UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Vitor Belarmino Rolim

Análise Automatizada de Discussões Online Baseada no Framework de Comunidade de Investigação: Classificação da Presença Cognitiva com Técnicas de Processamento de Linguagem Natural

Recife

2024

Vitor Belarmino Rolim

Análise Automatizada de Discussões Online Baseada no Framework de Comunidade de Investigação: Classificação da Presença Cognitiva com Técnicas de Processamento de Linguagem Natural

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Rafael Dueire Lins

Coorientador (a): Rafael Ferreira Leite de Melo

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Rolim, Vitor Belarmino.
Análise automatizada de discussões online baseada no framework de comunidade de investigação: classificação da presença cognitiva com técnicas de processamento de linguagem natural / Vitor Belarmino Rolim. - Recife, 2024.
117f.: il.

Inclui referências.
Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Informática, Pós-graduação em Ciência da Computação, 2024.
Orientação: Rafael Dueire Lins.
Coorientação: Rafael Ferreira Leite de Mello.

1. Comunidade de investigação; 2. Fóruns de discussão; 3. Presença cognitiva; 4. Analíticas de aprendizagem; 5. Processamento de linguagem natural; 6. Aprendizado ativo. I. Lins, Rafael Dueire. II. Mello, Rafael Ferreira Leite de. III. Título.

UFPE-Biblioteca Central

Folha de aprovação: Inserir a folha de aprovação enviada pela Secretaria do curso de Pós-Graduação. A folha deve conter a **data de aprovação**, estar **sem assinaturas** e em formato **PDF**.

Dedico este trabalho à minha esposa Laís, aos meus pais e a todos os meus familiares, cujo amor e apoio foram minha inspiração e fortaleza.

AGRADECIMENTOS

Primeiramente, gostaria de expressar minha profunda gratidão à minha esposa, Laís, cujo amor e apoio não apenas me mantiveram de pé durante este desafio acadêmico, mas também enriqueceram cada dia da minha jornada. Seu companheirismo foi meu refúgio e sua fé em mim, uma fonte constante de motivação.

Aos meus queridos pais, Flávio e Ilma, ofereço meu eterno agradecimento. O amor incondicional e o esforço incansável de vocês para me proporcionar uma educação de qualidade foram a fundação sobre a qual construí minhas aspirações e realizações. Cada sacrifício que vocês fizeram por mim não passou despercebido e é profundamente valorizado.

Um agradecimento especial à minha irmã, Flávia, e ao meu cunhado, Junior, pelos momentos de descontração e encorajamento. As pausas que compartilhamos e o apoio que vocês me ofereceram foram essenciais para manter meu ânimo e foco.

Não posso deixar de agradecer imensamente a meus orientadores, Rafael Dueire e Rafael Ferreira. A confiança que vocês depositaram em mim e o apoio constante nos momentos mais desafiadores foram cruciais para minha perseverança e sucesso. Vocês não apenas acreditaram em mim nos momentos de desmotivação, como também guiaram habilmente cada passo do meu desenvolvimento acadêmico com paciência e sabedoria.

Por fim, um agradecimento a todos os amigos e familiares que, de várias formas, contribuíram para minha jornada. Seja através de palavras de incentivo ou de atos de cuidado, cada um de vocês teve um papel indispensável em me ajudar a alcançar este objetivo.

A todos vocês, meu sincero obrigado.

