá-las.

Figura 1 – Modelo de uma comunidade de investigação. (adaptado de (GARRISON et al., 2001)



2.1.1 Presença Social

De acordo com a definição de (ANDERSON et al., 2001), a presença social é a capacidade dos participantes de uma comunidade de investigação de se projetarem social e emocionalmente, como pessoas reais, ou seja, sua personalidade completa, através do meio de comunicação em uso. Os autores reiteram que esta presença deve ser bem mais que estabelecer, de forma simples, a presença sócio-emocional e as relações pessoais. A coesão do grupo e da comunidade, relaciona-se com a intensidade em que os participantes possuem objetivos em comum e a demonstração de amizade mútua. Além disso, em contraste com a interação face a face, nas discussões online, é essencial expressar textualmente tais habilidades, a fim de estabelecer uma comunicação socio-emocional (GARRISON; ARBAUGH, 2007). Para que os relacionamentos sejam desenvolvidos, é necessário um certo tempo para que seja alcançado um nível de conforto, confiança e um sentimento de pertencimento.

Diante desse contexto, a presença social torna-se necessária para o desenvolvimento da presença cognitiva, apoiando no trabalho de facilitar o processo de pensamento critico, visto que os participantes de uma comunidade precisam estar socialmente envolvidos e emocionalmente

motivados para interagir no contexto virtual (GARRISON et al., 2001). Por exemplo, a correlação da coesão do grupo (presença social) e evento desencadeante (presença cognitiva) foi de 0,662 (KOZAN, 2016). Dessa forma a presença social demonstra ser um elemento relevante para o sucesso da experiência educacional. Dada a importância da presença social, (GARRISON et al., 2001) desenvolveram um esquema de classificação que a divide em três categorias fundamentais: Expressões de Afetividade (Afetiva), Comunicação Aberta (Interativa) e Coesão de Grupo (Coesiva).

A primeira categoria, Expressões de Afetividade ou Afetiva, é composta por expressões de emoção, sentimentos, crenças e valores, apresentados através de repetições de pontuação, frase inteira escrita em caixa alta ou ainda uso de emoticons. Também é levado em consideração o senso de humor das pessoas, através de ironias, piadas ou sarcasmos. E ainda as demonstrações pessoais sobre aspectos não relacionados ao curso em si, como particularidades da vida pessoal, gostos pessoais e até expressões de vulnerabilidades. A expressão das emoções, bem como a persistência e o interesse, constituem-se como fatores essenciais a toda a experiência educativa e são facilitadores do diálogo autêntico e necessário a uma experiência formativa (GARRISON, 2011).

A segunda categoria, Comunicação aberta ou Interativa é definida como a troca de comunicação reciproca e respeitosa (GARRISON et al., 2001). Os autores citam a consciência mútua e reconhecimento da contribuição, como exemplos desta categoria. A consciência mútua contribui modelando as atividades de aprendizagem de cada um dos participantes e o reconhecimento é o processo que serve para promover o desenvolvimento e a manutenção da troca de relacionamentos. O principal objetivo desta categoria é melhorar a comunicação aberta entre os alunos. Os autores também afirmam que a interatividade apresenta-se nos ambientes assíncronos quando é utilizado o recurso de respostas, disponibilizados pela ferramenta, para postar mensagens; seja por meio da citação de mensagens de outros integrantes, quando um comentário é direcionado a uma pessoa em particular e também por meio de referencias explicitas as mensagens de outros integrantes.

Por fim, a ultima categoria conhecida como Coesão de Grupo ou coesiva é exemplificada pelos autores através de praticas que constroem e sustentam o senso de envolvimento do grupo. Seu principio é que, quando os alunos sentem-se como membros do grupo e não apenas indivíduos isolados, a qualidade do discurso se torna fácil e otimizada. Os autores afirmam que a categoria Coesiva apresenta-se através da utilização de vocativos, o emprego de pronomes inclusivos e as saudações. Estas características mencionadas são importantes expressões de

coesão, pois possibilitam a participação e a afinidade e indicam uma tentativa de estreitamento do relacionamento com o destinatário da mensagem.

A Tabela 1 apresenta as categorias, indicadores e definições sugeridos por (ROURKE et al., 2001), que auxiliam na identificação da presença social em discursos escritos.

Tabela 1 – Categorias indicadores e definição da Presença Social

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	Referir-se ao grupo por "nós "nosso grupo".
	Saudações	Comunicação social, saudações, despedidas.

Fonte: Adaptado de (ROURKE et al., 2001).

2.1.2 Presença de Ensino

Segundo (ANDERSON et al., 2001) definem a presença de ensino como a ação de projetar, facilitar e direcionar os processos cognitivo e social com o objetivo de obter resultados

de aprendizagem significativos e de valor educacional. Essa presença refere-se ao papel dos professores e dos alunos antes, ou seja, na concepção do curso e durante o curso desde a facilitação até a instrução direta.

Para (GARRISON, 2011), a atuação do professor na presença de ensino é um componente fundamental no processo de ensino-aprendizagem, principalmente no ensino a distância, que se torna um desafio ainda maior quando comparamos ao ensino presencial. Para os autores, além de conhecer o assunto do curso, o professor deve ter uma estrategia educacional e ser um facilitador social. Dessa forma, cabe a ele promover ações, comportamentos e estratégias, com o objetivo de proporcionar um espaço mais acessível, confortável e seguro para apoiar a experiência geral de aprendizagem do aluno. Sendo assim, três categorias caracterizam a presença de ensino: Desenho Instrucional e Organização, Facilitação do Discurso e Instrução Direta.

A primeira categoria, Desenho Instrucional e Organização, esta associada com o planejamento e projeto das questões de estrutura, processo, interação e avaliação do curso online. (GARRISON; ARBAUGH, 2007) trazem alguns exemplos de atividades que integram esta categoria, a recriação de apresentações PowerPoint e anotações das aulas; elaborar cronograma para tarefas individuais e de grupo; desenvolvimento de pequenas atividades em áudio/vídeo; indicações pessoais sobre o material do curso. Essas atividades cujo professores realizam, precisam ser bastante claras pois o sucesso dos cursos online depende bastante de uma estrutura consistente e transparente, fornecendo o apoio necessário aos envolvidos no processo. A segunda categoria, a Facilitação do Discurso, trata-se do meio pelo qual os estudantes estão envolvidos na interação e na elaboração do conhecimento mediante os recursos oferecidos pelo curso. Segundo (GARRISON; ARBAUGH, 2007), esta categoria deve ser coerente com os resultados que apoiam a importância da interação dos alunos para a efetividade da aprendizagem online. De acordo com (RICHARDSON et al., 2015), a participação do professor é vital na modalidade online. Dessa forma, cabe a ele a responsabilidade de rever e comentar as respostas dos alunos; sugerir novas indagações e fazer comentários com o intuito de direcionar a discussão ao objetivo desejado, de forma eficaz; estimular os alunos menos ativos e estabelecer o número de interações daqueles que dominam as discussões, impedindo que esses prejudiquem o processo de aprendizagem em grupo (ANDERSON et al., 2001). Por fim, na terceira categoria, a Aprendizagem Direta, os professores exercem a liderança intelectual e acadêmica, e partilham seu conhecimento sobre o assunto com os alunos (ANDERSON et al., 2001). Os autores complementam que a função do professor vai além de simples facilitador do conteúdo, eles precisam analisar os comentários, inserir fontes de informação, direcionar as discussões e instruir os alunos para alcançarem novos conhecimentos.

(GARRISON; ARBAUGH, 2007), ressaltam que o professor possui a responsabilidade com foco em facilitar a reflexão e o discurso dos alunos, através da divulgação do conteúdo, utilizando os meios disponíveis para a avaliação e feedback. Os autores também afirmam que a presença social do professor é fundamental para o êxito desse tipo de comunicação, exigindo-o conhecimentos técnicos e pedagógicos para correlacionar as contribuições, identificar conceitos errados e introduzir o conhecimento consolidado em referencias, por exemplo, em livros, artigos científicos e recursos educativos disponíveis na Web.

A Tabela 2 apresenta um resumo, elaborado por (ANDERSON et al., 2001) , referente as categorias da presença de ensino e seus respectivos indicadores e definições.

Tabela 2 – Categorias, indicadores e definição da Presença de Ensino

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 da 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.
	Induzir 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.

Estudos apontam que as percepções dos alunos sobre a presença de ensino e a presença social em cursos online podem influenciar a experiência geral de aprendizagem (GARRISON et al., 2001) e (SWAN, 2002) , sugerindo que uma deficiência resulta em um afastamento do curso (BOLLIGER; INAN, 2012), enquanto níveis mais altos promovem persistência (HART, 2012).

2.1.3 Presença Cognitiva

A presença cognitiva é uma medida em que os participantes em qualquer configuração específica de uma comunidade de indivíduos são capazes de construir significado por meio de uma comunicação sustentada (GARRISON et al., 2001). Segundo (GARRISON et al., 1999), é o elemento do modelo CoI mais básico para o sucesso no ensino superior. A presença cognitiva está definida em um modelo prático de investigação com o objetivo de descrever e compreender sua presença em processos educacionais e o desenvolvimento do processo de pensamento critico (GARRISON et al., 2001). Estes autores a dividem em quatro fases, são elas: Evento desencadeador, Exploração, Integração e Resolução.

A primeira fase, Evento desencadeador, descreve o início das discussões e representa a fase inicial do processo de investigação crítica. Durante esta fase, um dado problema ou uma questão, oriundo da experiência ou da análise do contexto educacional, é apresentado no fórum de discussões. Em fóruns mais democráticos, não somente professor/moderador podem lançar um evento desencadeador, o aluno também pode desempenhar esse papel, por exemplo, exprimindo alguma dúvida (GARRISON et al., 2001). Mas, os autores enfatizam a importância do envolvimento do professor no inicio da interação garantindo que não haja um desvio do objetivo da atividade.

Por sua vez, a segunda fase, Exploração refere-se ao momento em que os alunos alternam entre o mundo particular, reflexivo e a exploração social das ideias (GARRISON et al., 2001). Nesta fase, primeiramente os alunos precisam perceber ou compreender a origem do problema para, em seguida, começar a explorar novas fontes para coletar informações mais relevantes. A fase de exploração tem como característica o aumento da divergência, o questionamento, e a troca de informações. Dessa forma os alunos tornam-se mais seletivos quanto a relevância para a questão ou problema.

Na terceira fase, Integração é caracterizada como a fase em que os alunos constroem significados baseados em ideias desenvolvidas durante a exploração, conectando informações e descobertas relevantes e formulam hipóteses (GARRISON et al., 2001). Segundo os autores, no decorrer da passagem da fase exploratória para a de integração, os alunos começam o processo de avaliação da aplicabilidade para o problema descrito ou evento desencadeador em questão. Ressaltam também, a importância da presença de ensino atuando na descoberta de conceitos errados, concedendo perguntas,informações adicionais e comentários, com o objetivo de garantir o desenvolvimento cognitivo e modelar o processo de pensamento crítico.

Na quarta e última fase, Resolução é onde acontece a concepção do conhecimento e sua

possível aplicação em problemas práticos (GARRISON et al., 2001). Nesse contexto educacional, esta abordagem mostra-se mais complexa, pois requer outros atores dispostos a fazer os testes das hipóteses e a construir um entendimento 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. Levando em consideração que os alunos adquiriram conhecimento relevante, eles são induzidos a aplicá-los na resolução de novos problemas, fazendo com que o ciclo lógico do modelo de investigação se inicie novamente.A Tabela 3 apresenta as quatro fases da presença cognitiva, elaborada por (GARRISON et al., 2001), com seus respectivos indicadores e definições.

Tabela 3 – Categorias, indicadores e definição da Presença Cognitiva

Categoria	**Indicador**	**Definição**
Evento desencadeador	Reconhecer o problema	Apesentar 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).
	Brainstorming	Acrescenta novas ideias mas não 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 ideais, 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 no 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 ideais que defende no que diz respeito ao problema proposto.

2.2 MINERAÇÃO DE TEXTO

A mineração de texto, também conhecida como descoberta de conhecimento a partir de texto (Knowledge Discovery from Text - KDT), é um campo da linguística de computador que, e por sua vez, trata da aplicação da ciência da computação à análise e fala da fala síntese (KLAHOLD; FATHI, 2020). Segundo (REZENDE et al., 2011), é uma área da mineração de dados, podendo ser definida como uma coleção de técnicas e processos para descoberta de conhecimento inovador a partir de dados textuais. Seu objetivo é extrair informações relevantes, através da identificação e exploração de padrões interessantes em bases de textos não estruturadas, ou semi-estruturadas (FELDMAN et al., 2007). Na Mineração de Textos, geralmente a informação que pretende-se extrair é clara, sendo explicitada nos textos, mas, o problema é que a informação não é representada de uma maneira que seja passível de processamento automático (WITTEN; FRANK, 2002).Sendo assim, o processo geralmente envolve diversas técnicas, incluindo recuperação de informações, mineração de dados, aprendizado de máquina, processamento de linguagem natural (PLN), classificação de texto e gerenciamento de conhecimento (FELDMAN et al., 2007).

A mineração de textos apresenta alto grau de dependência com relação a processamento de linguagem natural (PLN), que é uma coleção de técnicas computacionais para análise automática e representação de linguagens humanas, motivada pela teoria (CHOWDHARY, 2020).

A importância da Mineração de Texto está relacionada com a grande volume de dados produzidos e disponibilizados em formato digital atualmente (TURNER et al., 2014), pois uma parcela desses dados está no formato textual, como e-mails, relatórios, boletins, artigos,

registros de pacientes e conteúdo de páginas Web (ROSSI, 2016). A descoberta de conhecimento a partir de dados disponíveis significa identificar esses dados, encontrar o que é relevante e ter a possibilidade de processá-los. Realizar as atividades de organizar, analisar, e extrair o conhecimento incorporado nesses dados manualmente, é uma tarefa extremamente difícil de ser executada por um ser humano (CHOWDHARY, 2020).

Comumente, para extrair 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 atividades abordadas neste trabalho serão detalhadas nas próximas subseções.

2.2.1 Coleta de Dados

De acordo com (ARANHA et al., 2007) , quando estamos diante de um problema de classificação de textos, é necessário obter um conjunto de dados para treinamento. Dessa forma, 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 denominada corpus e o conjunto destes é chamado de corpora. Um corpus é que uma coleção de textos, que representa uma ou um conjunto de linguagens naturais e, a criação deste conjunto de treino revela-se uma tarefa custosa, uma vez que na maioria dos casos exige-se processos manuais através de anotação de especialistas (INDURKHYA; DAMERAU, 2010)

Para a construção do corpus,é necessário a coleta previa de uma grande quantidade de documentos. Os documentos podem ser coletados de diversas fontes, como redes sociais, e-mails, campos textuais em banco de dados, páginas da Web, chats e fóruns de ambientes online, de acordo com a relevância de cada um perante o domínio de estudo.

Esta é uma etapa muito importante, que exige esforço e cuidado, com o objetivo de obter material de qualidade e que sirva de insumo para a aquisição de conhecimento.

2.2.2 Pré-processamento

Esta etapa é executada após a coleta dos dados, geralmente tem um alto custo de tempo pois exige a utilização de diversos algoritmos e técnicas variadas que são aplicadas de acordo com o domínio do estudo. O principal objetivo de pré-processar um texto, consiste na filtragem e limpeza dos dados, eliminando redundâncias e informações desnecessárias para o conhecimento que se deseja extrair (GONÇALVES et al., 2006). De forma simples, esta fase prepara o conjunto de dados coletados para servir de insumo para a fase de extração de características.

Existem uma variedade de técnicas que podem ser utilizadas, de forma separadas ou combinadas, como:

- **Tokenização:** Esta é uma etapa muito importante do pré-processamento com a finalidade de extrair unidades minímas do texto. Cada unidade é denominada de token, normalmente, corresponde a uma palavra do texto mas pode ser um número, sinal, caracteres de pontuação (SCHÜTZE et al., 2008);

- **Remoção de stopwords:** Ao se trabalhar com uma base de textos, encontra-se diversos tokens que possuem pouco significância em um texto. Esses tokens sao conhecidos como stopwords, normalmente correspondem aos artigos, preposições, pontuação, conjunções e pronomes de uma língua (CAMBRIDGE, 2009);

- **Stemming:** Após a retirada das stopwords, pode-se aplicar a técnica de stemming para reduzir cada palavra do léxico. Essa técnica 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 de Conhecimento

É nesta etapa que de fato inicia-se a mineração de dados. A extração ou descoberta de conhecimento significa identificar, receber informações relevantes e poder computá-las e agregá-las ao seu conhecimento prévio, mudando o estado de conhecimento atual, com o objetivo de que determinada situação ou problema possa ser resolvido.(TALVEZ ADICIONAR CONCEITO DE KDDT) As técnicas utilizadas para a execução desta etapa na metodologia proposta são elencadas a seguir.

2.2.3.1 Classificação de texto

A enorme quantidade de documentos disponíveis em formato digital e a necessidade de encontrar informações relevantes são um grande desafio para os seres humanos. Segundo (BALI; GORE, 2015), o principal objetivo da classificação de texto é treinar o classificador com base em categorias predefinidas. Com a finalidade de utilizar o classificador para determinar se um novo documento pertence a uma das classes. Essa tarefa é descrita como uma função: $\phi : DxC \rightarrow \{T, F\}$, onde $D = \left\{d1, d2, d3, ...d_{|D|}\right\}$ é o conjunto que corresponde ao dominio de documentos e $C = \left\{C1, C2, C3, ...C_{|c|}\right\}$ é o conjunto preestabelecido de classes, nesse trabalho as categorias. O valor *T* atribuído a $< d_j, c_i >$ indica um proposito de classificar d_j

como c_i, e F indica que d_j não é classificado como c_i (SEBASTIANI, 2002). Dessa forma o classificador é a função que descreve como os documentos são classificados.

Para realizar a classificação automática de textos, uma das maneiras é utilizando técnicas e algoritmos de Inteligencia Artificial (LECUN et al., 2015). 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, 2016).

A aprendizagem supervisionada, dentre as categorias de aprendizagem de máquina, destaca-se em realizar a tarefa de classificação de texto. Algoritmos de aprendizagem supervisionada são algoritmos que aprendem por meio de um conjunto de treinamento (instâncias de exemplos), já classificados por um especialista de domínio, que possuem dados de entrada e saída (GOODFELLOW et al., 2016). A associação entre entrada e saída é definida por uma função, também chamada de modelo de classificação. Segundo (NORVIG; RUSSELL, 2014), quando o modelo de classificação verifica se um determinado atributo pertence ou não a uma classe, chamamos de classificação binária. Para que o modelo gerado seja capaz de predizer novas instâncias de forma eficiente é importante que o conjunto de dados de treinamento seja de qualidade.

2.2.3.2 Extração de Características

Para (SHAH; PATEL, 2016), extração de características é o processo de gerar novas características a partir de características existentes nos documentos. Essa técnica tem a finalidade de criar novas características que representem, mais satisfatoriamente, os padrões das instâncias de documentos. As características consideradas potencialmente significativas para um dado contexto são também denominadas como features, atributos, variáveis (WEISS; KULIKOWSKI, 1991).Os exemplos utilizados no processo de classificação de texto mencionado na seção anterior são representados por conjuntos de características.

Dessa forma, a etapa de extração de características é um processo primordial para o reconhecimento de padrões, de maneira que quanto mais relevantes as características utilizadas para representar os documentos, melhor será o classificador, consequentemente mais confiável será a classificação.

Para realização da extração de características, existem várias técnicas, como por exemplo **Bag of Words (BoW)** - cada documento é um vetor em um espaço multidimensional, e cada dimensão é um termo da coleção (FELDMAN et al., 2007) , **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 **Linguistic Inquiry and Word Count (LIWC)** - dicionário léxico (TAUSCZIK; PENNEBAKER, 2010) e **Coh-Metrix** - dicionário léxico, sintático e semântico (GRAESSER et al., 2004).

2.2.3.3 Linguistic Inquiry and Word Count (LIWC)

O LIWC é uma ferramenta desenvolvida por (PENNEBAKER et al., 2001) com o intuito de fornecer um método eficiente para estudos sobre fatores emocionais, psicológicos, cognitivos entre outros, presentes em trechos de falas verbais e escritas de indivíduos.

Desde o seu lançamento em 2001 recebeu atualizações em 2007 e 2015 com mudanças tanto na categorização de palavras quanto nas categorias em si. Aquelas categorias que possuíam uma taxa muito baixa de palavras foram removidas e outras adicionadas. A atualização de 2007 analisou mais de 100 milhões de palavras, com os detalhes do processo sendo documentado em (PENNEBAKER et al., 2007). Segundo os autores, a ferramenta funciona analisando os documentos e contando as palavras contidas nele, onde cada palavra é comparada ao dicionário definido pelo usuário e classificada em uma ou mais categorias deste modo a porcentagem da ocorrência das palavras e das categorias relacionadas sugere a tendência do texto a cada uma das categoria psicológicas.

Em sua versão original em inglês o dicionário possui aproximadamente 6.540 palavras, cada uma delas associada a uma ou mais categorias dentre as 73 disponíveis (TAUSCZIK; PENNEBAKER, 2010). 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.

Nesse trabalho foi utilizada a versão em português que possui cerca de 127.149 palavras que estão assinaladas a uma ou mais das 64 categorias (social, afetiva, concordância dentre

outras) disponíveis (FILHO et al., 2013). Essa versão foi de um projeto 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, desenvolvendo o dicionário do LIWC em português usando como base o dicionário original English LIWC Dictionary e disponibilizado-o no site PortLEX[1].

No dicionário, as palavras possuem rótulos que foram nomeados com códigos numéricos que representam as categorias as quais pertencem. Por exemplo, a palavra “amigo” no dicionário pertence a 5 categorias: discurso informal(gíria), amizade, humano, afeto e emoções positivas. Estas são representadas no dicionário pelos códigos: 22 , 123, 124, 125 e 126, respectivamente.

Sintetizando o que foi colocado na presente seção, o LIWC é uma ferramenta de análise textual em que o documento é estruturado em categorias, de acordo com um dicionário, onde cada palavra desse documento é atribuída uma ou mais categorias correspondentes ao dicionário.

2.2.3.4 Coh-Metrix

A ferramenta Coh-Metrix, que quer dizer cohesion metrics, foi desenvolvida para análise textual em inglês por pesquisadores da Universidade de Memphis, nos Estados Unidos (GRAESSER et al., 2004). Ela extrai de um texto características que influenciam em sua coesão, coerência e em sua facilidade ou dificuldade de leitura (SCARTON et al., 2010). Além disso, possui a função de indicar dados para identificar problemas textuais de ordem estrutural (FINATTO, 2011).

Os autores da ferramenta propõem uma diferenciação entre coesão e coerência em um texto, onde a coesão é uma característica do texto e a coerência é uma característica da representação mental do conteúdo do texto estabelecido pelo leitor (GRAESSER et al., 2004). Os autores também mencionam que a relação entre coesão e coerência é fortemente influenciado pelo conhecimento de mundo do leitor, assim como por suas habilidades de interpretação e raciocínio e pelos construtos coesivos do texto explícito.

O Coh-Metrix, integra a saída de diversas outras ferramentas de PLN podendo ser utilizada em diversos cenários de análise e classificação de textos. A ferramenta contém diversas métricas, coletadas e avaliadas ao longo do tempo pelos autores, que medem características do texto relacionadas a palavras, sentenças e à conexão entre sentenças (GRAESSER et al., 2011).

A ferramenta foi adaptada para o português do Brasil pelo Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo (USP), como consequência do

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

projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) que tinha como objetivo 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). Essa adaptação deu origem ao Coh-Metrix-PT [2], que possui 48 medidas implementadas de nível léxico, sintático em nível de sintagmas nominais, semântico, e discursivo (SCARTON et al., 2010).

Em sua versão original [3] a ferramenta possui 108 medidas para a língua inglesa. Com o objetivo de implementar mais medidas para a língua portuguesa, além das 48 medidas existentes, o grupo de pesquisa AIbox.edu composto por alunos do curso de Computação da Universidade Federal Rural de Pernambuco desenvolveu uma versão que implementa 94 medidas. Neste trabalho iremos utilizar essa versão, na seção 4.3.2, a Tabela 14 exibe as categorias disponíveis e uma descrição sobre cada uma delas.

2.2.3.5 Bag of Words (BoW)

Conforme mencionado em (ZHAO; MAO, 2017), a representação de documento é a peça fundamental para várias tarefas de mineração de texto e PLN. Para que os dados textuais sejam processados pelos algoritmos de aprendizagem de máquina, deve-se buscar essa estruturação ou representação do documento. O modelo frequentemente utilizado para representação de dados textuais é o modelo espaço-vetorial, onde cada documento é um vetor em um espaço multidimensional, e cada dimensão é um termo da coleção (FELDMAN et al., 2007).

Uma abordagem comumente adotada e eficaz para a representação de documentos é o modelo Bag of Words (BoW), onde o documento é representado por um vetor das contagens de palavras que aparecem nele. Dependendo do método de classificação, o vetor pode ser normalizado à unidade e dimensionado de modo que palavras comuns sejam menos importantes do que palavras raras, como na representação TF-IDF (BOULIS; OSTENDORF, 2005).

O TF-IDF foi desenvolvido a partir do IDF, proposto por (JONES, 1972) com a percepção heurística de que um termo de consulta que ocorre em muitos documentos não é um bom discriminador e deve receber um peso menor do que aquele que ocorre em poucos documentos. O TF-IDF é uma combinação de duas técnicas estatisticas, o TF (*Term Frequency*) e IDF (*Inverse Document Frequency*) suas respectivas equações são demonstradas a seguir.

A equação 1, representa o TF, que mede a frequência em que um termo ocorre em um documento. Ele é representado pela divisão entre a quantidade de vezes que um termo aparece

[2] http://143.107.183.175:22680/
[3] http://cohmetrix.com/

no documento (frequenciaT) e o total de termos que o documento possui (N).

$$tf = \frac{frequenciaT}{N} \tag{1}$$

Por sua vez, IDF mede a importância de um termo. Pois o TF pode dar importância para termos como preposições e artigos por serem termos muito comuns, aparecendo bastante no documento, mas possuem pouca importância. O IDF busca dar um peso maior para termos que ocorrem mais raramente. Pode ser representado pela equação 2, onde 'N' é a quantidade total de documentos, e 'n' é a quantidade de documentos que possui um determinado termo.

$$idf = log\left(\frac{N}{n}\right) \tag{2}$$

Por fim, uma medida razoável da importância de um termo pode ser obtida utilizando o produto da frequência do termo (TF) com a frequência inversa do documento (IDF) (SALTON; BUCKLEY, 1988). A fórmula resultante está representada na equação 3.

$$tfidf = tf * idf \tag{3}$$

A idéia básica do TF-IDF é da teoria da modelagem da linguagem de que os termos em um determinado documento podem ser divididos em duas categorias: vetores de palavras com importância e aquelas sem importância (ROBERTSON, 2004), ou seja, se um termo é relevante ou não com o tópico de um determinado documento. É um método simples, porém eficaz, para mapear um documento em um vetor de comprimento fixo.

2.2.3.6 Balanceamento de Dados

O problema de desbalanceamento de dados é muito comum no mundo real e também que esses dados estejam 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). Estas classes minoritárias também podem ser chamadas de classes raras (WEISS, 2004).

Algoritmos de classificação são bastante sensíveis ao desbalanceamento e tendem a valorizar as classes predominantes e a ignorar as classes de menor representação (PHUA et al., 2004).Segundo (CHAWLA et al., 2004), base de dados desbalanceadas prejudicam a acurácia dos modelos gerados pelos classificadores. Dessa forma, de acordo com (SPILIOPOULOS et al., 2010), o conjunto de treinamento deve esta balanceado em número, ou seja, as instâncias de

treinamento de todas as classes devem ser numericamente iguais, assim a classificação poderá alcançar uma maior eficiência. Corroborando com os autores, (BECKMANN, 2010) afirma que no processo de classificação, para que o modelo aprenda conhecimento novo representativo, os exemplos das classes precisam estar bem definidos e deve haver uma quantidade suficiente.

Dentre as técnicas existentes na literatura, destaca-se a oversampling. Essa técnica replica 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). Segundo o autor, considera-se que $|S| = |S_{pos}| + |S_{neg}|$, onde *S* representa um determinado conjunto de treinamento com *m* instancias, ou seja, $(|S|) = m$; e S_{pos} e S_{neg} correspondem aos conjuntos das classes positivas e negativas, respectivamente. Para o subconjunto $S_{pos'}$, é considerado os k-vizinhos mais próximos para cada instância $x_i \in S_{pos}$, para um determinado valor de k. Os k-vizinhos mais próximos são definidos como os k elementos de S_{pos} cuja distância euclidiana entre si e a instância x_i apresenta o menor valor. Para gerar uma nova instância sintética, é selecionado 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). Em resumo, $x_{new} = x_i + (y_i - x_i) * \delta$, onde x_i é uma instância da classe minoritária (positiva) em consideração, y_i é um dos seus k-vizinhos mais próximos de x_i e δ é o número aleatório (FERNÁNDEZ et al., 2018).

2.2.3.7 Técnicas e Algorítimos de Classificação

De forma geral, classificação é o processo de atribuir um rótulo de um conjunto predefinido a um objeto, por exemplo, classificando um filme de acordo com seu gênero. Entretanto, no contexto de mineração de dados, a classificação consiste em uma análise de dados em duas etapas, a primeira onde um modelo é gerado a partir de um conjunto de dados de treinamento usando algoritmos estatísticos e de aprendizado de máquina capazes de, em uma segunda etapa, prever qual a classe um novo dado até então desconhecido pertence (HAN et al., 2011)

Os problemas de classificação de texto foram amplamente estudados e abordados em muitas aplicações reais nas últimas décadas (JIANG et al., 2018). Por ser um tópico relevante em muitas áreas, a classificação tem sido objeto de intensa pesquisa nas últimas décadas, o que resultou na proposta de vários métodos para gerar, aprimorar e avaliar classificadores (SCHÜTZE

et al., 2008).

Segundo (ONAN, 2016), alguns métodos de *ensemble* estão se tornando populares em pesquisas de aprendizado de máquina como por exemplo, métodos de classificação por voto. Técnicas de classificação de votação, *boosting* e *bagging*, foram desenvolvidas com sucesso para a classificação de conjuntos de dados de documentos e textos (FARZI; BOLANDI, 2016).

Nas próximas seções iremos abordar os métodos de aprendizado em conjunto, *boosting* e *bagging* e também alguns algoritmos que implementam essas técnicas.

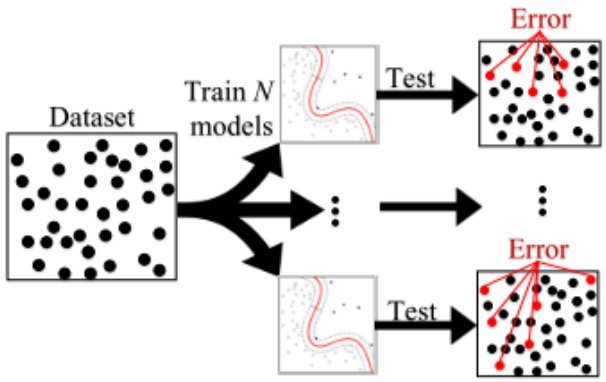
2.2.3.7.1 Bagging

O Bagging (*Bootstrap Aggregating*), é um método *ensemble* proposto por (BREIMAN, 1996) que visa criar um classificador robusto, com alto desempenho preditivo, combinando os classificadores treinados em diferentes conjuntos de treinamento.

No Bagging os classificadores são treinados de forma independente por diferentes conjuntos de treinamento através do método de inicialização. Para construí-los é necessário montar k conjuntos de treinamento idênticos e replicar esses dados de treinamento de forma aleatória para construir k redes independentes por re-amostragem com reposição. Em seguida, deve-se agregar as k redes através de um método de combinação apropriada, tal como a maioria de votos (BREIMAN, 1996). Para obter novos conjuntos de treinamento, é utilizada a amostragem aleatória simples com substituição. Este método produz a diversidade necessária para a aprendizagem do conjunto.

A Figura 2 mostra uma arquitetura simples da técnica de bagging que implementa N modelos.

Figura 2 – Arquitetura da técnica de bagging. (KOWSARI et al., 2019)



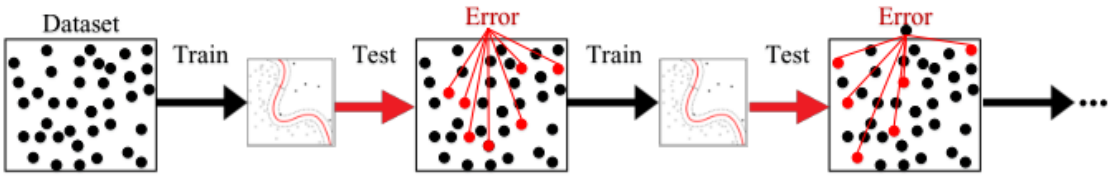
2.2.3.7.2 Boosting

A técnica de boosting foi desenvolvida por (FREUND, 1992) e introduzido pela primeira vez por (SCHAPIRE, 1990) como uma técnica para aumentar o desempenho de um algoritmo de aprendizado fraco.

No Boosting, de forma semelhante ao Bagging, cada classificador é treinado usando um conjunto de treinamento diferente. A diferença primordial em relação ao Bagging é que os conjuntos de dados, de re-amostragem, são construídos especificamente para gerar aprendizados complementares e a importância do voto é ponderado com base no desempenho de cada modelo, em vez de serem atribuídos pesos iguais para todos os votos. Basicamente, esse procedimento permite aumentar o desempenho de um limiar arbitrário simplesmente adicionando *learners* mais fracos. Segundo (LANTZ, 2013), Boosting é considerado uma das descobertas mais significativas em aprendizado de máquina.

A Figura 3 mostra o funcionamento de um algoritmo simples de boosting.

Figura 3 – Arquitetura da técnica de boosting. (KOWSARI et al., 2019)



A seguir serão apresentados alguns algoritmos que implementam uma dessas técnicas apresentadas.

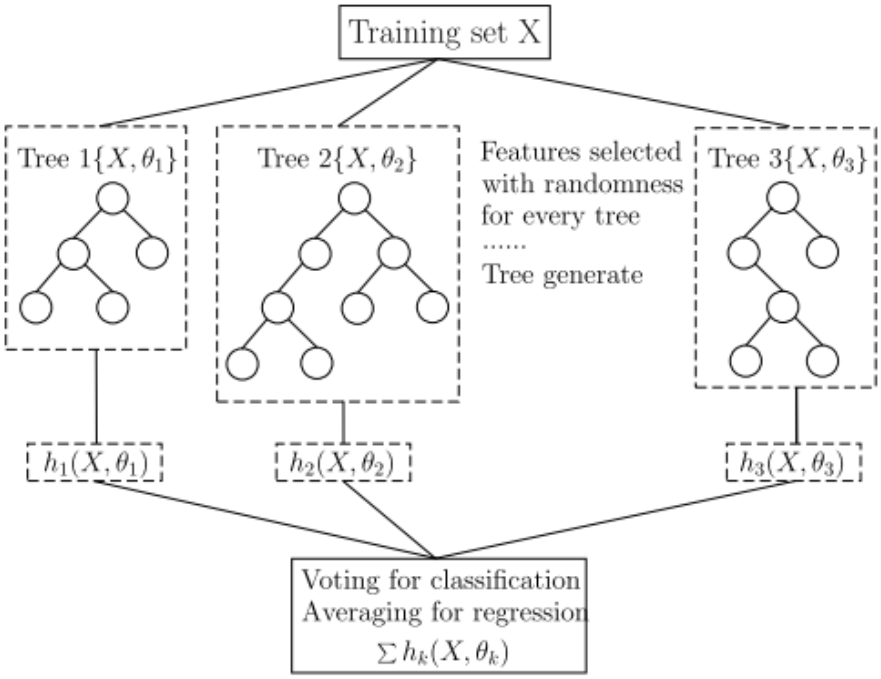
2.2.3.7.3 Random Forest

O *Random Forest* (Floresta Aleatória) é um algoritmo desenvolvido por (BREIMAN, 2001), baseado em um conjunto de árvores de decisão induzidas a partir de amostras *bootstrap* dos dados de treinamento. Esse algoritmo é composto por 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 em uma dada entrada X. As florestas aleatórias oferecem várias vantagens, a melhora da acurácia, em relação às árvores de decisão individuais e sua robustez à *overfitting* (BREIMAN, 2001). Lidam bem com dados mistos e ausentes (HASTIE et al., 2009), e naturalmente criam uma medida de importância variável e classificam-as pelo poder preditivo (BREIMAN, 2001).

Do ponto de vista prático, os RF são amplamente utilizados e exibem um desempenho extremamente alto, com apenas alguns parâmetros para ajustar (ZIEGLER; KÖNIG, 2014). Notando sua versatilidade, diversas pesquisas aplicaram com sucesso florestas aleatórias em várias tarefas de diferentes áreas disciplinas biotecnologia (BECHT et al., 2019), economia (VARIAN, 2014), ciência da computação (FERNÁNDEZ-DELGADO et al., 2014) e educação (AHMED; SADIQ, 2018).

Segundo (MENTCH; HOOKER, 2016), o algoritmo é construído em subamostras escolhidas aleatoriamente dos dados de treinamento e a previsão final é tomada como a média sobre os resultados individuais, formando assim um conjunto de árvores de decisão. De forma resumida, a classificação é definida com o maior número de votos. A Figura 4 demonstra o funcionamento simplificado do algoritmo.

Figura 4 – Random forest. (ZHANG et al., 2018)



Como mencionado anteriormente, o algoritmo de floresta aleatória utiliza o método *bagging* (*Bootstrap Aggregating*) para produzir amostras aleatórias de conjuntos de treinamento (amostras *bootstraps*) para cada árvore gerada. De acordo com (EFRON, 1992), amostragem *bootstrap* é uma técnica que faz amostragem com reposição, ou seja, a partir do conjunto de treinamento inicial, são escolhidos aleatoriamente instâncias para um novo subconjunto de treinamento. Essa variabilidade nos dados de treinamento induz diferentes construções de árvores de decisão (BREIMAN, 1996). A utilização deste método traz algumas vantagens como

a melhora na performance do algoritmo e a probabilidade de conseguir estimativas internas do erro de generalização do conjunto agregado de árvores, da força e correlação de uma árvore classificadora utilizando o método out-of-bag (OOB).

No método OOB, para um determinado 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 equivalem ao bootstrap de amostras que contenham c_i. Segundo (BREIMAN, 2001), Uma vantagem do método é que ele fornece estimativas de erro, OOB error, que dispensam o uso de um conjunto de teste, que corresponde à validação cruzada, porém realiza cálculos das estimativas internas durante o processo de treinamento. O erro OOB também é usado para calcular a importância da variável, que é um fator bastante decisivo para quando deseja-se selecionar as variáveis mais relevantes para o desenvolvimento do modelo (GENUER et al., 2010).

Diante desse contexto, (BREIMAN, 2001) propôs duas medidas de importância de variáveis, uma conhecida como *Mean Decrease in Gini (MDG)*, que mede a impureza de Gini e a outra denominada *Mean Decrease in Accuracy (MDA)* baseada na importância da permutação. Neste trabalho será aplicado o MDG, oriundo do classificador utilizado no experimento onde cada nó da árvore busca-se a divisão ótima segundo a impureza de Gini, ou seja, o quanto uma divisão é capaz de separar as amostras de um dado nó. Essa medida é calculada pela soma de todas as diminuições na impureza de Gini a cada divisão do nó, normalizada pelo número de árvores (BREIMAN, 2001)

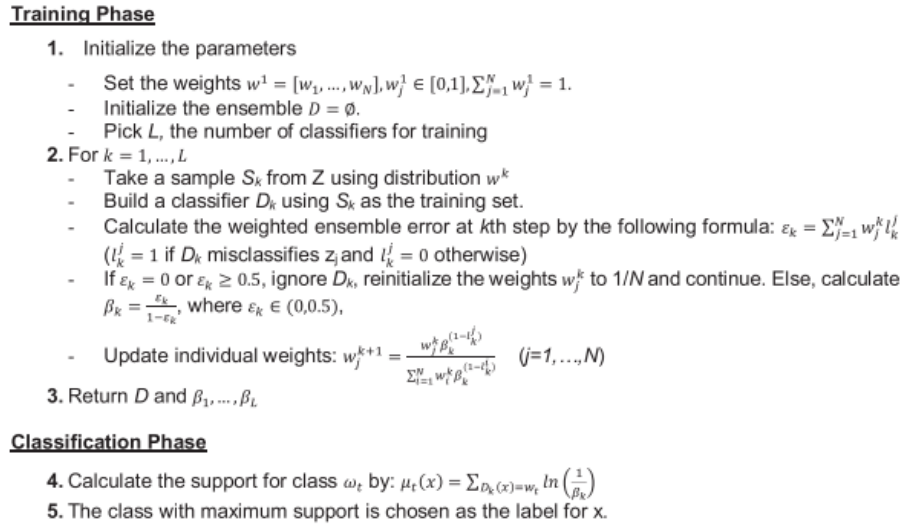
2.2.3.7.4 AdaBoost *(Adaptive Boosting*)

O algoritmo AdaBoost foi o primeiro algoritmo que implementou o método de boosting proposto por (SCHAPIRE, 1990), e continua sendo um dos mais amplamente utilizados e estudados, com aplicações em diversas áreas. Segundo (KUNCHEVA, 2014), este algoritmo é um método *ensemble* que aprimora o algoritmo por meio de um processo iterativo no qual o maior foco é dedicado aos padrões difíceis. O algoritmo agrega sequencialmente diversos modelos mais fracos, onde o *weak learner* subsequente leva em conta as predições do anterior, para então formar um classificador mais conciso.

De forma simples, primeiro todas as instâncias no conjunto de treinamento recebem o mesmo peso. Durante o processo, os valores dos pesos para instâncias classificadas incorretamente aumentam, por outro lado, os pesos para instâncias classificadas corretamente diminuem. Dessa maneira, o algoritmo mais fraco dedica mais iterações e classificadores as instâncias mais

difíceis de classificar (ROKACH, 2010).A estrutura geral do algoritmo é demonstrada na Figura 5.