RESUMO

Com o estabelecimento do ensino a distância como modelo educacional, diversas ferramentas foram desenvolvidas com o objetivo de proporcionar uma experiência de ensino semelhante à do ensino presencial. Entre essas ferramentas, destacam-se os fóruns de discussão, que oferecem aos alunos um ambiente para construção de conhecimento. Técnicas de aprendizagem de máquina vêm sendo empregadas para fornecer classificações dos níveis de desenvolvimento cognitivo dos alunos, baseadas nas interações ocorridas nos fóruns educacionais. A criação desses classificadores depende de diversos aspectos para aumentar a acurácia dos modelos treinados; contudo, esses modelos são altamente dependentes da quantidade e qualidade dos dados. A anotação desses dados é um trabalho intensivo que depende de especialistas de domínio e, além disso, há uma escassez de dados devido à dificuldade de aquisição de dados educacionais. Este trabalho investiga a aplicação de técnicas de aprendizagem de máquina para a análise automatizada de discussões online em AVAs, utilizando o modelo CoI. Explora-se a viabilidade de métodos automáticos para a identificação da presença cognitiva em fóruns de discussão, visando entender e otimizar a construção de conhecimento em contextos educacionais à distância. Foram utilizados diferentes modelos de aprendizado de máquina, incluindo Random Forest, XGBoost, MLP, além de abordagens de aumento de dados com BERT e GPT-4 para lidar com o desbalanceamento das categorias da presença cognitiva. As características textuais foram extraídas utilizando ferramentas como LIWC, Coh-Metrix e SNA, proporcionando uma representação abrangente das interações e conteúdos discutidos, além de modelos de linguagem focados em codificação, como o DeBERTa. Os resultados mostram que a combinação de técnicas de aprendizado ativo tem grande potencial para o problema abordado, considerando a limitação de dados na área educacional, especialmente em relação aos dados anotados. Conseguimos atingir um coeficiente de Cohen's Kappa de 0.43 e uma acurácia de 0.60 com aprendizado ativo utilizando Random Forest sem aumentação de dados e 0.42 e 0.62 de Cohen's Kappa e acurácia respectivamente ao utilizar modelos de linguagem para classificação e aumentação dos dados. Esta pesquisa contribui para o avanço das metodologias de análise automatizada em ambientes de aprendizagem online, abrindo possibilidades para a utilização das técnicas desenvolvidas no monitoramento e apoio ao desenvolvimento cognitivo dos alunos, promovendo uma melhor experiência de aprendizagem no ensino à distância.

Palavras-chaves: Comunidade de Investigação. Fóruns de Discussão. Analíticas de Aprendi-

zagem. Presença Cognitiva. Aprendizado Ativo.

ABSTRACT

With the establishment of distance learning as an educational model, various tools have been developed to provide an experience similar to that of in-person education. Among these tools, discussion forums stand out, offering students an environment for knowledge construction, social interaction, and information sharing. Machine learning techniques have also been employed to provide classifications of students' social and cognitive development levels based on their interactions in educational forums. The creation of these classifiers depends on various aspects (features) to increase the accuracy of the trained models; however, these models are highly dependent on the quantity and quality of the annotated data in the training set. Annotating this data is a labor-intensive task that relies on several domain experts, and there is also a scarcity of data due to the difficulty of large-scale acquisition of educational data. This work investigates the application of machine learning techniques for the automated analysis of online discussions in virtual learning environments, using the Community of Inquiry (CoI) model. It explores the feasibility of automatic methods for identifying cognitive presence in discussion forums, aiming to understand and optimize knowledge construction in distance education contexts. Different machine learning models were used, including Random Forest, XGBoost, and MLP, along with data augmentation approaches using BERT and GPT-4 to address the imbalance in cognitive presence categories. Textual features were extracted using tools like LIWC, Coh-Metrix, and social network analysis (SNA), providing a comprehensive representation of the interactions and content discussed, along with language models focused on encoding, such as DeBERTa. The results show that the combination of active learning techniques has great potential for the addressed problem, considering the limitation of data in the educational field, especially regarding annotated data. We achieved a Cohen's Kappa coefficient of 0.43 and an accuracy of 0.60 with active learning using Random Forest without data augmentation, and 0.42 and 0.62 of Cohen's Kappa and accuracy, respectively, when using language models for classification and data augmentation. This research contributes to the advancement of automated analysis methodologies in online learning environments, opening possibilities for the use of the developed techniques in monitoring and supporting students' cognitive development, promoting a better learning experience in distance education.

Keywords: Community of Inquiry. Discussion Foruns. Learning Analytics. Cognitive Presence. Active Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A ascensão do ensino à distância tem fomentado o desenvolvimento de diversas plataformas digitais de aprendizado, buscando replicar a interação e o engajamento típicos do ambiente presencial (KILAG et al., 2023; MURPHY, 2004; GARRISON; ANDERSON; ARCHER, 1999). Essas plataformas visam não apenas a transmissão de conteúdo, mas também a facilitação de interações ricas entre alunos e professores, essenciais para a construção do conhecimento e o desenvolvimento crítico dos estudantes (ANDERSON; SOSNIAK, 1994; BIGGS; COLLIS, 2014).