Figura 5 – Estrutura geral do algoritmo Adabbost. (KUNCHEVA, 2014)



2.2.3.7.5 XGBoost (*Extreme Gradient Boosting*)

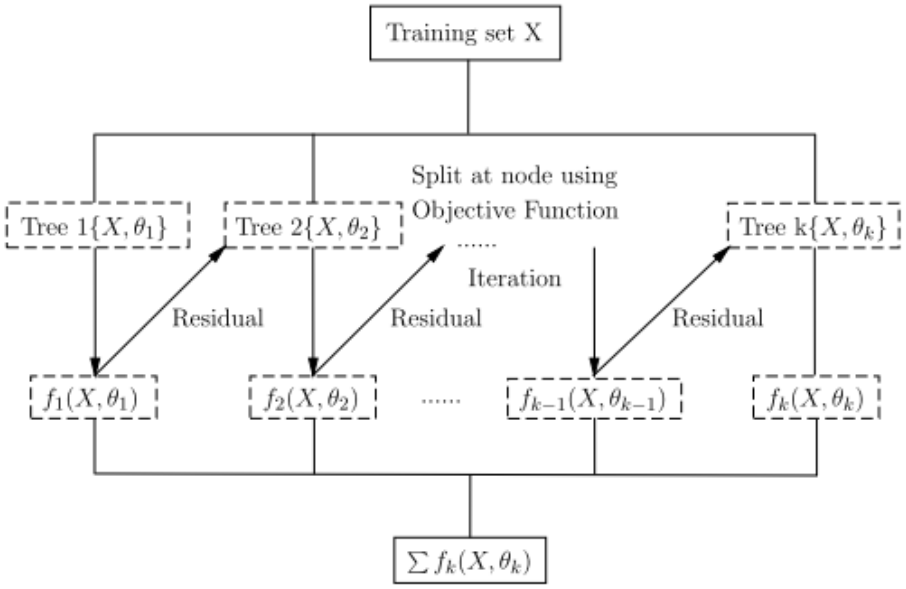
O XGBoost é um tipo de algoritmo de aprendizado de máquina baseado na técnica de Boosting. É uma biblioteca geral de aumento de gradiente desenvolvida pelo Dr. Chen da Universidade de Washington. Seu objetivo é obter resultados de previsão precisos com eficiência por meio do cálculo iterativo do classificador de árvore de decisão CART. O XGBoost é um modelo de otimização que aprimora o algoritmo de aumento de gradiente existente, que combina modelo linear e modelo de aprendizado em árvore (CHEN; GUESTRIN, 2016). De acordo com (NATEKIN; KNOLL, 2013), o aumento do gradiente é construído de forma sequencial. De fato, um novo *learner* fraco é construído para ser correlacionado ao máximo com o gradiente negativo da função de perda associado a todo o conjunto para cada iteração.

Para evitar o *overfitting*, o XGBoost reduz a variação do modelo adicionando termos regulares à função de custo, controlando assim a complexidade do modelo. A característica mais importante desse algoritmo é que ele pode usar automaticamente multithreads da CPU para computação paralela melhorando a precisão, consequentemente otimizando o algoritmo. O algoritmo tem sido amplamente utilizado nos campos de inteligência artificial, análise de dados, mineração de dados e estatística. É usado para resolver vários problemas práticos e alcançou uma precisão muito alta (CHEN; GUESTRIN, 2016).

O XGBoost pertence ao grupo de algoritmos de aprendizado de árvores amplamente utilizados (HE et al., 2014). Entre as 29 soluções vencedoras na competição de aprendizado

de máquina Kaggle em 2015, o XGBoost foi o método mais popular em que 17 soluções o utilizaram (ZHANG et al., 2018). O fator mais importante por trás do sucesso do XGBoost é sua escalabilidade em todos os cenários. O sistema roda mais de dez vezes mais rapidamente do que as soluções populares existentes em uma única máquina e pode escalar para bilhões de exemplos em configurações distribuídas ou com pouca memória (CHEN; GUESTRIN, 2016). A Figura 6 exemplifica o fluxo do algoritmo.

Figura 6 – Fluxograma do XGBoost. (ZHANG et al., 2018)



2.2.4 Avaliação e Interpretação dos Resultados

Esta etapa, também conhecida como Pós-processamento de dados, refere-se à verificação da eficiência da aplicação do algoritmo de classificação. Segundo (SEBASTIANI, 2002), a avaliação experimental de um classificador, em geral, mede sua eficácia, em outras palavras, sua capacidade de tomar as decisões corretas de classificação. De acordo (CAMILO; SILVA, 2009), é nesta etapa onde testes e validações, com o objetivo de obter a confiabilidade nos modelos, devem ser executados (*cross validation, suplied test set, use training set, percentage split*). E também, deve-se obter indicadores que permitam a análise dos resultados (matriz de confusão, *recall*, estatística kappa, acurácia, precisão, F-measure, dentre outros).

Nesta pesquisa, foram utilizadas a técnica de validação cruzada e as medidas: acurácia e kappa, pois, nas pesquisas educacionais essas métricas são bastante utilizadas para medir o desempenho de algoritmos de aprendizado de máquina supervisionado (MESSICK, 1995) e (KOVANOVIC et al., 2014b). Além disso os trabalho relacionados também as utilizam, dessa forma pode-se realizar comparações congruentes.

2.2.4.1 Validaçao Cruzada

Desde a introdução da validação cruzada de saída única (*leave-one-out cross-validation* - LOOCV) por (STONE, 1974) , a validação cruzada (CV) tem sido uma das técnicas mais populares para a seleção de modelos. De acordo com (KOHAVI et al., 1995), validação cruzada (*cross validation*), trata-se de uma técnica estatística que divide ma 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.

Em resumo a técnica é executada da seguinte maneira: um conjunto de dados é dividido de forma aleatoria em K subconjuntos (*folds*), ambos exclusivos, cujo tamanho é aproximadamente igual a n/K, onde n é o tamanho do conjunto de dados. Dessa forma, são realizados K iterações, e em cada uma, um subconjunto diferente é escolhido para o teste e os K-1 subconjuntos restantes são indicados para o treinamento. A medida de eficiência é a média das medidas de eficiência calculadas para cada um dos subconjuntos (JUNG, 2018).

A vantagem de utilizar-se desta técnica é que todos os dados do conjunto sao são utilizados tanto para treinamento quanto para teste. Então, de acordo com (PENG et al., 2004), a maneira como a divisão do conjunto foi realizada interfere menos no resultado final visto que, qualquer dado será usado exatamente uma vez para teste e $k1$ vezes para treinamento.

2.2.4.2 Acurácia

Acurácia é uma medida de desempenho relacionada com os percentuais dos acertos e erros, realizados pelo modelo de classificação gerado, perante novos exemplos. É uma das medidas mais comuns para avaliar um modelo de classificação (WITTEN; HELL, 2011). Ela pode ser calculada a partir de uma estrutura denominada matriz de confusão.

A matriz de confusão, Tabela 4, contém 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.

Tabela 4 – Matriz de Confusão

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

Fonte: Elaborado pelo autor.

Onde:

- **Verdadeiro Positivo (TP)** - quantidade de instâncias da classe positiva preditas corretamente;
- **Verdadeiro Negativo (TN)** - quantidade de instâncias da classe negativa preditas corretamente;
- **Falso Positivo (FP)** - quantidade de instâncias da classe negativa preditas como sendo da classe positiva;
- **Falso Positivo (FN)** - quantidade de instâncias da classe positiva preditas como sendo da classe negativa;

No contexto do presente estudo, classificação de textos, a acurácia é a quantidade dos documentos classificados corretamente (positivos e negativos) sobre o número total de documentos (HOTHO et al., 2005) e é definida pela equação 4 abaixo:

$$\mathbf{Acuracia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

2.2.4.3 Estatística Kappa

O coeficiente estatístico índice Kappa é uma medida de concordância em escalas nominais (COHEN, 1960). 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, ou seja, o quanto os dois estão de acordo quanto à classificação (CARLETTA, 1996).

Há uma escala com intervalo de 0 à 1, onde o valor máximo 1 indica total concordância, enquanto os valores próximos e até abaixo de 0, indicam nenhuma concordância.(LANDIS; KOCH, 1977) associam valores de Kappa à qualidade da classificação de acordo com a Tabela 5 abaixo:

Tabela 5 – Representação dos índices de Kappa

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

Fonte: Elaborado pelo autor (adaptado de (LANDIS; KOCH, 1977)

O índice Kappa (k) é calculado de acordo com a Equação 5 abaixo:

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{5}$$

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

3 TRABALHOS CORRELATOS

Neste capítulo serão apresentados trabalhos que possuem uma certa similaridade em relação a abordagem deste trabalho. Eles estarão divididos duas seções: trabalhos com aplicações para a língua inglesa e trabalhos com aplicações para a língua portuguesa. Por fim, será apresentada uma tabela com um resumo das principais características de cada estudo e quais os diferencias da proposta deste trabalho em relação aos demais.

3.1 APLICAÇÕES PARA O INGLÊS

De forma geral os três primeiros trabalhos apresentados nesta seção, (SAUDE et al., 2012), (KOVANOVIC et al., 2014b) e (ROLIM et al., 2019 no prelo) utilizam-se de técnicas manuais de codificação das categorias da presença social presentes na literatura e abordagens em contagem de frases e palavras. Enquanto nos trabalhos de (TANIGUCHI et al., 2019) e (FERREIRA et al., 2020) os autores tem como objetivo principal a classificação automática das categorias da presença social, através de ferramentas e técnicas de mineração de texto. Cada um deles será detalhado a seguir.

Em (SAUDE et al., 2012), tem como foco as três categorias e a densidade da presença social, examinando 826 respostas de 6 fóruns do MyLine, um portal de aprendizado de inglês para estudantes de universidades públicas da Malásia, para avaliar se o ambiente desses fóruns é aquele em que a aprendizagem profunda e significativa provavelmente ocorrerá. Foi utilizado o esquema de codificação que divide a presença social em categorias e indicadores e o cálculo de densidade de presença social (*Social Presence Density calculation - SPD*), que mede o número de ocorrências desses indicadores por 1.000 palavras, ambos desenvolvidos por (ROURKE et al., 1999). Cada uma das 826 mensagens foram codificadas manualmente, de acordo com os indicadores propostos, compilados e quantificados de acordo com as três categorias da presença social e em seguida foram submetidos ao cálculo de densidade de presença social. Por fim, os autores concluíram que as repostas interativas são as mais presentes com cerca de 48% do total de indicadores identificados, seguidos de 41% e 11% de ocorrências das categorias afetiva e coesiva, respectivamente. E que o SPD parece ser uma ferramenta valiosa para calcular a presença social, pois foi capaz de diferenciar aqueles transcrições que pareciam ter altos níveis de sociabilidade daqueles com baixa sociabilidade baseado nas impressões intuitivas do autor ao ler as mensagens.

No trabalho de (KOVANOVIC et al., 2014b), pretendia-se verificar a associação entre o

capital social, um dos aspectos centrais no estudo das redes sociais (BURT, 2001), e a presença social em comunidades de investigação, onde capital social é um valor resultante da ocupação de uma posição particularmente vantajosa dentro de uma rede social (BURT et al., 2002). Para isso os autores utilizaram um conjunto de dados oriundos de um fórum de discussão de uma universidade pública do Canadá com 1.747 mensagens produzidas por 81 estudantes. As mensagens foram codificadas por 2 anotadores, orientados pelo esquema de codificação proposto por (ROURKE et al., 1999) composto de doze indicadores, desses, três (Continuar uma conversa, Expressar apreço/concordância e Vocativos) foram excluídos pois continham um grande número de mensagens. Para medir o capital social dos alunos, foram extraídos gráficos das redes sociais dos alunos das interações nos fóruns de discussão e a partir deles foram extraídas três medidas de centralidade de rede, *Betweenness centrality, Degree centrality e Closeness centrality*, mais frequentemente utilizadas para o estudo das redes sociais educacionais (CAROLAN, 2013). Por fim para investigar a relação entre as rês categorias de presença social e o capital social foram realizadas análises de regressão linear múltipla para cada uma das medidas de centralidade de rede extraídas. Os resultados indicam que as três categorias da presença social predizem significativamente as medidas de centralidade da rede comumente usadas para medir o capital social.

Em (ROLIM et al., 2019 no prelo), é apresentada uma abordagem que utiliza modelagem de tópicos e análise de rede epistêmica para investigar como a presença social dos alunos é expressa em diferentes tópicos do curso, a partir de mensagens de fóruns de discussão online e como esse método pode ser adotado para examinar como a presença social dos alunos mudou devido a uma intervenção instrucional. A base de dados e o esquema de codificação das categorias da presença social foram os mesmos utilizados no trabalho mencionado anteriormente em (KOVANOVIC et al., 2014b). Foi realizada uma etapa de pré-processamento do texto, aplicando técnicas de PLN como a remoção de *sopwords* e *stemming* e em seguida foi aplicada a técnica de Latent Dirichlet Allocation (LDA), uma técnica de modelagem probabilística de tópicos amplamente usada para extração de tópicos latentes em corpus textuais (BLEI et al., 2003). Para verificar a relação entre presença social e os tópicos do curso, foi utilizada a Análise de Rede Epistêmica (*Epistemic Network Analysis - ENA*) (SHAFFER et al., 2016) que recebeu como codificação as três categorias da presença social e os 15 tópicos do curso extraídos através da LDA. Os resultados mostraram que a verificação da presença social em relação aos diferentes tópicos do curso revela dinâmicas sociais interessantes que foram visualizadas de forma mais fácil através da observação da rede, como por exemplo, que presença social não era distribuída

igualmente entre os diferentes tópicos do curso, com variações substanciais em sua intensidade, dependendo do tópico.

No trabalho realizado por (TANIGUCHI et al., 2019), foram desenvolvidos modelos para cada indicador das presenças cognitiva e social, um total de 26 indicadores, 14 da presença cognitiva e 12 da presença social. Dois tipos de algoritmos foram utilizados, o Random Forest e o XGBoost que são classificadores baseados em árvores de decisão, por terem bom desempenho nas tarefas de classificação de texto. Segundo os autores o experimento visa comparar métodos de tokenização, não o desempenho dos classificadores por isso não foi investido esforços no ajustes de seus hiperparâmetros. Foram comparados doi métodos de tokenização, o MeCab (KUDO et al., 2004) e SentencePiece (KUDO; RICHARDSON, 2018). O MeCab é um analisador morfológico popular para tokenizar frases em japonês, os parâmetros de seu modelo são pré-treinados a partir de um grande corpus de texto e fornecidos como um dicionário. O SentencePiece é um método relativamente novo, esse algoritmo não precisa de um conjunto de dados pré-tokenizado e, portanto, é essencialmente independente da linguagem. Antes do processo de tokenização houve a normalização dos textos, por exemplo a conversão de maiúsculas em minúsculas e normalização Unicode. Mas após a tokenização a etapa de pré-processamento de linguagem natural comumente utilizada foi omitida, como por exemplo a remoção de *stopwords*, pois desejava-se evitar ao máximo o pré-processamento dependente de idioma. Para representação do texto os *tokens* foram submetidos a técnica TF-IDF e em seguida submetidos aos classificadores utilizando-se da técnica de validação cruzada e como métrica de avaliação foi empregada a curva ROC. Apesar de alguns indicadores possuirem valores da AUC superiores a 0,9, os resultados sugerem que é fundamental uma análise mais avançada das mensagens do fórum, em vez de utilizar apenas os recursos tradicionais como o TF-IDF.

Por fim, em (FERREIRA et al., 2020), é proposta uma abordagem para classificar automaticamente as mensagens de discussão online escritas em inglês, de acordo com as categorias da presença social. Para alcançar esse objetivo foi utilizada a mesma base de dados e esquema de codificação utilizado em (KOVANOVIC et al., 2014b) onde dois codificadores especialistas codificaram o conjunto de dados de acordo com os 12 indicadores da presença social com um percentual de concordância entre os dois de 84%, e um terceiro codificador resolveu as discordâncias. Assim como em (KOVANOVIC et al., 2014b) foram removidos três indicadores (Continuar uma conversa, Expressar apreço/concordância e Vocativos) pois continham um grande número de mensagens. Além das características tradicionais de mineração de texto, como frequência de palavras, o trabalho combinou ferramentas linguísticas LIWC e Coh-Metrix para extrair

indicações de presença social através do texto. Após a extração de características as mensagens passaram por um pré-processamento onde foram aplicadas as técnicas de remoção de *stopwords*, *stemming* e utilizado o algoritmo SMOTE para resolver o problema de desbalanceamento dos dados. A base de dados foi dividida em conjunto de treino e teste, em seguida submetida a um processo de otimização utilizando o algoritmo Random Forest com o objetivo de encontrar bons valores para os parâmetros *n_estimator* e *max_features*, resultando em três classificadores, um para cada categoria da presença social. As métricas utilizadas para avaliar o modelo foram acurácia e kappa, onde o melhor modelo atingiu 0,95 e 0,88 respectivamente. Além disso o trabalho também apresentou as características mais relevantes para os classificadores durante o processo de classificação de cada categoria.

3.2 APLICAÇÕES PARA O PORTUGUÊS

Existem alguns trabalhos que aplicam os conceitos da presença social na língua portuguesa mas em relação a identificação automática da forma com que este trabalho é proposto não foi identificado nenhum. A seguir serão apresentados alguns trabalhos assim como uma tabela comparativa entre eles e serão apresentados os diferenciais desta pesquisa em relação aos mesmos.

No trabalho de (WIVES et al., 2010), foi apresentado um estudo das interações discursivas realizadas em um chat, que teve duração de aproximadamente uma hora e participação de dez alunos e um tutor, em contexto educacional de forma a verificar as três categorias da presença social dos sujeitos em ambiente de aprendizagem a distância utilizando a análise textual convencional e automática. Para avaliar o grau de Presença Social nesse estudo, foram utilizados os critérios indicados por (ROURKE et al., 2001): aspectos afetivos, interativos e coesivos. Basicamente os autores utilizaram-se de uma ferramenta de mineração de texto denominada Eurekha (WIVES, 1999), submeteram o texto extraído do chat e aplicaram a técnica de tokenização e remoção de stopwords e através de recursos gráficos da ferramenta analisaram a ocorrência e relevância das palavras. Concluíram que o software permite o processamento de parte das pistas indicadoras de presença social, e que o levantamento do número de ocorrências dos termos pode fornecer dados quantitativos auxiliares na interpretação dos dados mas ao mesmo tempo o processamento automático realizado não contempla algumas das pistas indicativas das categorias da presença social como, por exemplo, udo de emoticons e uso expressivo de sinais de pontuação.

Em (BASTOS et al., 2010), foi realizado um estudo de caso feito em chats e fóruns no Curso de Informática Instrumental para Professores do Ensino Básico oferecido pelo convênio UAB-II- UFRGS. O objetivo era verificar como os usuários expressam sua presença social em suas interações, nas duas ferramentas, com base no mapeamento de pistas discursivas das três categorias da presença social indicadas por (ROURKE et al., 2001) acrescidas do subcampo "Força" (sinalizadora de intensidade) encontrado no Sistema de Avaliatividade Linguística (*Appraisal System*) de (MARTIN; WHITE, 2003). Através desse mapeamento os autores identificaram e extraíram as pistas textuais referentes aos indicadores e realizaram a análise comparativa entre o número de ocorrências em cada uma das ferramentas, chat e fórum.

Em (SILVA, 2011), foi proposta uma abordagem automática para encontrar o grau de presença social do aluno, desenvolvendo-se uma ferramenta e realizado-se quatro experimentos, comparados com a análise feita por um perito. A ferramenta consistia em: a) um conversor HTML para XML, utilizado para converter os arquivos extraídos pelo professor dos fóruns/chat; b) Construtor de Categorias: que permite ao professor criar e editar categorias, subcategorias de presença social e especificar suas respectivas pistas textuais; e c) Analisador: baseado no arquivo de categorias, analisa o arquivo convertido e retorna a presença social dos alunos. Para os dois experimentos foram utilizados dois fóruns, um composto por 66 tópicos e 142 postagens e o outro com 74 tópicos e 7.309 postagens. Os melhores resultados, onde houve uma taxa de acerto entre 89% a 94%, foram no experimento onde primeiramente o perito analisou de forma manual cada fórum e em seguida cadastrou as pistas encontradas. Mas no experimento onde o perito cadastrou as pistas de forma mais genérica de acordo com as classes cadastradas no sistema os resultados ficaram entre 5% a 58,90%. Sendo assim fica clara, nessa abordagem, que é essencial a presença do perito para cadastrar os padrões a serem procurados pelo software.

Em (PEREZ et al., 2012), tinha como objetivo analisar características da presença social em um curso de capacitação docente, na modalidade à distância, mais precisamente na ferramenta de fórum de discussão que gerou 303 postagens. Para realizar a análise foi utilizada a Matriz Padrão de escala da presença social, desenvolvida por (KIM, 2011) construída com base em quatro fatores: ligação afetiva, senso de comunidade, comunidade aberta e atenção e apoio mútuo. Especificamente para reconhecer características relativas à presença social foram identificadas em cada um dos quatro fatores 20 questões dentro de cada um dos variados contextos no reconhecimento das impressões dos alunos sobre o curso e a proposta de avaliar se foram estabelecidas as ligações de presença social esperadas. A pesquisa permitiu medir a presença social com discussões baseadas na matriz utilizada, o que possibilitou corroborar no quanto a

presença social é importante para entender a percepção do aluno,tanto no seu sentimento de pertencimento ao curso quanto na sua aprendizagem concreta e satisfação com o curso.

No trabalho de (GOMES; PESSOA, 2012) foram utilizadas a presença social e cognitiva para analisar a experiência de aprendizagem na área da supervisão pedagógica de nove indivíduos, através das publicações nas funcionalidades Mural e Discussões de um grupo de estudo criado na rede social Facebook. A metodologia de análise para categorizar as mensagens combinou as categorias e indicadores das presenças. Dessa forma, analisaram separadamente as mensagens do Mural e das Discussões. Por meio dessas análises, foram definidos os indicadores de cada uma das categorias de cada presença e de cada seção (Mural e Discussões), logo após, foi realizada a classificação das mensagens segundo esses indicadores.

Em (NETO et al., 2018), foi apresentado um método que permite a análise automatizada das mensagens trocadas em fóruns online de ensino a distância escritas em português brasileiro com foco na classificação das mensagens segundo os níveis da presença cognitiva. Para isso o autor utilizou um corpus com um total de 2.234 mensagens e realizou a extração de 127 características de diferentes recursos,em especial o LIWC e Coh-Metrix, 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 da presença cognitiva. O modelo desenvolvido atingiu 76% de acurácia e o de 0,55, o que representa uma concordância moderada, e está acima do nível de puro acaso. O trabalho também forneceu uma análise detalhada da relevância das características propostas, baseadas principalmente no Coh-Metrix e LIWC, 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.

A Tabela 6 apresenta um comparativo dos trabalhos relacionados com a abordagem proposta. Ela está organizada da seguinte forma: Trabalhos - referência da literatura analisada; Presenças - quais as presenças do CoI são abordadas; Bases de Dados - origem do conjunto de dados utilizado; Categorias da Presença Social (PS) - se utilizam as categorias da presença social que foram abordadas, em caso de uso; Ferramentas e Técnicas de Mineração de Texto (MT) - quais as principais ferramentas e técnicas de mineração de texto que foram utilizadas, em caso de uso; Identificação da Presença Social (PS) - como se deu a identificação da presença social; Idioma - qual o idioma utilizado.

Tabela 6 – Trabalhos Correlatos

Trabalho	Presenças	Base de Dados	Categ. da PS	Ferramentas e Técnicas de MT	Ident. PS	Idioma
SAUDE et al.,2012	Social	Fórum de discussões	Sim	Não	Manual	Inglês
KOVANOVIC et al.,2014b	Social	Fórum de discussões	Sim	Não	Manual	Inglês
ROLIM et al., 2019	Social	Fórum de discussões	Sim	Latent Dirichlet Allocation (LDA); stopwords; stemming	Manual	Inglês
TANIGUCHI et al., 2019	Social e Cognitiva	Chat	Sim	TF-IDF; RandomForest; XGBoost	Autom.	Inglês
FERREIRA et al., 2020	Social	Fórum de discussões	Sim	LIWC;Coh-Metrix; BoW; SMOTE; validação cruzada; Random Forest	Autom.	Inglês
WIVES et al., 2010	Social	Chat	Sim	stopwords e número de ocorrência de palavras	Manual	Português
BASTOS et al., 2010	Social	Chat e Fórum de discussões	Sim	Não	Manual	Português
SILVA, 2011	Social	Chat e Fórum de discussões	Sim	Não	Autom.	Português
PEREZ et al., 2012	Social	Fórum de discussões	Não	Não	Manual	Português
GOMES; PESSOA, 2012	Social e Cognitiva	Rede Social (Facebook)	Sim	Não	Manual	Português
NETO, 2018	Cognitiva	Fórum de discussões	Não	LIWC;Coh-Metrix; Word Embedding; Entidade Nomeada; contexto de discussão;SMOTE; validação cruzada; RandomForest	Autom.	Português
Proposta da Dissertação	Social	Fórum de discussões	Sim	LIWC;Coh-Metrix; BoW; SMOTE;validação cruzada; Random Forest; AdaBoost;XGBoost	Autom.	Português

Fonte: Elaborado pelo autor.

De acordo com o que foi apresentado na Tabela 6, observando-se principalmente os trabalhos com aplicações para o português conseguimos identificar que a proposta deste trabalho utiliza-se de diversas ferramentas e técnicas de mineração de textos com bom desempenho na

literatura diferenciando-a das demais apresentadas. Apesar de (WIVES et al., 2010) aplicar algumas técnicas de MT, elas são aplicadas apenas no pré-processamento do texto, por exemplo remoção de *stopwords*, e a identificação da presença social é manual. Outro ponto que diferencia este trabalho dos demais é em relação a identificação da presença social, onde é proposta a automatização da classificação das postagens segundo as categorias da presença através de mensagens de textos geradas pelos alunos.

Como foi apresentado nesta seção, os trabalhos analisados com aplicação na língua portuguesa realizaram esse processo manualmente, tornando-o uma tarefa difícil de se realizar, principalmente em contextos maiores. Apesar de que em (SILVA, 2011), utilize-se de uma abordagem automática, ela não faz uso de técnicas de mineração de texto e o próprio autor chega a conclusão de que fica clara a importância do perito para auxiliar o software a obter um bom desempenho e que a ferramenta possui uma limitação, pois é baseada em lexemas fixos, sem dispor de um dicionário morfológico ou regras mais complexas que envolvam um complexo analisador sintático da língua natural. Vale ressaltar que, em geral, as mensagens desses estudos foram classificadas com o objetivo de avaliar o grau da presença social, do aluno ou do grupo, e não com o foco de se automatizar a etapa de identificação. É importante destacar que existe um trabalho desenvolvido por (NETO et al., 2018), que possui maior semelhança com a abordagem proposta nesta dissertação mas com o foco na identificação automática presença cognitiva. Além da diferença da presença estudada, outro diferencial importante do método apresentado neste trabalho é que foram utilizados três tipos de algoritmos, bastante utilizados na literatura, com o objetivo de obter os melhores resultados. Não apenas o Random Forest foi utilizado mas também o AdaBoost e o XGBoost que aprensetam bons resultados como citado na seção 2.2.3.7.

Também, um dos objetivos desse trabalho é realizar uma análise das características mais relevantes relacionadas a presença social em duas turmas de cursos EaD de contextos diferentes. Assim, este trabalho apresentará um método para automatizar a identificação da presença social, através de técnicas de Mineração de Texto e analisará perspectivas da presença em dois contextos distintos, conforme será detalhado nas próximas seções.

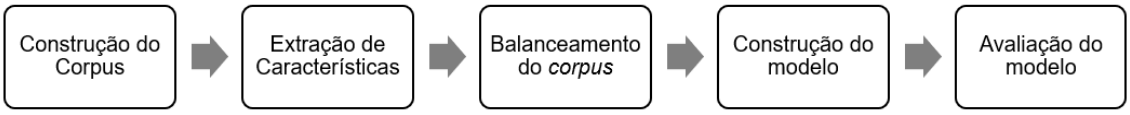
4 METODOLOGIA ADOTADA

Este capítulo descreve a metodologia utilizada para automatizar a análise da presença social em mensagens de fóruns de discussão assíncronas escritas em português, e está organizado em cinco seções: na seção 4.1 são apresentadas as etapas realizadas; a seção 4.2 contém a descrição do *corpus* utilizado nesta pesquisa; a seção 4.3 mostra as características extraídas; na seção 4.4 explica o balanceamento realizado no *corpus*; e por fim, na seção 4.5 demonstra a construção, otimização dos parâmetros e a avaliação do modelo proposto.

4.1 ETAPAS DA METODOLOGIA

A Figura 7, apresenta as etapas realizadas para a construção do modelo proposto.

Figura 7 – Etapas da metodologia



- **Construção do corpus -** processo de coleta das mensagens provenientes de fóruns online, bem como a construção do corpus e sua anotação, segundo as fases da presença social;
- **Extração de Características -** submissão das mensagens do corpus nas ferramentas utilizadas na pesquisa para extração das características utilizadas na construção do modelo;
- **Balanceamento do corpus -** balanceamento do corpus para treinamento e avaliação do modelo;
- **Construção do modelo -** seleção, otimização e treinamento 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

Neste trabalho foram utilizadas as mensagens extraídas de fóruns de discussão, provenientes de dois cursos totalmente online, de uma universidade pública brasileira. Ambos os fóruns tinham como objetivo promover discussões sobre um tema proposto pelo professor, onde

a participação do aluno representava 20% da nota final do curso. Em sua maioria, as discussões foram do tipo perguntas e respostas, ou seja, o professor iniciava o fórum com uma pergunta e os alunos deveriam responde-las deixando suas contribuições, seja respondendo a pergunta inicial e/ou novas perguntas relacionadas ao tema proposto pelo professor.

A primeira base, denominada pelo autor de BioBase, foi gerada com as mensagens enviadas em fóruns de discussão de um curso de graduação em Biologia. Foram extraídas 1.500 mensagens, produzidas por 215 alunos (64 homens e 151 mulheres) durante quatro semanas de curso, conforme descrito na Tabela 7 a seguir.

Tabela 7 – Temas do curso por semana (BioBase)

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%)

Fonte: Elaborado pelo autor.

A segunda base de dados, denominada pelo autor de TecBase, foi gerada através de mensagens retiradas de fóruns de discussão de um curso de graduação em Tecnologia. Foram extraídas 734 mensagens de discussão produzidas por 216 alunos (169 homens e 57 mulheres) durante três semanas de curso, como especificado na Tabela 8.

Tabela 8 – Temas do curso por semana (TecBase)

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%)

A Tabela 9 abaixo, mostra um resumo das principais informações relacionadas a alunos e mensagens nos fóruns. Em relação a quantidade média de mensagens e palavras por mensagem, na BioBase, cada aluno produziu 6,97 mensagens, com cerca de 89 palavras cada uma. Na TecBase, foram geradas 3,39 mensagens por aluno, contendo cerca de 113 palavras em cada uma. Existe uma predominância de alunos do gênero masculino na TecBase, cerca de 75%, enquanto na BioBase ocorre o inverso onde cerca de 70% dos alunos são do gênero feminino. Vale ressaltar que devido a pouca quantidade de mensagens, foram utilizadas todas as mensagens

presentes nas bases, incluindo a dos professores/tutores onde na BioBase esse tipo de mensagem corresponde a 25% e na TecBase 6%.

Tabela 9 – Resumo das bases

	BioBase	**TecBase**
Número de alunos (homens)	64	169
Número de alunos (mulheres)	151	57
Quantidade total de mensagens	1.500	734
Quantidade de mensagens do professor/tutor	386	44
Quantidade média de mensagens por aluno	6,97	3,39
Quantidade média de palavras por mensagem	89	113

Fonte: Elaborado pelo autor.

Ao unir as duas bases, foi gerado um corpus contendo 2.234 mensagens, que foi anotado por dois codificadores independentes, ambos conhecedores dos conceitos referentes a presença social e com conhecimento sobre o processo de análise de conteúdo, que analisaram as mensagens em separado, segundo os indicadores das três categorias da presença social (Afetiva, Interativa e Coesiva).

Para o processo de codificação escolheu-se a mensagem como unidade de análise, conforme indicado por (GARRISON et al., 2001). Assim os codificadores anotaram o conjunto de dados levando em consideração os 12 indicadores da presença social assim como foi feito em (KOVANOVIC et al., 2014b). Cada mensagem do conjunto de dados para cada indicador recebeu o valor "um"(possui o indicador) ou o valor "zero"(não possui o indicador). Assim como em (KOVANOVIC et al., 2014b), três indicadores (I1, I5 e C1) foram removidos pois continham um grande número de mensagens. A Tabela 10 mostra o resultado da anotação da base por indicador.

Tabela 10 – Anotação dos Indicadores da Presença Social

Categoria	**Código**	**Indicador**	**Quantidade**	**Concordância (%**
Afetiva	A1	1. Expressar emoções	190 (8.50%	84.29%
	A2	2. Uso de humor	3 (0.13%	99.06%
	A3	3. Informações sobre si	18 (0.81%	98.25%
Interativa	I1	4. Continuar uma conversa	2.224 (99.55%	100.00%
	I2	5. Citar mensagens de outros	5 (0.22%	98.39%
	I3	6. Fazer referência às mensagens de outros	920 (41.18%	87.65%
	I4	7. Fazer perguntas	158 (7.07%	86.57%
	I5	8. Saudar, expressar apreço	421 (18.85%	82.41%
	I6	9. Concordar	215 (9.62%	88.99%

Coesiva	C1	10. Vocativos	578 (25.87%	77.57%
	C2	11. Refere-se ao grupo usando pronomes	8 (0.36%	97.36%
	C3	12. Saudações	267 (11.95%	89.26%

Por fim, como o objetivo deste trabalho foi construir classificadores binários para cada categoria da presença social, as categorias foram compostas pelos indicadores anotados. Para que uma mensagem seja classificada como positiva (1), ela deve ter ao menos um indicador da respectiva categoria anotado com o valor "um". Por exemplo, se uma mensagem continha os indicadores A1 = 0, A2 = 0 e A3= 1 , então ela era considerada positiva para a categoria afetiva.

A Tabela 11 apresenta mensagens retiradas do corpus e suas respectivas classes, para exemplificar o resultado da anotação.

Tabela 11 – Exemplos de mensagens e suas respectivas anotações

Indicador	**Mensagem**
A1	QUANDO COMEÇA E QUANDO TERMINA A VIDA DE UM SER VIVO? Apesar de ser uma pergunta simples e bem clara, é difícil de ser especificada, e não se é certo falar que realmente ocorre esse processo igualmente...
A2	Clonagem de órgão sim, do ser vivo não... fila de espera aí vou eu.... kkkkk
A3	Recentemente li uma matéria na Revista Galileu em que mostra o desenvolvimento do microscópio nos EUA onde desenvolveram um microscópio 4D, em que cientistas foram capazes de observar pela primeira vez, o movimento das nanoestruturas do DNA.
I2	Os últimos anos têm sido marcados pelo desenvolvimento acelerado das tecnologias da informação e da comunicação na nossa sociedade, causando um forte impacto em todas as áreas da atividade humana.
I3	Parabéns pela colocação "Fulana"!
I4	Seria nas lentes ópticas em geral?
I6	Concordo com você "Fulana", o avanço da tecnologia é realmente muito eficiente no mundo modernizado em que vivemos.
C2	Parabéns a todos os colegas do curso, pois a cada postagem do fórum adquirimos novos conhecimentos, enriquecendo ainda mais nosso entendimento. Desejo a todos um bom estudo.
C3	Bom dia! O desenvolvimento tecnológico se tornou uma grande aliada no processo ensino/aprendizagem.

Fonte: Elaborado pelo autor

Ao final do processo de codificação, calculou-se o grau de concordância obtido entre os avaliadores por meio do índice Kappa. Dessa forma, obteve-se no *corpus* uma concordância percentual de 90.82% e k =0.82, um ótimo índice (Tabela 5). Os desacordos, foram resolvidos por um terceiro avaliador que seguiu os mesmos critérios que os demais.

A Tabela 12 a distribuição das três categorias da presença social, nos conjuntos de dados BioBase, TecBase e corpus, respectivamente.

Tabela 12 – Distribuição de classes nas bases e corpus

Categoria	**Mensagens (%) BioBase**	**Mensagens (%) TecBase**	**Mensagens (%) *Corpus***
Afetiva	176 (11.73%)	35 (4.76%)	211 (9,44%)
Interativa	1.053 (70.2%)	245 (33.37%)	1.298 (58.10%)
Coesiva	233 (15.53%)	52 (7.08%)	275 (12,30%)

Fonte: Elaborado pelo autor

A categoria com mais ocorrência de mensagens foi a Interativa, correspondendo por mais de 58.10% dos dados, as categorias Afetiva e Coesiva foram as que tiveram menos ocorrência com 9.44% e 12.30%, respectivamente. No trabalho de (FERREIRA et al., 2020) e (ROLIM et al., 2019 no prelo) também ocorreu essa diferença entre as categorias.

4.3 EXTRAÇÃO DE CARACTERÍSTICAS

Neste trabalho utilizou-se características extraídas da ferramenta LIWC e Coh-Metrix em suas versões para o português e também das próprias características extraídas do texto através da técnica de Bag of Words. Apesar de que alguns trabalhos não levarem em consideração essas características tradicionais, por exemplo em (KOVANOVIC et al., 2014b), devido sua dependência em relação ao domínio e gerar um vetor de alta dimensionalidade, decidimos inicialmente mantê-las.

Dessa forma, a extração das características foi orientada pela literatura e também baseada em estudos empíricos e após realizar os experimentos iniciais caso constatado que essas características impactassem de forma negativa o resultado elas seriam retiradas do conjunto de características. A seguir teremos o detalhamento das características utilizadas.

4.3.1 Características LIWC

A ferramenta LIWC, conforme detalhada na seção 2.2.3.3, conta palavras em categorias psicologicamente significativas de acordo com um dicionário (TAUSCZIK; PENNEBAKER, 2010), permitindo a extração de características de diferentes processos psicológicos, como por exemplo, afetivo, social e perceptual. Alguns estudos apontam bons resultados obtidos ao utilizar a ferramenta em contextos importantes para este trabalho, como (BARBOSA et al., 2020), para classificar automaticamente mensagens de discussão online, na língua inglesa, para os níveis de

presença cognitiva; (MACHADO et al., 2015) na busca dos sentimentos associados às palavras; e (FERREIRA et al., 2020) para classificar automaticamente mensagens, na língua inglesa, de discussão online para as categorias de presença social.

Sendo assim, foram utilizadas nesta pesquisa as 64 categorias disponíveis na versão em português brasileiro da ferramenta (FILHO et al., 2013). A Tabela 13 detalha as categorias, em suas respectivas dimensões, utilizadas como características.

Tabela 13 – Características LIWC

Nº	**Características**	**Descrição**
Linguística		
1	Função de palavras	Total de palavras com função
2	Total de pronomes	Número total de pronomes
3	Pronome pessoal	Número de pronomes pessoais
4	1ª pessoa do singular	Número de pronomes em primeira pessoa do singular
5	1ª pessoa do plural	Número de pronomes em primeira pessoa do plural
6	2ª pessoa	Número de pronomes em segunda pessoa
7	3ª pessoa do singular	Número de pronomes em terceira pessoa do singular
8	3ª pessoa do plural	Número de pronomes em terceira pessoa do plural
9	Pronomes impessoais	Número de pronomes impessoais
10	Artigos	Número de artigos
11	Verbos comuns	Número de verbos
12	Verbos auxiliares	Número de verbos auxiliares
13	Passado	Número de verbos no passado
14	Presente	Número de verbos no presente
15	Futuro	Número de verbos no futuro
16	Advérbios	Número de advérbios
17	Preposição	Número de preposições
18	Conjunções	Número de conjunções
19	Negação	Número de palavras que expressam negação
20	Quantificadores	Número de quantificadores
21	Números	Total de palavras numéricas
22	Palavrões	Número de palavrões
Psicológica		
23	Social	Número de palavras relacionadas a interação entre as pessoas (ex.: companheiro, conversa)
24	Família	Número de palavras que fazem referência a membros da família (ex.: pai, tia)

25	Amigos	Número de palavras que fazem referência a amizade (ex.: amigo, vizinho)
26	Humano	Número de palavras que fazem referência a seres humanos (ex.: adulto, garoto)
27	Afetivo	Número de palavras que expressam sentimentos (ex.: triste, chorão)
28	Emoções positivas	Número de palavras que expressam emoções positivas (ex.: feliz, bondade)
29	Emoções negativas	Número de palavras que expressam emoções negativas (ex.: detesto, odioso)
30	Ansiedade	Número de palavras que expressam ansiedade (ex.: medo, neurose)
31	Raiva	Número de palavras que expressam raiva (ex.: matar, aborrecido)
32	Tristeza	Número de palavras que expressam tristeza (ex.: triste, chorando)
33	Cognitivos	Número de palavras que fazem referência a características cognitivos (ex.: porque, sabedoria)
34	Intuição	Número de palavras que expressam intuição (ex.: penso, conheço)
35	Causa	Número de palavras que expressam causa (ex.: executado, efeito)
36	Discordância	Número de palavras que expressam discordância (ex.: expectativa, deveria)
37	Tentativa	Número de palavras que expressam tentativa (ex.: talvez, adivinhe)
38	Certeza	Número de palavras que expressam certeza (ex.: sempre, nunca)
39	Inibição	Número de palavras que expressam inibição (ex.: restringir, bloquear)
40	Inclusivo	Número de palavras que expressam inclusão (ex.: com, e)
41	Exclusivo	Número de palavras que expressam exclusão (ex.: mas, excluir)
42	Perceptivo	Número de palavras que expressam percepção (ex.: olhando, ouvindo)
43	Ver	Número de palavras que fazem referência a visão (ex.: ver, visto)

44	Ouvir	Número de palavras que fazem referência a audição (ex.: ouça, ouvindo)
45	Sentir	Número de palavras que fazem referência ao sentir (ex.: sente, toque)
46	Biológico	Número de palavras que fazem referência a características biológicas (ex.: sangue, dor)
47	Corpo	Número de palavras que fazem referência ao corpo (ex.: bochecha, mãos)
48	Saúde	Número de palavras que fazem referência à saúde (ex.: gripe, pílula)
49	Sexual	Número de palavras que fazem referência ao sexo (ex.: tesão, amor)
50	Ingestão	Número de palavras que fazem referência a ingestão (ex.: prato, bolos)
51	Relatividade	Número de palavras que fazem referência ao relativo (ex.: dobre, pare)
52	Movimento	Número de palavras que fazem referência a movimento (ex.: carro, fui)
53	Espaço	Número de palavras que fazem referência a espaço (ex.: abaixo, fina)
54	Tempo	Número de palavras que fazem referência ao tempo (ex.: fim, temporada)
55	Trabalho	Número de palavras que fazem referência a trabalho (ex.: trabalho, abandono)
56	Conquista	Número de palavras que fazem referência a conquistas (ex.: ganhe, herói)
57	Lazer	Número de palavras que fazem referência a lazer (ex.: bate-papo, filme)
58	Casa	Número de palavras que fazem referência ao lar (ex.: quarto, cozinha)
59	Dinheiro	Número de palavras que fazem referência a dinheiro (ex.: dinheiro, auditoria)
60	Religião	Número de palavras que fazem referência a religião (ex.: santo, igreja)
61	Morte	Número de palavras que fazem referência a morte (ex.: morrer, mate)
Falada		

62	Concordância	Número de palavras que fazem referência a concordância (ex.: concordo, de acordo)
63	Sem fluência	Número de palavras do tipo onomatopeia (figura de linguagem que permite o uso de vocábulos para representar som) (ex.: er, hm)
64	Enchimento	Número de palavras que fazem referência a enchimento (ex.: blá, sacou)

Fonte: Elaborado pelo autor (adaptado de (LANDIS; KOCH, 1977)

4.3.2 Características Coh-Metrix

O Coh-Metrix, descrito na seção 2.2.3.4, calcula medidas e índices de coesão e coerência de textos em um grande contexto de medidas. Neste trabalho utilizou-se a versão desenvolvida pelo grupo de pesquisa AIbox.edu, composto por alunos do curso de Computação da Universidade Federal Rural de Pernambuco, que implementa 94 medidas. A Tabela 14 descreve as medidas utilizadas como características para a classificação e uma descrição superficial de cada uma delas.