No entanto, essa transição para o ambiente virtual traz desafios únicos, como a necessidade de monitorar e analisar efetivamente as interações entre os participantes. Os fóruns de discussão, em particular, desempenham um papel crucial nesse contexto, proporcionando um espaço onde os alunos podem compartilhar ideias, esclarecer dúvidas e colaborar na resolução de problemas (KARUMBAIAH et al., 2021). Além disso, o acompanhamento do desenvolvimento cognitivo dos alunos nesse tipo de ambiente é de extrema importância, pois permite aos educadores identificar dificuldades, adaptar estratégias de ensino e fornecer feedback personalizado (BOZKURT; SHARMA, 2023; PRATAMA; SAMPELOLO; LURA, 2023; WHITESIDE; DIKKERS; SWAN, 2023).

Monitorar a evolução cognitiva dos estudantes ajuda a garantir que eles estejam não apenas absorvendo o conteúdo, mas também desenvolvendo habilidades críticas de pensamento, resolução de problemas e capacidade de aplicar o conhecimento de maneira prática. Sem um acompanhamento adequado, há o risco de que as necessidades individuais dos alunos passem despercebidas, comprometendo a eficácia do processo de ensino e aprendizagem. Portanto, desenvolver métodos eficazes para analisar as interações nos fóruns de discussão é essencial para promover uma educação de qualidade e maximizar o potencial de aprendizado em ambientes virtuais (FARROW; MOORE; GAŠEVIĆ, 2019).

Para enfrentar esses desafios, o modelo de Comunidade de Investigação (CoI) emerge como um framework teórico robusto (GARRISON; ANDERSON; ARCHER, 1999), projetado para auxiliar os educadores na compreensão e melhoria das interações em ambientes de ensino online. O CoI destaca três tipos de presenças essenciais: a presença social, que se refere à capacidade dos alunos de se projetarem como indivíduos reais em um ambiente online; a presença de ensino, que envolve o papel dos educadores na facilitação e direção das interações educacionais; e, mais importante para este estudo, a presença cognitiva, que foca na capacidade dos alunos

de construir e confirmar significado por meio de reflexão e diálogo (GARRISON; ANDERSON; ARCHER, 1999).

A presença cognitiva é fundamental para avaliar o desenvolvimento crítico e a construção de conhecimento dos estudantes. Utilizando o CoI, os educadores podem monitorar e analisar essas dimensões e acompanhar o progresso dos alunos, identificando áreas que necessitam de intervenção (GARRISON; ARBAUGH, 2007). Desta forma, o CoI não apenas fornece um mapa teórico para entender as dinâmicas de aprendizagem, mas também serve como uma ferramenta prática para melhorar a eficácia do ensino à distância, especialmente em termos de desenvolvimento cognitivo.

Com a vasta quantidade de dados textuais gerados nesses fóruns, surge a oportunidade de realizar análises detalhadas e automatizadas para compreender melhor o processo de aprendizagem. Entretanto, a análise manual dessas interações torna-se impraticável devido ao volume crescente de dados, exigindo o desenvolvimento de técnicas automatizadas, como a mineração de texto e o uso de aprendizagem de máquina, para apoiar os educadores na tarefa de promover um aprendizado eficaz e engajador (KOVANOVIĆ et al., 2016). Este cenário destaca a importância de desenvolver métodos eficientes para a classificação e análise das categorias da presença cognitiva, visando otimizar a experiência educacional em ambientes de ensino à distância.

Diante do desafio de disponibilidade e qualidade dos dados anotados para treinamento dos modelos de aprendizagem automática, esta pesquisa explora estratégias para mitigar essas limitações. Propõe-se o uso de aumentação de dados e técnicas de aprendizagem de máquina com poucos dados, visando superar a escassez de dados anotados e reduzir a dependência de um grande volume de dados anotados, o que frequentemente impede a aplicação prática dos modelos. A exploração dessas técnicas busca melhorar a generalização e aplicabilidade dos modelos desenvolvidos, permitindo a análise automática e eficaz das presenças cognitivas em ambientes de ensino à distância (BARBOSA et al., 2020; ROLIM et al., 2021).

1.1 MOTIVAÇÃO

A transformação digital no ensino superior, acelerada por eventos globais recentes, ressaltou a importância do ensino à distância como um elemento crucial para a continuidade da educação em situações adversas. Neste contexto, os fóruns de discussão online, integrados aos Ambientes Virtuais de Aprendizagem (AVA), emergiram como espaços essenciais para a

interação e construção de conhecimento entre estudantes e professores (MILMAN, 2015; LIM et al., 2017). A eficácia desses fóruns depende significativamente da capacidade de monitorar e analisar as interações neles contidas, visando otimizar o suporte educativo e promover um ambiente de aprendizado colaborativo e engajador (XIA; FIELDER; SIRAGUSA, 2013).