Tabela 14 – Características Coh-Metrix

Nº	**Características**	**Descrição**
Descritiva		
1	DESPC	Nº de parágrafos
2	DESSC	Nº de frases
3	DESWC	Nº de palavras
4	DESPL	Nº de frasespor parágrafo, média
5	DESPLd	Nº de frasespor parágrafo, desvio padrão
6	DESSL	Nº de palavras por frase, média
7	DESSLd	Nº de palavras por frase, desvio padrão
8	DESWLsy	Nº de sílabas por palavra, média
9	DESWLsyd	Nº de sílabas por palavra, desvio padrão
10	DESWLlt	Nº de letras por palavra, média
11	DESWLltd	Nº de letras por palavra, desvio padrão
Coesão Referencial		
12	CRFNO1	Sobreposição de substantivos, sentenças adjacentes
13	CRFAO1	Sobreposição de argumento, frases adjacentes
14	CRFSO1	Sobreposição de radical, frases adjacentes

15	CRFNOa	Sobreposição de substantivos, todas as frases
16	CRFAOa	Sobreposição de argumentos, todas as frases
17	CRFSOa	Sobreposição de radical, todas as frases
18	CRFCWO1	Sobreposição de palavra de conteúdo, frases adjacentes, média
19	CRFCWO1d	Sobreposição de palavra de conteúdo, frases adjacentes, desvio padrão
20	CRFCWOa	Sobreposição da palavra de conteúdo, todas as frases, média
21	CRFCWOad	Sobreposição da palavra de conteúdo, todas as frases, desvio padrão
22	CRFANP1	Sobreposição de anáforas, frases adjacentes
23	CRFANPa	Sobreposição de anáforas, todas as frases
LSA		
24	LSASS1	Sobreposição de LSA, sentenças adjacentes
25	LSASS1d	Sobreposição de LSA, frases adjacentes
26	LSASSp	Sobreposição de LSA, todas as frases no parágrafo, media
27	LSASSpd	Sobreposição de LSA, todas as frases no parágrafo, desvio padrão
28	LSAPP1	Sobreposição de LSA, parágrafos adjacentes, média
29	LSAPP1d	Sobreposição de LSA, parágrafos adjacentes, desvio padrão
30	LSAGN	LSA dado / novo, frases, média
31	LSAGNd	LSA dado / novo, frases, desvio padrão
Diversidade Lexical		
32	LDTTRc	Diversidade lexical, proporção de tipo-token, lemas de palavras de conteúdo
33	LDTTRa	Diversidade lexical, proporção de tipo-token, todas as palavras
34	LDMTLDa	Diversidade lexical, MTLD, todas as palavras
35	LDVOCDa	Diversidade lexical, VOCD, todas as palavras
Conectivos		
36	CNCAll	Ocorrência de todos os conectivos
37	CNCCaus	Ocorrência de conectivos causais
38	CNCLogic	Ocorrência de conectivos lógicos

39	CNCADC	Ocorrência de conectivos adversos e contrastivos
40	CNCTemp	Ocorrência de conectivos temporais
41	CNCAdd	Ocorrência de conectivos aditivos
42	CNCPos	Ocorrência de conectivos positivos
43	CNCNeg	Ocorrência de conectivos negativos
44	CNCAlter	Ocorrência de conjunções alternativas
45	CNCConclu	Ocorrência de conjunções conclusivas
46	CNCExpli	Ocorrência de conjunções explicativas
47	CNCConce	Ocorrência de conjunções concessiva
48	CNCCondi	Ocorrência de conjunções condicional
49	CNCConfor	Ocorrência de conjunções conformativas
50	CNCFinal	Ocorrência de conjunções finais
51	CNCProp	Ocorrência de conjunções proporcionais
52	CNCComp	Ocorrência de conjunções comparativas
53	CNCConse	Ocorrência de conjunções consecutivas
54	CNCInte	Ocorrência de conjunções integrantes
Modelo de Situação		
55	SMCAUSv	Ocorrência de verbo causal
56	SMCAUSvp	Verbos causais e incidência de partículas causais
57	SMCAUSr	Razão de partículas casuais para verbos causais
58	SMCAUSlsa	Sobreposição de LSA, verbo
59	SMCAUSwn	Sobreposição de verbos do WordNet
Complexidade Sintática		
60	SYNLE	Ocorrência de palavras antes do verbo principal,
61	SYNMEDpos	Distância mínima de edição, classe gramatical
62	SYNMEDwrd	Distância mínima de edição, todas as palavras
63	SYNMEDlem	Distância mínima de edição, lemas
64	SYNSTRUTa	Semelhança de sintaxe de sentença
65	SYNSTRUTt	Semelhança de sintaxe de frase
Densidade do Padrão Sintático		
66	DRNP	Densidade de frase nominal
67	DRVP	Densidade de frase de verbo
68	DRAP	Densidade de frase adverbial,
69	DRPP	Densidade de frase de preposição
70	DRPVAL	Densidade de voz passiva sem agente
71	DRNEG	Densidade de negação
72	DRGERUND	Densidade de gerúndios
73	DRINF	Densidade infinitiva

Informações sobre Palavras		
74	WRDNOUN	Ocorrência de substantivo
75	WRDVERB	Ocorrência de verbo
76	WRDADJ	Ocorrência de adjetivo
77	WRDADV	Ocorrência de advérbio
78	WRDPRO	Ocorrência de pronome
79	WRDPRP1s	Ocorrência de pronome na primeira pessoa do singular
80	WRDPRP1p	Ocorrência de pronome na primeira pessoa do plural
81	WRDPRP2	Ocorrência de pronome na segunda pessoa
82	WRDPRP3s	Ocorrência de pronome na terceira pessoa do singular
83	WRDPRP3p	Ocorrência de pronome na primeira pessoa do plural
84	WRDFRQc	CELEX frequência de palavras para palavras de conteúdo
85	WRDFRQa	CELEX Log de frequência para todas as palavras
86	WRDFRQmc	CELEX Log de frequência mínima para palavras
87	WRDAOAc	Idade de aquisição para palavras de conteúdo
88	WRDFAMc	Familiaridade para palavras de conteúdo
89	WRDCNCc	Concretude para palavras de conteúdo
90	WRDIMGc	Capacidade de imaginação para palavras de conteúdo
91	WRDMEAc	Significância, normas do Colorado
Legibilidade		
92	RDFRE	Índice de facilidade de leitura
93	RDFKGL	Índice Flesch (medida de complexidade do texto associada à sua inteligibilidade para diferentes tipos de leitores)
94	RDL2	Pontuação de legibilidade do segundo idioma

4.3.3 *Bag of Words* (BoW)

Conforme detalhado na subseção 2.2.3.5, o BoW é uma técnica onde o documento é representado por um vetor das contagens de palavras que aparecem nele.vA técnica foi aplicada após a etapa de pré-processamento onde foram normalizadas as palavras, removidas as *stopwords* e reduzido as palavras para a sua raiz (*stemming*). Ao final do processo obtivemos um total de 4.877 características.

A Tabela 15 abaixo contem um resumo das principais informações relacionadas aos recursos utilizados para extração de características.
Fonte: Elaborado pelo autor.

Tabela 15 – Resumo dos recursos

Nº	Recurso	Descrição	Nº de Características
1	Dicionário Léxico LIWC	Contagens de palavras que são indicativas de diferentes processos psicológicos.	64
2	Coh-Metrix	Calcula medidas e índices de coesão e coerência de textos.	95
3	Bow	Frequência de palavras	4.877
Total			5.036

4.4 BALANCEAMENTO DO CORPUS

Com o objetivo de separar os dados que o algoritmo de classificação utilizará para treinar e testar o modelo, adotou-se a divisão comumente utilizada no aprendizado de máquina onde 75% dos dados foram reservados para treino e 25% para testes (HASTIE et al., 2009). Essa etapa é executada para evitar que o desempenho do modelo se torne superestimada visto que a acurácia será calculada com os mesmos dados, ou seja, tanto o treinamento quanto o teste são realizados com o mesmo conjunto de dados.

Para conseguir uma amostragem estratificada dos dados de treinamento e teste foi utilizada a biblioteca *scikit-learn*, que fornece ferramentas de aprendizado de máquina eficientes e bem estabelecidas em várias áreas científicas (KRAMER, 2016). Após essa divisão obtivemos um conjunto de dados com 1.675 instâncias para a etapa de treinamento e 559 instâncias para etapa de testes. A Tabela 16 exibe a distribuição de cada categoria da presença social nas bases de treino e teste.

Tabela 16 – Distribuição das mensagens entre o grupo de treino e teste

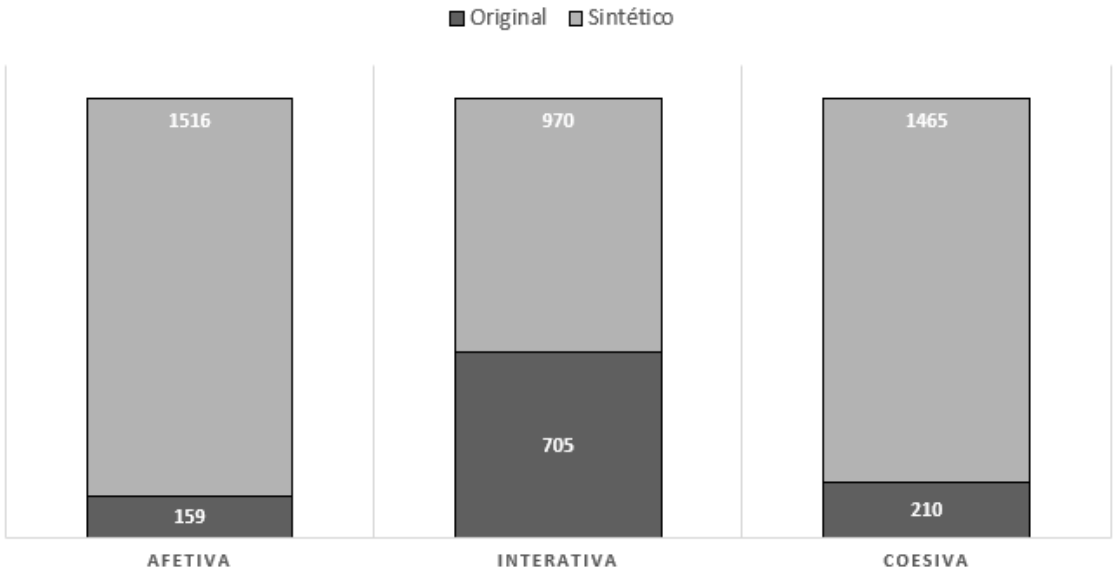
Categoria	Grupo	Negativa (0)	Positiva (1)	Total
Afetiva	Treino	1.516 (90.5%)	159 (9.5%)	1.675
	Teste	509 (91%)	50 (9%)	559
Interativa	Treino	970 (58%)	705 (42%)	1.675
	Teste	316 (56.52%)	243 (43.48%)	559
Coesiva	Treino	1.465 (87.46%)	210 (12.54%)	1.675
	Teste	498 (89.08%)	61 (10.92%)	559

Fonte: Elaborado pelo autor

Após essa divisão da base de dados, identificamos o problema do desbalanceamento das classes conforme exibido na Tabela 16 . Esse é um problema bastante estudado em pesquisas de aprendizagem de máquina clássica (BUDA et al., 2018) e de acordo com (JAPKOWICZ; STEPHEN, 2002), uma distribuição de dados com menor desbalanceamento, pode levar a melhora nos resultados dos algoritmos de classificação.

Para resolver esse problema foi utilizado o algoritmo SMOTE (CHAWLA et al., 2002), que tem como papel principal a criação sintética de instâncias de classes adicionais combinando linearmente as instâncias existentes. A Figura 8 apresenta o resultado final da aplicação do algoritmo SMOTE no conjunto de treinamento do *corpus*.

Figura 8 – Balanceamento de classe com SMOTE



4.5 CONSTRUÇÃO E OTIMIZAÇÃO DO MODELO

A literatura apresenta diversas abordagens de aprendizado de máquina que alcançaram resultados superiores no processamento de linguagem natural. O sucesso desses algoritmos depende de sua capacidade de gerar modelos através de um conjunto de dados e conseguir classificar novas instâncias. Porém, encontrar técnicas apropriadas para classificação de texto é um desafio para os pesquisadores (KOWSARI et al., 2019). A escolha do algoritmo demanda esforço e compreensão do problema abordado.

Existem vários algoritmos de aprendizado de máquina com o objetivo de desenvolver modelos supervisionados. Em (FERNÁNDEZ-DELGADO et al., 2014), é apresentada uma analise que compara 179 algoritmos de classificação de propósito geral em 121 conjuntos de dados diferentes que apontam o algoritmo random forest e SVM como os melhores em desempenho. Na revisão da literatura sobre algoritmos de classificação de texto realizado por (KOWSARI et al., 2019), o autor destaca que técnicas de classificação por voto, como *bagging* e *boosting*, foram desenvolvidas com sucesso para classificação de dados textuais (FARZI; BOLANDI, 2016). Dentre os algoritmos que implementam essas técnicas o autor aponta o

random forest e adaboost. Segundo o autor, as vantagens de utilizar-se de algoritmos *bagging* e *boosting* é que eles melhoram a estabilidade e a acurácia, tirando proveito do aprendizado de conjuntos (ensemble learning) e também reduzem a variância que ajuda a evitar problemas de sobreajsute (overfitting). Outro algoritmo de *boosting* que vem sendo utilizado atualmente é o XgBoost que combina modelo linear e modelo de aprendizado em árvore (CHEN; GUESTRIN, 2016).

Neste trabalho foram utilizados os algoritmos Random Forest, Adaboost e XGBoost, devido aos seus ótimos desempenhos apresentados na literatura e também pretendia-se analisar o quanto cada característica contribui para a classificação, ou seja, quais as características mais influentes para classificar cada categoria da presença social. Para avaliar a importância das características durante a classificação, foi escolhida a medida *Mean Decrease Gini* (MDG), uma medida amplamente utilizada que explica a separabilidade de uma determinada característica em relação às categorias (BREIMAN, 2001).

Ainda de acordo com (BREIMAN, 2001), os principais parâmetros do algoritmo Random Forest são o número de variáveis de entrada escolhidas aleatoriamente em cada divisão (max_features) e o número de árvores na floresta (n_estimators). A primeira parte do experimento buscou otimizar esses parâmetros, para isso foi utilizada a técnica de validação cruzada para cada conjunto de treinamento (Afetivo, Interativo e Coesivo), onde cada partição (*fold*) era composta por uma das semanas do curso, quatro semanas no curso de biologia e três no de tecnologia, no respectivo conjunto de treinamento. Inicialmente buscou-se ajustar o parâmetro *max_features*, portanto foi definido o número de árvores em 1.500 e em cada execução da validação cruzada eram verificados valores para o parâmetro. Foi definido de forma aleatória o valor de 140 e para cada execução eram acrescidos mais 140 características ao parâmetro obedecendo o número máximo de características possíveis (5.036) e sem repetição. Após definido o *max_features*, o próximo parametro a ser otimizado foi o número de árvores (*n_estimator*), para isso foi utilizado o OOB error.

Após a etapa de otimização dos parâmetros foram realizados outros cinco experimentos, em cenários diferentes, utilizando não apenas o algoritmo Random Forest mas também o AdaBoost e o XGBoost. Os dados foram divididos em conjunto de treinamento (75%) e teste (25%), conforme Tabela 16, e os valores dos parâmetros que foram os obtidos na otimização também foram utilizados. Os cenários foram: i) Treinamento e teste com o *corpus*; ii) Treinamento e teste com a base de tecnologia (TecBase); iii) Treinamento e teste com a base de biologia (BioBase); iV) Treinamento com a base de biologia e teste com a base de tecnologia

(Bio-Tec); v) Treinamento com a base de tecnologia e teste com a base de biologia (Tec-Bio). Durante os experimentos foi definida uma semente (*seed*) com o valor cinco para se obter um comportamento determinístico possibilitando assim a reprodutibilidade do experimento. Por fim, foi verificado quais as características mais relevantes para o classificador.

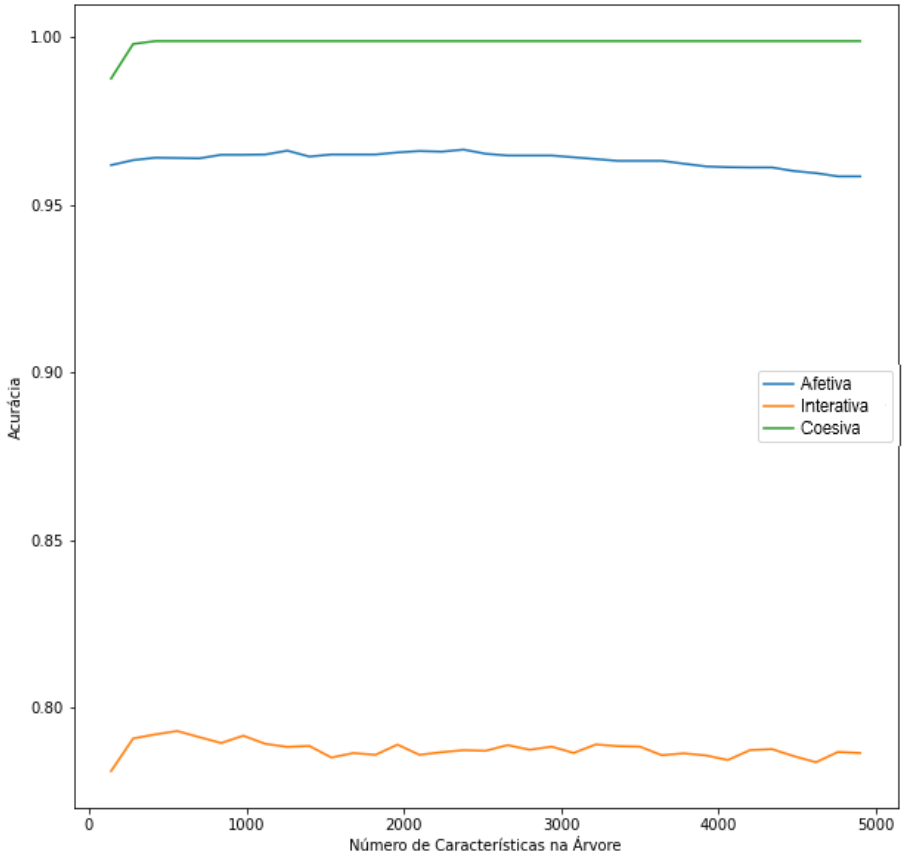
5 RESULTADOS

Neste capítulo serão exibidos os resultados alcançados com a aplicação do método exposto no capítulo anterior, através dos experimentos realizados em cada cenário, e as discussões a respeito. Além dos resultados dos cinco experimentos, serão apresentadas as 20 características mais relevantes utilizadas pelo classificador de cada categoria em 3 cenários distintos.

5.1 MODELO DE TREINAMENTO E AVALIAÇÃO - QP01

Como exposto anteriormente, selecionamos dois parâmetros para serem utilizados no algoritmo *random forest*: *max_features* e *n_estimator*. Foi realizado o procedimento de otimização desses parâmetros, conforme descrito na seção 4.5, o primeiro parâmetro a ser ajustado foi o *max_features*. A Figura 9 mostra o resultado desse procedimento, percebe-se que a acurácia média obtida em cada uma das três categorias começou a se estabilizar quando o número de características era em torno de 2.000.

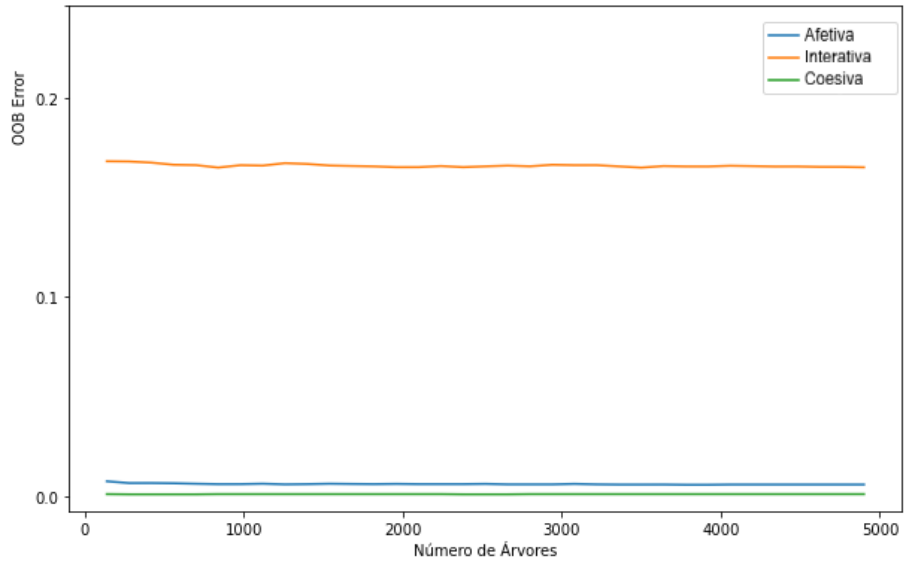
Figura 9 – Resultado do ajuste do parâmetro *max_feature*.