O modelo CoI, aplicado em ambientes online, é amplamente sustentado pela literatura acadêmica, destacando as presenças social, cognitiva e de ensino como componentes fundamentais para uma experiência educacional eficiente no ensino superior (GARRISON; ANDERSON; ARCHER, 1999). A análise das interações nesse modelo oferece visão ampla sobre o processo de aprendizagem, engajamento dos alunos e a qualidade da educação oferecida. Contudo, as metodologias convencionais de análise, principalmente a Análise de Conteúdo Quantitativa (QCA), enfrentam desafios de escalabilidade e aplicação em tempo real devido à natureza laboriosa e prolongada da codificação manual dos dados (DONNELLY; GARDNER, 2011; KOVANOVIC et al., 2014b).

A necessidade de métodos de análise mais ágeis e adaptativos justifica a investigação de abordagens automatizadas que possam processar grandes volumes de dados de forma eficiente. A automação não apenas acelera o processo de análise, mas também proporciona uma resposta educativa mais rápida e informada, potencialmente transformando a dinâmica do aprendizado em cursos online (GAšEVIć; KOVANOVIć; JOKSIMOVIć, 2017). Além disso, a implementação de técnicas de aprendizagem de máquina e análise de redes em fóruns de discussão pode descobrir padrões complexos de interação e conhecimento que são difíceis de identificar manualmente (FERREIRA-MELLO et al., 2019).

No entanto, a eficácia dessas técnicas depende crucialmente da disponibilidade de dados anotados de alta qualidade, que são escassos e onerosos para produzir, especialmente em um domínio tão especializado como o do ensino superior (FERREIRA et al., 2020). Portanto, este estudo também se dedica a explorar abordagens inovadoras para a geração de dados, como o uso de dados sintéticos, para superar algumas essas limitações (BARBOSA et al., 2020; ROLIM et al., 2019).

Dessa forma, a motivação desta pesquisa é dupla: melhorar a compreensão e a gestão das interações em fóruns online no ensino superior através da aplicação de métodos automatizados de análise de texto e, simultaneamente, enfrentar os desafios associados à preparação de dados para aprendizagem de máquina em ambientes educacionais. Este estudo busca contribuir para a literatura acadêmica e a prática educativa, fornecendo soluções viáveis e escaláveis para a análise de discussões online, essenciais para o desenvolvimento de práticas educativas mais

eficazes e inclusivas no ensino superior à distância.

1.2 PROPOSTA

Esta pesquisa tem como objetivo desenvolver uma metodologia para a análise automatizada de discussões online, com foco na classificação das categorias da presença cognitiva, utilizando o modelo Comunidade de Investigação (CoI).

1.2.1 Estrutura da Metodologia

Inicia-se com a extração de características textuais a partir das postagens dos estudantes nos fóruns de discussão. As características incluem, mas não estão limitadas a, contagens de palavras, análises linguísticas utilizando ferramentas como LIWC e Coh-Metrix, e elementos derivados das análises de redes sociais (SNA). Essas características são fundamentais para capturar a riqueza dos dados textuais e fornecer um contexto detalhado para a análise subsequente (TAUSCZIK; PENNEBAKER, 2010; MCNAMARA et al., 2014).

1.2.2 Modelagem e Classificação

Utilizaremos técnicas de aprendizagem de máquina para desenvolver modelos capazes de classificar automaticamente as diferentes categorias da presença cognitiva, conforme definido pelo modelo CoI. Algoritmos como Random Forest, XGBoost e MLP serão implementados e comparados para determinar o mais eficaz na identificação das categorias cognitivas em fóruns de discussão.

Para superar a escassez de dados anotados e o desbalanceamento entre as categorias, aplicaremos métodos de aumentação de dados. Utilizaremos modelos de linguagem como BERT e GPT-4 para gerar dados sintéticos que enriquecem o conjunto de treinamento, melhorando a generalização e a precisão dos modelos de classificação.

A extração de características textuais será realizada utilizando ferramentas avançadas como LIWC e Coh-Metrix, que fornecem uma análise detalhada dos aspectos linguísticos e psicológicos das interações nos fóruns. Além disso, modelos baseados na arquitetura Transformer, como o DeBERTa, serão utilizados para gerar *embeddings* textuais ricos em contexto ("*embeddings*" são representações numéricas que ajudam o computador a entender o significado e

a relação entre palavras), que serão fundamentais para a etapa de classificação.

Implementaremos técnicas de aprendizagem ativa para otimizar o treinamento dos modelos com um volume limitado de dados anotados. Esta abordagem permitirá que os modelos selecionem de forma inteligente os exemplos mais informativos para anotação adicional, melhorando a eficiência do processo de treinamento e a precisão das classificações.

Realizaremos uma série de experimentos comparando o desempenho dos modelos treinados com e sem a utilização de dados aumentados. Essa comparação permitirá avaliar a eficácia das técnicas de aumentação de dados na melhoria da acurácia e robustez dos modelos de classificação da presença cognitiva. A Figura 1 ilustra o fluxo de trabalho proposto para a condução dos experimentos.

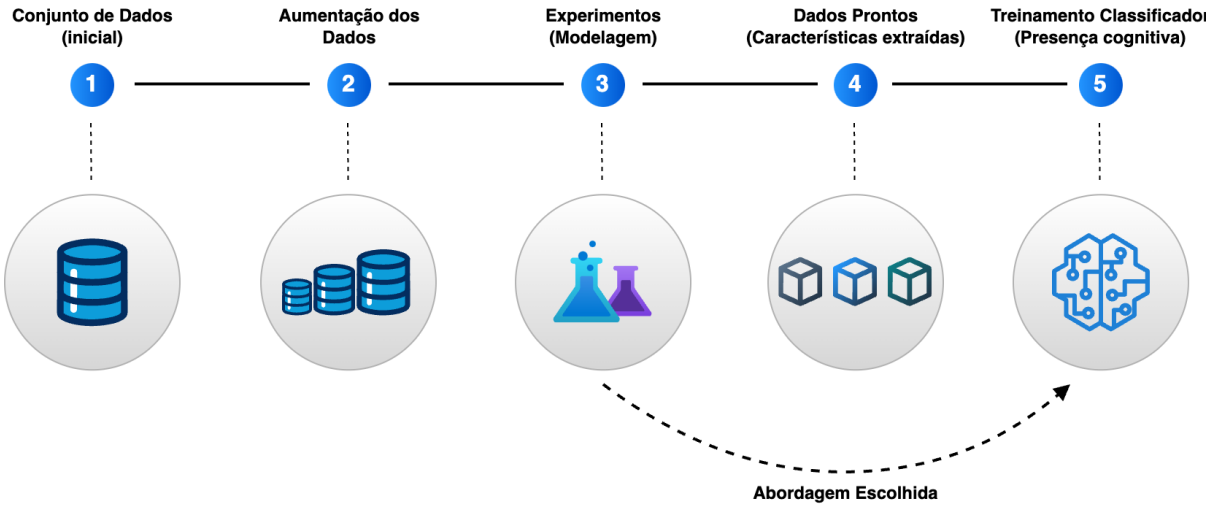

Figura 1 – **Representação visual da proposta.**

1.2.3 Implicações e Aplicações Práticas

A aplicação prática da metodologia desenvolvida nesta pesquisa oferece uma visão aprofundada sobre as dinâmicas cognitivas que ocorrem nos fóruns de discussão online no contexto do ensino superior. Com os resultados obtidos, espera-se que professores e pesquisadores sejam munidos de ferramentas eficazes para realizar intervenções informadas e ajustar o planejamento de cursos futuros, melhorando assim a qualidade e eficácia da educação online.

Ao proporcionar uma análise automatizada e detalhada das interações em fóruns, a proposta apresentada neste trabalho permitirá uma compreensão mais precisa de como os estudantes se envolvem com o conteúdo. Isso facilita a identificação de padrões de engajamento e

possíveis áreas de melhoria no processo educativo. A capacidade de realizar intervenções rápidas e baseadas em evidências é especialmente valiosa em ambientes de aprendizagem dinâmicos e diversificados, comuns em universidades que atendem a uma ampla gama de disciplinas e populações estudantis.