Uma vez definido o parâmetro *max_features*, buscou-se encontrar o número adequado de árvores (*n_estimator*) para o problema abordado. Para isso foi utilizado o OOB error, conforme

podemos observar na Figura 10, mesmo realizando diversas alterações no parâmetro não houve uma diferença significativa, dessa forma, foi definido o número de árvores como 800.

Figura 10 – Melhor desempenho da configuração do classificador Random Forest.



A tabela 17 mostra os resultado dos classificadores Random Forest, com os parâmetros padrão e otimizado, usando validação cruzada no conjunto de treinamento. Os resultados em relação a acurácia, aumentaram 15%, 18% e 10% para as categorias Afetiva, Interativa e Coesiva, respectivamente. Em relação ao kappa, houve uma melhora de 140%, 27.2% e 46.2% para as três categorias citadas anteriormente. Os resultados dessa primeira parte do experimento demonstram a importância de se realizar o ajuste dos parâmetros utilizados.

Tabela 17 – Resultado da otimização dos parâmetros do Random Forest.

Categoria	**Parâmetros**	**Acurácia**	**Kappa**
Afetiva	Padrão	0.80	0.20
	Otimizado	0.92	0.48
Interativa	Padrão	0.72	0.55
	Otimizado	0.85	0.70
Coesiva	Padrão	0.90	0.67
	Otimizado	0.99	0.98

Nas seções 5.2 à 5.4 serão apresentados os resultados de cada cenário, para os três algoritmos, contendo além das métricas de acurácia e kappa a matriz de confusão (Tabela 4). Na seção 5.8, serão discutidos com mais detalhes os resultados dos cenários I, II e III.

5.2 RESULTADOS DO CENÁRIO I (BIOBASE + TECBASE)

Para o cenário I foi utilizado o *corpus* completo (BioBase + TecBase), contendo 2.234 mensagens, onde 1.675 (75%) foram utilizadas para treinamento e 559 (25%) foram utilizadas para teste. Nesse cenário os resultados entre os algoritmos Random Forest e XGBoost tiveram um desempenho parecido, quando se trata da acurácia, mas o XGBoost obteve um melhor resultado principalmente na categoria Afetiva com 0.92 e 0.49 de acurácia e kappa, respectivamente. Na Tabela 18 podemos verificar os resultados de cada classificador em relação a cada uma das três categorias.

A Tabela 19 mostra a matriz de confusão gerada para cada categoria. A taxa de falso positivo na classe afetiva confirma os resultados de kappa, pois o XGBoost apresentou a menor taxa ficando com 2.36% , em seguida o Random Forest com 2.95% e o AdaBoost com a maior taxa de 4.52% devido aos seus 23 exemplos falso positivos.

Tabela 18 – Resultado Corpus

	Random Forest		**AdaBoost**		**XGBoost**	
Categoria	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**
Afetiva	0.92	0.48	0.90	0.36	0.92	0.49
Interativa	0.85	0.70	0.81	0.62	0.85	0.71
Coesiva	0.99	0.98	0.99	0.98	0.99	0.98

Tabela 19 – Matriz de confusão (Corpus)

		Afetiva		**Interativa**		**Coesiva**	
		neg	pos	neg	pos	neg	pos
Random Forest	neg	494	15	281	35	498	0
	pos	27	23	46	197	2	59
AdaBoost	neg	486	23	274	42	498	0
	pos	31	19	60	183	2	59
XGBoost	neg	497	12	289	27	498	0
	pos	28	22	53	190	2	59

5.3 RESULTADOS DO CENÁRIO II (BIOBASE)

No cenário II foi utilizada apenas a base de dados de biologia, composta por 1.500 mensagens, onde 1.125 (75%) foram utilizadas para treinar o modelo e 375 (25%) foram utilizadas para testá-lo. Na Tabela 20 são apresentados os resultados de cada classificador em relação a cada

uma das três categorias. Pode-se observar que na categoria Interativa ocorreu uma uniformidade entre os resultados de todos os classificadores mas na categoria Afetiva o classificador XGBoost se destacou com resultado de 0.90 e 0.40 para acurácia e kappa, respectivamente. Enquanto isso par a categoria Coesiva o Random Forest obteve melhor desempenho principalmente em relação ao kappa com valor de 0.67.

A matriz de confusão gerada conforme ilustrada na Tabela 21, indica que apesar do Random Forest ter obtido a maior precisão entre os demais, 97.12%, sua taxa de falso positivo foi cerca de 70% maior que a dos outros algoritmos em relação a categoria Coesiva. Em relação ao XGBoost além de uma precisão de 93.37% sua taxa de falso positivo foi a menor com 3.86% em relação a categoria Afetiva, respectivamente.Vale ressaltar que a taxa de falso positivo para todos os algoritmos em relação a categoria Interativa foi bastante alta, com valores de 16.96, 19.88% e 20.47% para os classificadores Random Forest, AdaBoost e XGBoost respectivamente.

Tabela 20 – Resultado BioBase

	Random Forest		**AdaBoost**		**XGBoost**	
Categoria	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**
Afetiva	0.88	0.25	0.88	0.29	0.90	0.40
Interativa	0.83	0.66	0.82	0.65	0.83	0.65
Coesiva	0.91	0.67	0.91	0.60	0.91	0.60

Tabela 21 – Matriz de confusão (BioBase)

		Afetiva		**Interativa**		**Coesiva**	
		neg	pos	neg	pos	neg	pos
Random Forest	neg	322	15	142	29	304	22
	pos	28	10	35	169	9	40
AdaBoost	neg	320	17	137	34	313	13
	pos	26	12	31	173	19	30
XGBoost	neg	324	13	136	35	314	12
	pos	23	15	30	174	20	29

5.4 RESULTADOS DO CENÁRIO III (TECBASE)

Nesse cenário foi utilizada apenas a base de tecnologia, composta por 734 mensagens, dessas, 550 (75%) foram utilizadas para a fase de treinamento e 184 (25%) para a fase de teste. Diante dos resultados apresentados na Tabela 22 pode ser observado que o algoritmo Random Forest obteve o melhor resultado na categoria Coesiva com acurácia e kappa de 0.97 e 0.76,

respectivamente. Enquanto isso o algoritmo XGBoost teve o melhor desempenho nas demais categorias.

Na Tabela 23 é apresentada a matriz de confusão gerada. Apesar de todos os algoritmos acertarem 7 exemplos da categoria Afetiva, ou seja, uma precisão igual para todos a taxa de falso positivo do Random Forest para a mesma categoria foi a menor, 0.57%, confirmando seu bom resultado dentre os demais.

Tabela 22 – Resultado TecBase

	Random Forest		**AdaBoost**		**XGBoost**	
Categoria	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**
Afetiva	0.94	0.39	0.96	0.51	0.96	0.57
Interativa	0.90	0.68	0.88	0.61	0.91	0.70
Coesiva	0.97	0.76	0.97	0.72	0.97	0.68

Tabela 23 – Matriz de confusão (TecBase)

		Afetiva		**Interativa**		**Coesiva**	
		neg	pos	neg	pos	neg	pos
Random Forest	neg	169	5	143	4	173	1
	pos	6	4	13	24	3	7
AdaBoost	neg	173	1	138	9	172	2
	pos	6	4	13	24	3	7
XGBoost	neg	172	2	141	6	171	3
	pos	5	5	11	26	3	7

5.5 RESULTADOS DO CENÁRIO IV (BIO-TEC)

No cenário IV, o conjunto de treinamento foi composto pela base de biologia e o conjunto de teste pela base de tecnologia. Apesar de que nas categorias Afetiva e Interativa todos os algoritmos obtiveram um resultado razoável em relação a acurácia, os resultados do índice kappa foram extremamente baixos, ou seja, uma concordância fraca e em alguns casos insignificante. Isso aconteceu devido ao desbalanceamento Apenas na categoria Coesiva todos os algoritmos tiveram bons resultados, inclusive com os mesmos valores de 0.99 e 0.95 para acurácia e kappa, respectivamente. Na Tabela 24 podemos visualizar todos os resultados desse experimento.

Na Tabela 25 é apresentada a matriz de confusão, confirma o bom resultado da categoria Afetiva onde sua taxa de falso positivo é 0%. Sobre a categoria Afetiva, todos os algoritmos tiveram dificuldades para identificar a classe positiva.

Tabela 24 – Resultado Bio-Tec

	Random Forest		**AdaBoost**		**XGBoost**	
Categoria	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**
Afetiva	0.94	0.01	0.94	0.10	0.95	0.19
Interativa	0.73	0.25	0.44	0.12	0.60	0.24
Coesiva	0.99	0.95	0.99	0.95	0.99	0.95

Tabela 25 – Matriz de confusão (Bio-Tec)

		Afetiva		**Interativa**		**Coesiva**	
		neg	pos	neg	pos	neg	pos
Random Forest	neg	695	5	463	123	686	0
	pos	34	0	75	73	4	44
AdaBoost	neg	691	9	193	393	686	0
	pos	31	3	14	134	4	44
XGBoost	neg	698	2	312	274	686	0
	pos	33	1	21	127	4	44

5.6 RESULTADOS DO CENÁRIO V (TEC-BIO)

O cenário V teve como conjunto de treinamento a base de tecnologia e o modelo gerado foi testado utilizando-se a base de biologia. A categoria Afetiva teve o pior desempenho pois todos os classificadores obtiveram o índice kappa insignificante. Em relação a categoria Interativa apenas os classificadores Random Forest e XGBoost alcançaram um kappa com valores 0.52 e 0.43, respectivamente, valores esses considerados como uma concordância moderada. A categoria Coesiva alcançou o valor máximo para ambas as métricas em todos os classificadores, os resultados podem ser observado na Tabela 26.

Assim como no resultado anterior, todos os algoritmos apresentaram excelente resultado para a categoria Coesiva com uma taxa de falso positivo de 0% como pode ser visualizado na matriz de confusão exibida na Tabela 27. O mesmo acontece com a categoria Afetiva, apesar de uma acurácia alta o modelo não conseguiu uma assertividade na classificação das classes positivas.

Tabela 26 – Resultado Tec-Bio

	Random Forest		**AdaBoost**		**XGBoost**	
Categoria	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**
Afetiva	0.84	0.03	0.84	0.02	0.85	0.02
Interativa	0.76	0.52	0.64	0.29	0.71	0.43
Coesiva	1.00	1.00	1.00	1.00	1.00	1.00

Tabela 27 – Matriz de confusão (Tec-Bio)

		Afetiva		**Interativa**		**Coesiva**	
		neg	pos	neg	pos	neg	pos
Random Forest	neg	1264	61	579	121	1277	0
	pos	171	4	234	566	0	223
AdaBoost	neg	1255	70	509	191	1277	0
	pos	162	13	339	461	0	233
XGBoost	neg	1263	62	545	155	1277	0
	pos	163	12	274	526	0	223

5.7 CARACTERÍSTICAS IMPORTANTES - QP02

O presente trabalho também verificou quais características contribuíram para o desempenho do modelo de cada categoria. Para essa análise foi escolhido o melhor classificador dos cenários I, II e III que utilizaram o corpus, base de biologia e base de tecnologia, respectivamente. Para cada um dos três modelos gerados pelo classificador em cada cenário mencionado anteriormente, foram selecionadas as vinte características com maior pontuação MDG. Em cada experimento além das 159 características oriundas do LIWC e Coh-Metrix, também foram utilizadas as relacionadas com a frequência de palavras de cada conjunto de dados.

Para identificar a fonte das características analisadas, serão utilizados os prefixos *liwc* e *cm* para o LIWC e Coh-Metrix, respectivamente e sem prefixo para a frequência de palavras. No geral, desconsiderando repetições de características entre as categorias de cada modelo, houve uma ocorrência de 90 características de frequência de palavras, 58 do Coh-Metrix e 32 do LIWC, ou seja, 50% de características tradicionais (BoW) e 50% dos recursos relacionados aos outros dois recursos. A característica do LIWC com maior pontuação foi *liwc.article* com MDG de 12.58, em relação ao Coh-Metrix foi a *cm.LSAGN* com a pontuação de 17.79 e por fim, na frequência de palavras a característica *boa* com MDG de 90.93.

A seguir, nas seções 5.7.1, 5.7.2 e 5.7.3 serão apresentadas com mais detalhes essas vinte melhores características utilizadas pelo classificador de maior desempenho em cada cenário.

5.7.1 Características Importantes do Cenário I (BioBase + TecBase)

Para o cenário I, o algoritmo com maior desempenho foi o XGBoost. Para a categoria afetiva o modelo considerou um total de 35 característica da frequência de palavras, 13 características do Coh-Metrix e 12 do LIWC. Apesar de 50% das características serem compostas do recurso BoW, suas pontuações são baixas com exceção da palavra boa que na categoria Coesiva ficou com MDG de 90.93. Nas demais categorias a predominância de características com maiores pontuações foi de recursos do LIWC e Coh-Metrix, por exemplo, a categoria afetiva contem nas suas 8 primeiras variáveis desses recursos mencionados. As Tabelas 28, 29 e 30 contem os detalhes das variáveis em suas respectivas categorias.

Tabela 28 – Vinte características mais importantes da categoria Afetiva (XGBoost)

Nº	**Variável**	**Descrição**	**MDG**
1	liwc.posemo	Nº de palavras que expressam emoções positivas	3.23
2	entant	Frequência de palavras	2.88
3	cm.LSASSp	Índice de semelhança da frase com as demais	2.72
4	efici	Frequência de palavras	2.64
5	liwc.see	Nº de palavras que fazem referência a visão	2.54
6	precoc	Frequência de palavras	2.50
7	vam	Frequência de palavras	2.21
8	liwc.health	Nº de palavras relacionadas a saúde	2.16
9	cm.CNCAll	Nº de todos os conectivos que aparecem em um texto	2.03
10	liwc.negate	Nº de palavras que expressam negação	2.02
11	ola	Frequência de palavras	1.81
12	cm.CRFCWOa	Índice global de palavras de conteúdo explícito que se sobrepõem entre pares de frases	1.75
13	coleg	Frequência de palavras	1.67
14	turm	Frequência de palavras	1.60
15	bem	Frequência de palavras	1.59
16	cm.SMCAUSv	Nº de verbos causais.	1.59
17	cm.LDVOCDa	Diversidade lexical, VOCD	1.53
18	rapid	Frequência de palavras	1.38
19	cm.DRVP	Nºde frases verbais	1.37
20	caus	Frequência de palavras	1.32

Tabela 29 – Vinte características mais importantes da categoria Interativa (XGBoost)

Nº	Variável	Descrição	MDG
1	cm.WRDNOUN	Nº de substantivos	10.55
2	liwc.relativ	Nº de palavras que fazem referência ao relativo	7.60
3	liwc.incl	Nº de palavras que fazem referência a inclusão	5.06
4	liwc.cogmech	Número de palavras que fazem referência a características cognitivas	2.10
5	liwc.article	Nº de artigos	1.90
6	liwc.preps	Nº de preposições	1.63
7	liwc.achieve	Nº de palavras que fazem referência a realização	1.47
8	cm.WRDPRP2	Nº de pronomes em segunda pessoa	1.31
9	silv	Frequência de palavras	1.01
10	tecnolog	Frequência de palavras	1.00
11	ola	Frequência de palavras	0.97
12	liwc.you	Nº de pronomes em segunda pessoa	0.97
13	concord	Frequência de palavras	0.93
14	process	Frequência de palavras	0.92
15	individu	Frequência de palavras	0.91
16	int	Frequência de palavras	0.88
17	gene	Frequência de palavras	0.74
18	popul	Frequência de palavras	0.72
19	fim	Frequência de palavras	0.72
20	marcel	Frequência de palavras	0.70

Tabela 30 – Vinte características mais importantes da categoria Coesiva (XGBoost)

Nº	Variável	Descrição	MDG
1	boa	Frequência de palavras	90.93
2	cm.RDL2	Pontuação de legibilidade do segundo idioma.	0.35
3	cm.WRDMEAc	Média de significância das palavras em um corpus	0.28
4	cm.SYNSTRUTa	Proporção de nós da árvore de interseção entre todas as sentenças adjacentes.	0.27
5	liwc.adverb	Nº de advérbios	0.26
6	cm.WRDIMGc	Índice de como é fácil construir uma imagem mental da palavra	0.25
7	cm.DESSL	Tamanho das sentenças	0.25
8	impact	Frequência de palavras	0.24

9	mid	Frequência de palavras	0.23
10	const	Frequência de palavras	0.23
11	dess	Frequência de palavras	0.23
12	desaf	Frequência de palavras	0.23
13	real	Frequência de palavras	0.23
14	futur	Frequência de palavras	0.22
15	marc	Frequência de palavras	0.22
16	profss	Frequência de palavras	0.22
17	segund	Frequência de palavras	0.22
18	nest	Frequência de palavras	0.22
19	coleg	Frequência de palavras	0.22
20	apresent	Frequência de palavras	0.22

5.7.2 Características Importantes do Cenário II (BioBase)

O classificador Random Forest obteve os melhores resultados no conjunto de dados composto apenas pela base de biologia (BioBase). Das 60 características, 40 (66.6%) tem como fonte as duas bibliotecas utilizadas, 23 do Coh-Metrix e 17 do LIWC, as 20 restantes são oriundas da frequência de palavras. A característica do LIWC com maior pontuação foi a liwc.article com MDG de 12.58, no caso do Coh-Metrix com 6.28 de MDG a característica cm.LDVOCDa foi a que mais se destacou e a característica olá ficou com MDG de 28.45, maior pontuação entre as características resultantes da frequência de palavras. Na categoria Afetiva e Interativa a predominância de variáveis com maior pontuação tem como origem o LIWC e Coh-Metrix, enquanto na categoria Coesiva as cinco primeiras variáveis são de frequência de palavras. As Tabelas 31, 32 e 33 a seguir possuem mais detalhes em relação as melhores características de cada categoria.

Tabela 31 – Vinte características mais importantes da categoria afetiva (Random Forest)

Nº	**Variável**	**Descrição**	**MDG**
1	coleg	Frequência de palavras	6.39
2	cm.LDVOCDa	Diversidade lexical, VOCD	6.28
3	cm.CNCAll	Nº total de conectivos	6.14
4	liwc.insight	Nº de palavras que expressam intuição	3.25
5	liwc.bio	Nº de palavras que fazem referência a características biológicas	2.81
6	liwc.relativ	Nº de palavras que fazem referência ao relativo	2.35
7	microscopi	Frequência de palavras	2.15

8	cm.WRDADV	Nº de advérbios	1.81
9	liwc.negemo	Nº de palavras que expressam emoções negativas	1.28
10	liwc.see	Nº de palavras que fazem referência a visão	1.28
11	cm.WRDADJ	Nº de adjetivos	1.24
12	liwc.certain	Nº de palavras que expressam certeza	1.19
13	rapid	Frequência de palavras	1.15
14	agor	Frequência de palavras	1.10
15	liwc.friend	Nº de palavras que fazem referência a amizade	1.09
16	ola	Frequência de palavras	1.09
17	cm.WRDVERB	Nº de verbos	0.91
18	rim	Frequência de palavras	0.84
19	liwc.time	Nº de palavras que fazem referência ao tempo	0.82
20	abort	Frequência de palavras	0.81

Tabela 32 – Vinte características mais importantes da categoria Interativa (Random Forest)

Nº	**Variável**	**Descrição**	**MDG**
1	liwc.article	Nº de artigos	12.58
2	cm.WRDNOUN	Nº de substantivos	5.29
3	liwc.incl	Nº de palavras que expressam inclusão	4.74
4	liwc.funct	Nº total de palavras funcionais	3.83
5	cm.WRDADJ	Nº de adjetivos	3.45
6	liwc.preps	Nº de preposições	2.91
7	celul	Frequência de palavras	1.93
8	liwc.cogmech	Número de palavras que fazem referência a características cognitivas	1.83
9	cm.WRDPRP2	Nº de pronomes em segunda pessoa	1.68
10	liwc.relativ	Nºde palavras que fazem referência ao relativo	1.44
11	cm.DESWC	Nº total de palavras	1.26
12	cm.DESWLltd	Desvio padrão da medida para o número médio de letras nas palavras dentro do texto	1.04
13	estud	Frequência de palavras	0.86
14	citolog	Frequência de palavras	0.86
15	liwc.you	Nº de pronomes em segunda pessoa	0.85
16	ola	Frequência de palavras	0.69
17	cm.DESWLsyd	Desvio padrão da medida para o número médio de sílabas nas palavras do texto	0.69
18	liwc.achieve	Nº de palavras que expressam realização	0.61

19	cm.RDFKGL	Índice Flesch (medida de complexidade do texto associada à sua inteligibilidade para diferentes tipos de leitores)	0.58
20	microscopi	Frequência de palavras	0.56

Tabela 33 – Vinte características mais importantes da categoria Coesiva (Random Forest)

Nº	**Variável**	**Descrição**	**MDG**
1	ola	Frequência de palavras	28.45
2	boa	Frequência de palavras	26.42
3	noit	Frequência de palavras	10.11
4	bom	Frequência de palavras	3.37
5	tard	Frequência de palavras	2.83
6	cm.RDFKGL	Índice Flesch (medida de complexidade do texto associada à sua inteligibilidade para diferentes tipos de leitores)	1.24
7	dia	Frequência de palavras	1.24
8	cm.RDFRE	Índice de facilidade de leitura	0.68
9	cm.WRDNOUN	Nº de substantivos	0.49
10	cm.WRDADJ	Nº de adjetivos	0.38
11	cm.WRDVERB	Nº de advérbios	0.36
12	cm.RDL2	Pontuação de legibilidade do segundo idioma.	0.36
13	cm.SYNMEDpos	pontuação média mínima da distância editorial entre sentenças adjacentes calculadas a partir de parte das tags de fala	0.33
14	cm.WRDIMGc	Índice de como é fácil construir uma imagem mental da palavra	0.32
15	cm.DESWLlt	Nº médio de letras para todas as palavras no texto	0.32
16	girlen	Frequência de palavras	0.29
17	curi	Frequência de palavras	0.28
18	cm.WRDFAMc	Classificação de quão familiar uma palavra parece para um adulto	0.27
19	liwc.they	Nº de pronomes em terceira pessoa do plural	0.26
20	cm.DESWLsyd	Desvio padrão da medida para o número médio de sílabas nas palavras do texto	0.25

5.7.3 Características Importantes do Cenário III (TecBase)

No cenário III onde foi utilizado apenas o conjunto de dados de tecnologia, o algoritmo XGBoost obteve o melhor resultado entre os demais. De forma geral houve uma maior ocorrência de características oriundas da frequência de palavras, um total de 35 (58.3%). As características

do Coh-Metrix e LIWC tiveram ocorrência de 22 (36.6%) e 3 (5.1%), respectivamente. Em relação as características da franquênia de palavras, a característica boa obteve maior pontuação MDG, atingindo 26.76. No Coh-Metrix a característica cm.LSAGN destacou-se com um MDG de 17.79. Por fim, no LIWC, a característica liwc.home teve uma pontuação MDG de 2.35, a maior dentre as características da mesma fonte. Nas três categorias ao observar as cinco primeiras características com melhor pontuação existe uma maior ocorrência de características com origem na frequência de palavras,com uma pequena exceção, pois na categoria afetiva a primeira característica faz parte do Coh-Metrix. Nas tabelas 34, 35 e 36 as características são melhor detalhadas.

Tabela 34 – Vinte características mais importantes da categoria Afetiva (XGBoost)

Nº	**Variável**	**Descrição**	**MDG**
1	cm.LSAGN	Disponibilidade média de cada sentença	17.79
2	ava	Frequência de palavras	9.42
3	prepar	Frequência de palavras	6.16
4	alun	Frequência de palavras	4.55
5	seguint	Frequência de palavras	4.10
6	cm.CRFCWOa	Índice global de palavras de conteúdo explícito que se sobrepõem entre pares de frases	3.74
7	izabel	Frequência de palavras	3.48
8	educ	Frequência de palavras	2.99
9	cm.DESWLsy	Nº médio de sílabas em todas as palavras do texto	2.59
10	conhec	Frequência de palavras	2.50
11	fal	Frequência de palavras	2.48
12	cm.SYNMEDwrd	Pontuação mínima da distância editorial entre sentenças adjacentes calculadas a partir de palavras	2.42
13	cm.LSASSp	Índice de semelhança da sentença com as demais	2.18
14	particip	Frequência de palavras	2.15
15	cm.LSASS1	Sobreposição de argumentos em sentenças adjacentes.	2.02
16	cm.WRDPRP3s	Nº de pronomes em terceira pessoa do singular	1.96
17	cm.SMCAUSlsa	Sobreposição de verbos LSA	1.85
18	cinem	Frequência de palavras	1.74
19	cm.DRPVAL	Nº de formas de voz passivas sem agente	1.64
20	pod	Frequência de palavras	1.20

Tabela 35 – Vinte características mais importantes da categoria Interativa (XGBoost)

Nº	Variável	Descrição	MDG
1	turm	Frequência de palavras	11.16
2	concord	Frequência de palavras	11.15
3	primord	Frequência de palavras	4.38
4	tem	Frequência de palavras	3.77
5	son	Frequência de palavras	3.04
6	cm.WRDPRP2	Nº de pronomes em segunda pessoa	3.01
7	medi	Frequência de palavras	3.00
8	inform	Frequência de palavras	2.64
9	comunic	Frequência de palavras	2.36
10	liwc.home	Nº palavras que fazem referência ao lar	2.35
11	real	Frequência de palavras	1.60
12	cm.CRFNO1	medidas de sobreposição local e global entre sentenças em termos de substantivos	1.50
13	dentr	Frequência de palavras	1.41
14	cm.LSASSp	Cossenos médios de LSA	1.28
15	liwc.incl	Nº de palavras que expressam inclusão	1.24
16	sent	Frequência de palavras	1.23
17	dinam	Frequência de palavras	1.16
18	cm.CNCTemp	Nº de conectivos temporais	1.16
19	cm.SMCAUSlsa	Sobreposição de LSA entre verbos.	1.11
20	import	Frequência de palavras	1.02

Tabela 36 – Vinte características mais importantes da categoria Coesiva (XGBoost)

Nº	Variável	Descrição	MDG
1	boa	Frequência de palavras	26.76
2	turm	Frequência de palavras	15.78
3	discuss	Frequência de palavras	7.11
4	cm.SMCAUSlsa	Sobreposição de LSA entre verbos	6.50
5	cm.LDTTRc	Taxa de frequência de token	5.29
6	ola	Frequência de palavras	4.95
7	cm.DESSL	Nº médio de palavras em cada frase do texto	2.67
8	const	Frequência de palavras	2.09
9	noit	Frequência de palavras	2.09

10	cm.LSASS1	mede quão conceitualmente semelhante cada sentença é à próxima sentença.	1.92
11	human	Frequência de palavras	1.79
12	bom	Frequência de palavras	1.37
13	desej	Frequência de palavras	1.19
14	prof	Frequência de palavras	1.10
15	cm.CRFSOa	Sobreposição das frases passadas	1.07
16	cm.CNCConfor	Incidência de conjunções conformativas	1.02
17	educ	Frequência de palavras	0.85
18	cm.WRDFAMc	Classificação de quão familiar uma palavra parece ara um adulto	0.83
19	cm.DESWLsy	Nº médio de sílabas nas palavras dentro do texto	0.81
20	liwc.friend	Nº palavras que fazem referência a amizade	0.80

5.8 DISCUSSÃO

Nesta seção serão apresentadas as discussões dos resultados dos experimentos nos cenários I (BioBase + TecBase), II (BioBase) e III (TecBase), que tiveram como objetivo automatizar a identificação da presença social através de modelos preditivos. Também será demonstrada uma análise comparativa sobre as características da presença social em dois contextos diferentes.

5.8.1 Análise dos Resultados dos Modelos (Cenário I)

Ao tratar da Questão de Pesquisa 01, apresentada na seção 1.2, os resultados alcançados, utilizando-se o *Corpus* (BioBase + TecBase) como conjunto de dados, com a classificação automática das três categorias da presença social foram satisfatórios. Foram obtidos índices kappa (k) de 0.49, 0.71 e 0.98, para as categorias Afetiva, Interativa e Coesiva, respectivamente. Assim como nos trabalhos encontrados na literatura, a categoria Afetiva possui um baixo valor de concordância, isso deve-se ao fato de existir pouca quantidade de amostras dessa categoria no conjunto de dados. De qualquer forma, de acordo com (LANDIS; KOCH, 1977), esses valores representam uma taxa de concordância entre avaliadores de média (0.49) a quase perfeita (0.98), ver Tabela 5. Esses valores demonstram que as características baseadas no LIWC e Coh-Metrix, em conjunto com as características referentes a frequência de palavras são eficazes para identificar as mensagens de fóruns de discussão em português segundo as três categorias.

Sobre a otimização realizada para o parâmetro *max_features* (número máximo de características utilizadas em cada árvore da floresta) e *n_estimator* (número de árvores na floresta),

ao avaliar a classificação podemos observar que em média houve uma melhora de 0.11 para a acurácia e 0.25 para o kappa (Tabela 17). Apesar de não ter sido encontrado trabalhos, em mensagens na língua portuguesa para a presença social, que houvessem conduzido uma análise parecida para que pudesse ser realizada uma comparação, é interessante mencionar que os resultados alcançados neste trabalho obteve resultados melhores aos desenvolvidos para o inglês em (FERREIRA et al., 2020) e aos de (NETO et al., 2018), este, voltado para a presença cognitiva em mensagens em português.

Vale ressaltar que o presente trabalho utilizou três algoritmos bastante utilizados na literatura, o Random Forest, AdaBoost e XGBoost,diferente dos trabalhos citados a pouco que utilizaram apenas o Random Forest. Eesse diferencial que proporcionou os melhores resultados pois o melhor algoritmo, no geral, foi o XGBoost.

5.8.2 Análise das Características Importantes

Retomando a Questão de Pesquisa 02, apresentada na seção 1.2, este estudo conduziu uma análise detalhada das características mais relevantes para classificar cada categoria. Ao observar os três cenários, de forma geral, houve uma ocorrência de 50% de características oriundas da frequência de palavras, 32% do Coh-Metrix e 18% do LIWC. Em cada categoria, houve ocorrência de características relacionadas aos três conjuntos mas com uma maioria de características do LIWC e Coh-Metrix nas primeiras posições, ou seja, com um ganho de informação considerável. Vale ressaltar que as caracteristicas relacionadas com a frequência de palavras que tiveram maior pontuação MDG foram palavras que podem ocorrer em diferentes domínios, como por exemplo boa, turma, ola, noit, devido a isso as chances de ocorrer *overfitting* são menores. A seguir serão detalhadas as características do cenário I e uma análise comparativa entre as características de domínios diferentes (cenário II e III).

5.8.2.1 Análise das Características no *Corpus* (BioBase + TecBase)

De forma geral, as características mais relevantes apresentadas nas Tabelas 28, 29 e 30 estão alinhadas com a teoria da presença social proposta por (GARRISON et al., 1999). Por exemplo, podemos destacar mensagens que: i) palavras que expressam emoções positivas; ii) quantidade de pronomes em segunda pessoa; iii) palavras que expressam concordância; iv) palavras que fazem referência a inclusão; v) palavras que indicam saudações.

Especificamente na categoria Afetiva, as características *liwc.posemo* (número de palavras que expressam emoções) e *bem* (frequência de palavras), estão ligadas com os indicadores

de expressão de emoções e uso de humor. Isso revela uma possível correlação entre essas características identificadas e os indicadores considerados mais preditivos de presença social (ROURKE et al., 1999). Comparando com as características encontradas em (FERREIRA et al., 2020), houve uma coincidência de 10%, pois no trabalho os autores também encontraram as características *liwc.posemo* e *liwc.negate*.

Para a categoria Interativa, as características *liwc.you* (quantidade de pronomes em segunda pessoa), *liwc.incl* (número de palavras que fazem referência a inclusão) e *cm.WRDPRP2* (número de pronomes em segunda pessoa) estão relacionadas ao indicador que cita ou referencia outras mensagens ou pessoas na discussão (ver Tabela **??**). A característica *concor* (concordo), originada da frequência de palavras, está relacionada ao indicador que expressa concordância com os outros ou suas mensagens confirmando a interação entre os alunos com base no modelo CoI (ROURKE et al., 1999). Nesta categoria houve uma coincidência de 15% em relação as características encontradas no trabalho de (FERREIRA et al., 2020), são elas: *concor* (concordo), *cm.WRDPRP2* e *liwc.you*.

Por fim, para a categoria Coesiva, a característica *boa* (frequência de palavras) foi a mais relevante, a maior pontuação dentre todas as características das demais categorias, ela está relacionada com o indicador de comunicação social/saudações. Outra característica interessante é a *coleg* (colega), levando a uma percepção de que poderia está ocorrendo uma harmonia entre os participantes do fórum. Não houve características em comum em relação ao trabalho de (FERREIRA et al., 2020), para esta categoria.

5.8.2.2 Análise das Características das Bases de Dados de Domínios Diferentes (BioBase Vs. TecBase)

Ao verificar-se os resultados referentes as bases de dados de Biologia e Tecnologia,conseguimos identificar que houve cerca de 16.6% de coincidência entre as características. Comparando-as com o *corpus*, na BioBase houve uma coincidência de 33.3% e, na TecBase , 21.6% com uma distribuição equilibrada entre características do LIWC, Coh-Metrix e frequência de palavras.

Em relação as diferenças, na categoria Afetiva foi observado que na na BioBase existem ocorrências de palavras que expressam emoções negativas, grande número de conectivos, palavras que expressam intuição e adjetivos. Já na TecBase além de não serem percebidas características condizentes com os indicadores propostos pela literatura, não houve nenhuma ocorrência de características oriundas do LIWC demonstrando a falta de mensagens de cunho emocional, por exemplo. Ao observar algumas características das bases, temos que, no curso de tecnologia

a maioria dos alunos são do sexo masculino (78%), por outro lado, no curso de Biologia a predominância é de alunos do sexo feminino (70%). Então, uma possível explicação é de que a ocorrência de expressão de emoções e humor nesses ambientes, diferem de acordo com o gênero dos alunos. Em (GARRISON; ARBAUGH, 2007), o autor comenta possíveis diferenças na forma como os alunos do sexo masculino e feminino se comunicam.

Ao observar a categoria Interativa, ambas as bases possuem características condizentes com os indicadores da categoria como por exemplo, palavras que expressam concordância (concord) e número de pronomes em segunda pessoa (cm.WRDPRP2). Mas na base de biologia existem mais indicativos da categoria em questão, como as características *cm.DESWC* (quantidade total de palavras), *liwc.you* (quantidade de pronomes em segunda pessoa), *liwc.article* (quantidade de artigos) e *cm.WRDNOUN* (quantidade de substantivos). A suposição anterior também cabe nessa constatação pois a predominância de alunos do sexo feminino pode resultar em uma maior interatividade. De acordo com (ROVAI, 2001), as mulheres são mais interativas do que os homens, pois procuram construir um senso de comunidade online, estimulando mais participações e interações nos fóruns, como verificado na BioBase.

Por fim, ao analisar as características da categoria Coesiva em ambas as bases de dados, constatou-se que as características com maior pontuação tem origem a frequência de palavras,por exemplo, *ola, boa, noit, bom, turm, tard*. Essas características são ligadas diretamente com o indicador da categoria Coesiva que diz respeito a comunicação social, saudações e despedidas. Vale ressaltar que as características destacadas anteriormente são independentes do domínio, isso explica o porque em praticamente todos os resultados apresentados em cada cenário os modelos referentes a categoria Coesiva apresentam resultados satisfatórios (Tabelas 18, 20, 22, 24 e 26).

6 CONSIDERAÇÕES FINAIS

Esta dissertação de mestrado teve como principal objetivo desenvolver e validar um método para automatizar a identificação da presença social em mensagens, oriundas de fóruns de discussões, escritas em português do Brasil, através de técnicas de Mineração de Texto e utilizando-se de recursos como o LIWC e Coh-Metrix. Existem três contribuições principais. Primeiro, foram realizados diversos experimentos incluindo uma etapa de otimização dos parâmetros do classificador. No principal cenário onde foi utilizado o *corpus*, foram desenvolvidos três classificadores, um para cada categoria da presença social, onde o algoritmo XGBoost obteve os melhores resultados. Os modelos desenvolvidos atingiram uma acurácia de 0.92, 0.85 e 0.99 e k de 0.49, 0.71 e 0.98 para as categorias, Afetiva, Interativa e Coesiva respectivamente. Segundo (LANDIS; KOCH, 1977), esses valores de k são considerados uma concordância de moderada (0.49), substancial (0.71) e quase perfeita (0.98). Vale destacar que este trabalho fez uso de três tipos de algoritmos, Random Forest, AdaBoost e XGBoost, com o objetivo de melhorar os resultados de cada modelo diferenciando-o sua metodologia dos demais trabalhos apresentados.

Como segunda contribuição temos a análise detalhada da importância das características propostas, baseadas tanto nos recursos Coh-Metrix e LIWC quanto na frequência de palavras comumente utilizada em técnicas de Mineração de Texto. No contexto analisado, observou-se que houve uma distribuição equilibrada das características entre os recursos mencionados mas que a predominância das características com maior puntação MDG são oriundas co Coh-Metrix e LIWC com exceção da categoria Coesiva onde um grupo pequeno de características da frequência de palavras se sobressaem.

A terceira contribuição é a analise sobre as principais diferenças relacionadas a presença social em dois contextos diferentes: Biologia (BioBase) e Tecnologia (TecBase). De forma geral, na base de biologia foram utilizadas as características dos três recursos de forma proporcional, 17 do LIWC, 23 do Coh-Metrix e 20 de frequência de palavras. Por outro lado na base de Tecnologia teve uma baixa ocorrência de características do LIWC, apenas 3, seguido de 22 do Coh-Metrix e 35 de frequência de palavras. Em relação a categoria Coesiva não foram constatadas diferenças que chamassem a atenção, diferente das categorias Afetiva e Interativa, onde observou-se que possivelmente a diferença da predominância de alunos do sexo feminino pode acarretar em uma maior interatividade e expressões de emoções do que em ambientes com o número maior de alunos do sexo masculino.

Sendo assim, através do abordagem proposta neste trabalho para automatizar a identifica-

ção das categorias da presença social para a língua portuguesa, espera-se tornar o processo de codificação mais fácil, contribuindo para a percepção do professor/tutor em relação a capacidade dos alunos se projetarem social e emocionalmente, em sua personalidade completa, no meio de comunicação utilizado. Sendo assim, através dessa percepção os professores tutores podem aplicar estratégias que afetem os resultados de aprendizagem dos alunos.

6.1 ARTIGO SUBMETIDO/ACEITO

Com o objetivo de divulgar este trabalho, um artigo foi submetido e aceito para o Simpósio Brasileiro de Informática na Educação (SBIE 2020).

6.2 LIMITAÇÕES DA PESQUISA

Dentre as diversas limitações encontradas neste trabalho, podemos destacar o problema com o tamanho pequeno da base de dados utilizadas e as categorias desbalanceadas, apesar de que essa situação reflita com as encontradas na literatura, podem afetar o desempenho do classificador. Devido a esse problema foram utilizadas todas as mensagens presentes nas duas base de dados, incluindo as mensagens dos professores/tutores. Outro problema é que ambas as bases continham apenas quatro semanas de interação, podendo comprometer a quantidade de indicadores da presença social presentes nas mensagens visto que leva tempo para que os alunos sintam-se a vontade para iniciar interações por conta própria.

Outro ponto é sobre a dificuldade de se encontrar recursos voltados para a análise de textos em português, principalmente, que não sejam apenas uma análise estatística. Apesar dos resultados terem sido coerentes com as teorias do modelo CoI, a análise realizada neste trabalho não estende-se aos aspectos semânticos.

6.3 TRABALHOS FUTUROS

Como trabalhos futuro pretende-se:

- Incluir outros recursos e técnicas disponíveis para a língua inglesa, por exemplo, outros dicionários;
- Testar a generalização dos classificadores desenvolvidos, em outros contextos educacionais. Por exemplo graduação vs. pós-graduação;

- Aplicar a metodologia utilizada neste trabalho em base de dados de diferentes domínios dos utilizados (biologia e tecnologia);
- Verificar a eficácia das características recomendadas nesta pesquisa para outras línguas, o espanhol por exemplo.
- Excluir as características referentes a técnica de BoW para verificar se existe alteração no desempenho dos classificadores.

REFERÊNCIAS

ABED, A. B. de Educação a D. Censo ead.br: relatório analítico da aprendizagem a distância no brasil 2018. In: . Brasil: [s.n.], 2018.

AHMED, N. S.; SADIQ, M. H. Clarify of the random forest algorithm in an educational field. In: IEEE. **2018 International Conference on Advanced Science and Engineering (ICOASE)**. [S.l.], 2018. p. 179–184.

ANDERSON, T.; LIAM, R.; GARRISON, D. R.; ARCHER, W. Assessing teaching presence in a computer conferencing context. Journal of the Asynchronous Learning Network, 2001.

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

ARANHA, C. N.; VELLASCO, M.; PASSOS, E. Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional. **Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ**, p. 33–34, 2007.

AZEVEDO, B. F. T. Minerafórum: um recurso de apoio para análise qualitativa em fóruns de discussão. 2011.

BALI, M.; GORE, D. A survey on text classification with different types of classification methods. **International Journal of Innovative Research in Computer and Communication Enginearing**, v. 3, p. 4888–4894, 2015.

BARBOSA, G.; CAMELO, R.; CAVALCANTI, A. P.; MIRANDA, P.; MELLO, R. F.; KOVANOVIĆ, V.; GAŠEVIĆ, D. Towards automatic cross-language classification of cognitive presence in online discussions. In: **Proceedings of the Tenth International Conference on Learning Analytics & Knowledge**. [S.l.: s.n.], 2020. p. 605–614.

BASTOS, H. P. P.; BERCHT, M.; WIVES, L. K. Presença social em cursos a distância: Um estudo comparativo de postagens em chats e fóruns. **RENOTE-Revista Novas Tecnologias na Educação**, v. 8, n. 3, 2010.

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**, Wiley Online Library, v. 58, n. 2, p. 329–331, 2007.

BECHT, E.; MCINNES, L.; HEALY, J.; DUTERTRE, C.-A.; KWOK, I. W.; NG, L. G.; GINHOUX, F.; NEWELL, E. W. Dimensionality reduction for visualizing single-cell data using umap. **Nature biotechnology**, Nature Publishing Group, v. 37, n. 1, p. 38–44, 2019.

BECKMANN, M. **Algoritmos genéticos como estratégia de pré-processamento em conjuntos de dados desbalanceados**. [S.l.]: Master's thesis, Programa de Pós-Graduação em Engenharia Civil-COPPE-UFRJ, 2010.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

BOLLIGER, D. U.; INAN, F. A. Development and validation of the online student connectedness survey (oscs). **International Review of Research in Open and Distributed Learning**, Athabasca University Press (AU Press), v. 13, n. 3, p. 41–65, 2012.

BOULIS, C.; OSTENDORF, M. Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. In: CITESEER. **Proc. of the International Workshop in Feature Selection in Data Mining**. [S.l.], 2005. p. 9–16.

BOZKURT, A.; AKGUN-OZBEK, E.; YILMAZEL, S.; ERDOGDU, E.; UCAR, H.; GULER, E.; SEZGIN, S.; KARADENIZ, A.; SEN-ERSOY, N.; GOKSEL-CANBEK, N. et al. Trends in distance education research: A content analysis of journals 2009-2013. **International Review of Research in Open and Distributed Learning**, Athabasca University Press (AU Press), v. 16, n. 1, p. 330–363, 2015.

BRASIL, C. C. Decreto n. 5.622, de 19 de dezembro de 2005. **Acesso em 11/05/19**, v. 1, 2005.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BUDA, M.; MAKI, A.; MAZUROWSKI, M. A. A systematic study of the class imbalance problem in convolutional neural networks. **Neural Networks**, Elsevier, v. 106, p. 249–259, 2018.

BURT, R. S. Closure as social capital. **Social capital: Theory and research**, Aldine de Gruyter New York, NY, p. 31–55, 2001.

BURT, R. S. et al. The social capital of structural holes. **The new economic sociology: Developments in an emerging field**, v. 148, n. 90, p. 122, 2002.

CAMBRIDGE, U. Introduction to information retrieval. 2009.

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

CARLETTA, J. Assessing agreement on classification tasks: the kappa statistic. **arXiv preprint cmp-lg/9602004**, 1996.

CAROLAN, B. V. **Social network analysis and education: Theory, methods & applications**. [S.l.]: Sage Publications, 2013.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Special issue on learning from imbalanced data sets. **ACM SIGKDD explorations newsletter**, ACM New York, NY, USA, v. 6, n. 1, p. 1–6, 2004.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794.

CHOWDHARY, K. Natural language processing. In: **Fundamentals of Artificial Intelligence**. [S.l.]: Springer, 2020. p. 603–649.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.

CUNHA, A. L. V. d. **Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais**. Tese (Doutorado) — Universidade de São Paulo, 2015.

EFRON, B. Bootstrap methods: another look at the jackknife. In: **Breakthroughs in statistics**. [S.l.]: Springer, 1992. p. 569–593.

FARZI, R.; BOLANDI, V. Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: A case study of the south pars gas field, persian gulf, iran. **Modeling Earth Systems and Environment**, Springer, v. 2, n. 2, p. 105, 2016.

FAYRAM, J. **The nature and role of social presence in audiographic, synchronous online language learning contexts**. Tese (Doutorado) — The Open University, 2017.

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

FERNÁNDEZ, A.; GARCÍA, S.; GALAR, M.; PRATI, R. C.; KRAWCZYK, B.; HERRERA, F. **Learning from imbalanced data sets**. [S.l.]: Springer, 2018.

FERNÁNDEZ-DELGADO, M.; CERNADAS, E.; BARRO, S.; AMORIM, D. Do we need hundreds of classifiers to solve real world classification problems? **The Journal of Machine Learning Research**, JMLR. org, v. 15, n. 1, p. 3133–3181, 2014.

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

FILHO, P. B.; PARDO, T. A. S.; ALUÍSIO, S. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013.

FINATTO, M. J. B. Complexidade textual em artigos científicos: contribuições para o estudo do texto científico em português. **Organon**, v. 25, n. 50, 2011.

FRAU-MEIGS, D.; BOSSU, A. Towards e-presence at distance as a way to reach and share e-quality: The case of the eco smoocs. In: SPRINGER. **European Conference on Massive Open Online Courses**. [S.l.], 2017. p. 38–47.

FREUND, Y. An improved boosting algorithm and its implications on learning complexity. In: **Proceedings of the fifth annual workshop on Computational learning theory**. [S.l.: s.n.], 1992. p. 391–398.

GARRISON, D. R. **E-learning in the 21st century: A framework for research and practice**. [S.l.]: Taylor & Francis, 2011.

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

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. Critical thinking, cognitive presence, and computer conferencing in distance education. **American Journal of distance education**, Taylor & Francis, v. 15, n. 1, p. 7–23, 2001.

GARRISON, D. R.; ARBAUGH, J. B. Researching the community of inquiry framework: Review, issues, and future directions. **The Internet and higher education**, Elsevier, 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**, Elsevier, 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**, Taylor & Francis, v. 3, n. 1, p. 63–78, 2017.

GENUER, R.; POGGI, J.; TULEAU-MALOT, C. Variable selection using random forests pattern recognition letters, 31, 2225 doi: 10.1016. **J. PATREC**, v. 14, 2010.

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, 2012, Lisboa: ticEDUCA**. [S.l.: s.n.], 2012. p. 60–70.

GONÇALVES, T.; SILVA, C.; QUARESMA, P.; VIEIRA, R. Analysing part-of-speech for portuguese text classification. In: SPRINGER. **International Conference on Intelligent Text Processing and Computational Linguistics**. [S.l.], 2006. p. 551–562.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016.

GOODWIN; MILLER. Evidence on flipped classrooms is still coming in. **Educational Leadership**, 2013.

GOOS, M. Learning mathematics in a classroom community of inquiry. **Journal for research in mathematics education**, JSTOR, p. 258–291, 2004.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-metrix: Providing multilevel analyses of text characteristics. **Educational researcher**, Sage Publications Sage CA: Los Angeles, CA, v. 40, n. 5, p. 223–234, 2011.

GRAESSER, A. C.; MCNAMARA, D. S.; LOUWERSE, M. M.; CAI, Z. Coh-metrix: Analysis of text on cohesion and language. **Behavior research methods, instruments, & computers**, Springer, v. 36, n. 2, p. 193–202, 2004.

GRAçAS, B. Maria das; GOMES, C. A. B. As concepções de interatividade nos ambientes virtuais de aprendizagem. **Campina Grande: EDUEPB**, 2011.

GU, Q.; CAI, Z.; ZHU, L.; HUANG, B. Data mining on imbalanced data sets. In: IEEE. **2008 International Conference on Advanced Computer Theory and Engineering**. [S.l.], 2008. p. 1020–1024.

HAN, J.; KAMBER, M.; PEI, J. Data mining concepts and techniques third edition. **The Morgan Kaufmann Series in Data Management Systems**, p. 83–124, 2011.

HART, C. Factors associated with student persistence in an online program of study: A review of the literature. **Journal of Interactive Online Learning**, v. 11, n. 1, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer Science & Business Media, 2009.

HE, X.; PAN, J.; JIN, O.; XU, T.; LIU, B.; XU, T.; SHI, Y.; ATALLAH, A.; HERBRICH, R.; BOWERS, S. et al. Practical lessons from predicting clicks on ads at facebook. In: **Proceedings of the Eighth International Workshop on Data Mining for Online Advertising**. [S.l.: s.n.], 2014. p. 1–9.

HOLSTEIN, S.; COHEN, A. The characteristics of successful moocs in the fields of software, science, and management, according to students' perception. **Interdisciplinary Journal of e-Skills and Lifelong Learning**, Informing Science Institute. 131 Brookhill Court, Santa Rosa, CA 95409, v. 12, p. 247–266, 2016.

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

INDURKHYA, N.; DAMERAU, F. J. **Handbook of natural language processing**. [S.l.]: CRC Press, 2010. v. 2.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. **Intelligent data analysis**, IOS Press, v. 6, n. 5, p. 429–449, 2002.

JIANG, M.; LIANG, Y.; FENG, X.; FAN, X.; PEI, Z.; XUE, Y.; GUAN, R. Text classification based on deep belief network and softmax regression. **Neural Computing and Applications**, Springer, v. 29, n. 1, p. 61–70, 2018.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, 1972.

JUNG, Y. Multiple predicting k-fold cross-validation for model selection. **Journal of Nonparametric Statistics**, Taylor & Francis, v. 30, n. 1, p. 197–215, 2018.

KIM, J. Developing an instrument to measure social presence in distance higher education. **British Journal of Educational Technology**, Wiley Online Library, v. 42, n. 5, p. 763–777, 2011.

KLAHOLD, A.; FATHI, M. Knowledge discovery from text (kdt). In: **Computer Aided Writing**. [S.l.]: Springer, 2020. p. 83–115.

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

KOVANOVIC, V.; GASEVIC, D.; HATALA, M. Learning analytics for communities of inquiry. **Journal of Learning Analytics**, v. 1, n. 3, p. 195–198, 2014.

KOVANOVIC, V.; JOKSIMOVIC, S.; GASEVIC, D.; HATALA, M. What is the source of social capital? the association between social network position and social presence in communities of inquiry. Citeseer, 2014.

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

KOWSARI, K.; MEIMANDI, K. J.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D. Text classification algorithms: A survey. **Information**, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 150, 2019.

KOZAN, K. A comparative structural equation modeling investigation of the relationships among teaching, cognitive and social presence. **Online Learning**, ERIC, v. 20, n. 3, p. 210–227, 2016.

KRAMER, O. Scikit-learn. In: **Machine learning for evolution strategies**. [S.l.]: Springer, 2016. p. 45–53.

KUDO, T.; RICHARDSON, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. **arXiv preprint arXiv:1808.06226**, 2018.

KUDO, T.; YAMAMOTO, K.; MATSUMOTO, Y. Applying conditional random fields to japanese morphological analysis. In: **Proceedings of the 2004 conference on empirical methods in natural language processing**. [S.l.: s.n.], 2004. p. 230–237.

KUNCHEVA, L. I. **Combining pattern classifiers: methods and algorithms**. [S.l.]: John Wiley & Sons, 2014.

LAMBERT, J.; FISHER, J. Community building in a wiki-based distance education course. In: ASSOCIATION FOR THE ADVANCEMENT OF COMPUTING IN EDUCATION (AACE). **EdMedia+ Innovate Learning**. [S.l.], 2009. p. 1527–1531.

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

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

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LIPMAN, M. **Thinking in education**. [S.l.]: Cambridge University Press, 2003.

MACHADO, A.; LONGHI, M.; NUNES, M. A. S. N.; PARDO, T. Personalitatem lexicon: Um léxico em português brasileiro para mineraçao de traços de personalidade em textos. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2015. v. 26, n. 1, p. 1122.

MARNEFFE, M.-C. D.; MACCARTNEY, B.; MANNING, C. D. et al. Generating typed dependency parses from phrase structure parses. In: **Lrec**. [S.l.: s.n.], 2006. v. 6, p. 449–454.

MARTIN, J. R.; WHITE, P. R. **The language of evaluation**. [S.l.]: Springer, 2003. v. 2.

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

MCKERLICH, R.; ANDERSON, T. Community of inquiry and learning in immersive environments. **Journal of asynchronous learning networks**, v. 11, n. 4, 2007.

MENTCH, L.; HOOKER, G. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. **The Journal of Machine Learning Research**, JMLR. org, v. 17, n. 1, p. 841–881, 2016.

MESSICK, S. Standards of validity and the validity of standards in performance asessment. **Educational measurement: Issues and practice**, Wiley Online Library, v. 14, n. 4, p. 5–8, 1995.

MOISEY, S.; ARTHUR, P.; GARRISON, D. R.; GRAHAM, C. et al. A thematic synthesis of community of inquiry research 2000 to 2014. 2016.

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

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in neurorobotics**, Frontiers, v. 7, p. 21, 2013.

NETO, V.; ROLIM, V.; FERREIRA, R.; KOVANOVIĆ, V.; GAŠEVIĆ, D.; LINS, R. D.; LINS, R. Automated analysis of cognitive presence in online discussions written in portuguese. In: SPRINGER. **European conference on technology enhanced learning**. [S.l.], 2018. p. 245–261.

NORVIG, P.; RUSSELL, S. **Inteligência artificial: Traduçao da 3a ediçao (Vol. 1)**. [S.l.]: Elsevier Brasil, 2014.

ONAN, A. Classifier and feature set ensembles for web page classification. **Journal of Information Science**, SAGE Publications Sage UK: London, England, v. 42, n. 2, p. 150–165, 2016.

PALLOFF, R. M.; PRATT, K. **O aluno virtual-um guia para trabalhar com estudantes on-line**. [S.l.]: Penso Editora, 2004.

PENG, Y.; KOU, G.; CHEN, Z.; SHI, Y. Cross-validation and ensemble analyses on multiple-criteria linear programming classification for credit cardholder behavior. In: SPRINGER. **International Conference on Computational Science**. [S.l.], 2004. p. 931–939.

PENNEBAKER, J. W.; BOOTH, R. J.; FRANCIS, M. E. Linguistic inquiry and word count: Liwc [computer software]. **Austin, TX: liwc. net**, v. 135, 2007.

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

PEREZ, A. F.; BASSOLI, D. A.; LOPES, C. S. G.; NETO, J. D. de O.; CAZARINI, E. W. Identificação da presença social em curso a distância de capacitação docente para ead. **SIED: EnPED-Simpósio Internacional de Educação a Distância e Encontro de Pesquisadores em Educação a Distância 2012**, 2012.

PHUA, C.; ALAHAKOON, D.; LEE, V. Minority report in fraud detection: classification of skewed data. **Acm sigkdd explorations newsletter**, ACM New York, NY, USA, v. 6, n. 1, p. 50–59, 2004.

PRATI, R. C.; BATISTA, G.; MONARD, M. C.; SAO-CARLENSE, A. do T.; POSTAL, C.-C. Uma experiência no balanceamento artificial de conjuntos de dados para aprendizado com classes desbalanceadas utilizando análise roc. In: **Proceedings of IV Workshop on Advances & Trends in AI for Problem Solving, Chilán**. [S.l.: s.n.], 2003.

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**, v. 2, n. 12, p. 4536–4538, 2013.

RAMSDEN, P. **Learning to teach in higher education**. [S.l.]: Routledge, 2003.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Embrapa Informática Agropecuária-Artigo em periódico indexado (ALICE)**, Revista de Sistema de Informação da FSMA, Macaé, n. 7, p. 7-21, 2011., 2011.

RICHARDS, K. A. R.; VELASQUEZ, J. D. First-year students' perceptions of instruction in large lectures: The top-10 mistakes made by instructors. **Journal on Excellence in College Teaching**, v. 25, n. 2, 2014.

RICHARDSON, J. C.; KOEHLER, A. A.; BESSER, E. D.; CASKURLU, S.; LIM, J.; MUELLER, C. M. Conceptualizing and investigating instructor presence in online learning environments. **The International Review of Research in Open and Distributed Learning**, v. 16, n. 3, 2015.

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for idf. **Journal of documentation**, Emerald Group Publishing Limited, 2004.

ROKACH, L. Ensemble-based classifiers. **Artificial intelligence review**, Springer, v. 33, n. 1-2, p. 1–39, 2010.

ROLIM, V.; FERREIRA, R.; KOVANOVIC, V.; GASEVIC, D. Analysing social presence in online discussions through network and text analytics. 2019 no prelo.

ROSSI, R. G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. Tese (Doutorado) — Universidade de São Paulo, 2016.

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

ROURKE, L.; ANDERSON, T.; GARRISON, D. R.; ARCHER, W. Methodological issues in the content analysis of computer conference transcripts. 2001.

ROVAI, A. P. Building classroom community at a distance: A case study. **Educational technology research and development**, Springer, v. 49, n. 4, p. 33, 2001.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information processing & management**, Elsevier, v. 24, n. 5, p. 513–523, 1988.

SAUDE, S.; PUTEH, F.; AZIZAN, A. R.; HAMDAN, N. N.; SHUKOR, N. H. A.; ABDULLAH, K. I. Learning through the lounge: Using social presence to assess the learning environment in a myline online forum. **Procedia-Social and Behavioral Sciences**, Elsevier, v. 66, p. 448–459, 2012.

SCARTON, C.; GASPERIN, C.; ALUISIO, S. Revisiting the readability assessment of texts in portuguese. In: SPRINGER. **Ibero-American Conference on Artificial Intelligence**. [S.l.], 2010. p. 306–315.

SCHAPIRE, R. E. The strength of weak learnability. **Machine learning**, Springer, v. 5, n. 2, p. 197–227, 1990.

SCHÜTZE, H.; MANNING, C. D.; RAGHAVAN, P. **Introduction to information retrieval**. [S.l.]: Cambridge University Press Cambridge, 2008. v. 39.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 34, n. 1, p. 1–47, 2002.

SHAFFER, D. W.; COLLIER, W.; RUIS, A. R. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. **Journal of Learning Analytics**, v. 3, n. 3, p. 9–45, 2016.

SHAH, F. P.; PATEL, V. A review on feature selection and feature extraction for text classification. In: IEEE. **2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)**. [S.l.], 2016. p. 2264–2268.

SHIELDS, P. M. The community of inquiry: Classical pragmatism and public administration. **Administration & Society**, Sage Publications, v. 35, n. 5, p. 510–538, 2003.

SILVA, J. K. K. d. Automatização do processo de identificação de presença social em fóruns e chats. 2011.

SOARES, F. B. M.; MACHADO, C. J. R.; DINIZ, D.; MACIEL, A. M. A. Educational data mining to support distance learning students with difficulties in the portuguese grammar. In: **Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016)**. Brasil: [s.n.], 2016. p. 956–965.

SPILIOPOULOS, V.; VOUROS, G. A.; KARKALETSIS, V. On the discovery of subsumption relations for the alignment of ontologies. **Journal of Web Semantics**, Elsevier, v. 8, n. 1, p. 69–88, 2010.

STONE, M. Cross-validatory choice and assessment of statistical predictions. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 2, p. 111–133, 1974.

STRIJBOS, J.-W.; MARTENS, R. L.; PRINS, F. J.; JOCHEMS, W. M. Content analysis: What are they talking about? **Computers & Education**, Elsevier, v. 46, n. 1, p. 29–48, 2006.

SUHANG, J.; WILLIAMS, A.; SCHENKE, K.; WARSCHAUER, M.; ODOWD, D. Predicting mooc performance with week 1 behavior. **Educational Data Mining**, 2014.

SWAN, K. Building learning communities in online courses: The importance of interaction. **Education, Communication & Information**, Taylor & Francis, v. 2, n. 1, p. 23–49, 2002.

TANIGUCHI, Y.; KONOMI, S.; GODA, Y. Examining language-agnostic methods of automatic coding in the community of inquiry framework. In: IADIS PRESS. **16th International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2019**. [S.l.], 2019. p. 19–26.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. **Journal of language and social psychology**, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 24–54, 2010.

TURNER, V.; GANTZ, J. F.; REINSEL, D.; MINTON, S. The digital universe of opportunities: Rich data and the increasing value of the internet of things. **IDC Analyze the Future**, v. 16, 2014.

VARIAN, H. R. Big data: New tricks for econometrics. **Journal of Economic Perspectives**, v. 28, n. 2, p. 3–28, 2014.

VIEIRA, R.; LIMA, V. L. Lingüística computacional: princípios e aplicações. In: SN. **Anais do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial**. [S.l.], 2001. v. 3, p. 47–86.

WEISS, G. M. Mining with rarity: a unifying framework. **ACM Sigkdd Explorations Newsletter**, ACM New York, NY, USA, v. 6, n. 1, p. 7–19, 2004.

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.

WITTEN, e. f. ian h.; HELL mark a. Data mining: practical machine learning tools and technique, by ian h. witten, eibe frank, mark a. hell. **ACM SIGSOFT Software Engineering Notes**, ACM New York, NY, USA, v. 36, n. 5, p. 51–52, 2011.

WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with java implementations. **Acm Sigmod Record**, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002.

WIVES, L. K. Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de"clustering". 1999.

WIVES, L. K.; BERCHT, M.; BASTOS, H. P. P. Análise manual e automática de pistas lexicais de presença social em chat educacional. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2010. v. 1, n. 1.

YU, F.; ZHENG, D. Education data mining: How to mine interactive text in moocs using natural language process. In: IEEE. **2017 12th International Conference on Computer Science and Education (ICCSE)**. [S.l.], 2017. p. 694–699.

ZHANG, D.; QIAN, L.; MAO, B.; HUANG, C.; HUANG, B.; SI, Y. A data-driven design for fault detection of wind turbines using random forests and xgboost. **IEEE Access**, IEEE, v. 6, p. 21020–21031, 2018.

ZHAO, R.; MAO, K. Fuzzy bag-of-words model for document representation. **IEEE transactions on fuzzy systems**, IEEE, v. 26, n. 2, p. 794–804, 2017.

ZIEGLER, A.; KÖNIG, I. R. Mining data with random forests: current options for real-world applications. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 4, n. 1, p. 55–63, 2014.