A flexibilidade da metodologia proposta permite sua adaptação e aplicação a diferentes contextos educacionais, desde cursos de graduação até programas de pós-graduação e educação continuada. Essa adaptabilidade amplia seu impacto e utilidade na comunidade de educação superior à distância. Tal adaptação e aplicação contribuem para o desenvolvimento de práticas educacionais mais eficazes e inclusivas, essenciais para o sucesso do ensino superior na era digital. Além disso, a análise automatizada das categorias da presença cognitiva pode ser integrada a sistemas de gestão de aprendizado, proporcionando feedback contínuo e em tempo real para educadores e estudantes, promovendo um ambiente de aprendizado mais responsivo e personalizado.

Embora este trabalho tenha sido desenvolvido no contexto do ensino superior, há a possibilidade da aplicação dessa proposta no ensino fundamental e médio, etapas cruciais do desenvolvimento cognitivo. Essa aplicação potencial permitiria identificar rapidamente desvios de aprendizagem, possibilitando intervenções imediatas e personalizadas (KONOPKA; ADAIME; MOSELE, 2015). Assim, os alunos poderiam receber o suporte necessário para otimizar seu processo educacional desde as fases iniciais de sua formação.

1.3 HIPÓTESE E PROBLEMA DE PESQUISA

A crescente adoção de plataformas de ensino à distância tem apresentado novos desafios e oportunidades para a análise e compreensão das interações educacionais. Dentre essas interações, as discussões em fóruns virtuais são fundamentais para a construção do conhecimento e desenvolvimento cognitivo dos alunos. No entanto, a análise manual desses dados textuais é impraticável devido ao grande volume de informações geradas. Para enfrentar esses desafios, este estudo propõe a aplicação de técnicas de aprendizagem de máquina para a análise automática das categorias da presença cognitiva no modelo de Comunidade de Investigação (CoI). Diante disso, formulamos as seguintes perguntas de pesquisa para guiar nossa investigação:

> **RQ1:** Como as técnicas de aprendizagem de máquina, combinadas com diferentes abordagens de mineração de texto e representação textual, podem ser utilizadas

para classificar as categorias da presença cognitiva e qual é o impacto dessas abordagens na acurácia dos modelos?

Esta pergunta (RQ1) busca explorar a aplicação de algoritmos de aprendizado de máquina na identificação das diferentes categorias da presença cognitiva, enquanto avalia o impacto de diversas técnicas de extração de características e representações textuais, incluindo modelos baseados em *Transformer* como o DeBERTa, no desempenho dos modelos em termos de acurácia e coeficiente de Cohen's Kappa.

RQ2: Quais são as melhores práticas para lidar com a escassez de dados anotados e o desbalanceamento das categorias da presença cognitiva?

Considerando a dependência de grandes volumes de dados anotados para o treinamento de modelos precisos, esta pergunta (RQ2) investiga métodos como a aumentação de dados e o aprendizado ativo para melhorar a generalização e aplicabilidade dos modelos em cenários com dados limitados.

Ao abordar estas perguntas de pesquisa, esperamos avançar o conhecimento sobre o uso de técnicas de aprendizagem de máquina para a análise automática das interações educacionais em ambientes de ensino à distância, contribuindo para a melhoria da qualidade do ensino e aprendizagem.

1.4 OBJETIVOS

Esta seção delineia os objetivos gerais e específicos da pesquisa proposta, visando estabelecer um direcionamento claro para as atividades de desenvolvimento e análise dentro do âmbito deste trabalho.

1.4.1 Objetivo Geral

Desenvolver uma metodologia para a análise automatizada da presença cognitiva em fóruns de discussão online, utilizando o modelo Comunidade de Investigação (CoI), com o intuito de melhorar o entendimento das interações e o desenvolvimento cognitivo dos estudantes em ambientes de ensino superior à distância.

1.4.2 Objetivos Específicos

Os objetivos específicos estão projetados para garantir uma cobertura abrangente dos aspectos técnicos e educacionais da pesquisa, permitindo não apenas a criação de uma solução eficaz para a análise de discussões online, mas também o aprofundamento da compreensão científica e prática sobre as dinâmicas de aprendizado em ambientes virtuais de ensino superior. Os objetivos específicos são:

1. Implementar técnicas de aprendizagem de máquina para classificar automaticamente as categorias da presença cognitiva em fóruns de discussão.
2. Explorar e aplicar métodos de aumentação de dados, para lidar com a escassez e o desbalanceamento dos dados anotados.
3. Avaliar a eficácia de diferentes representações textuais, para a extração de características das interações nos fóruns.
4. Realizar experimentos utilizando técnicas de aprendizagem ativa para melhorar a precisão dos modelos de classificação, mesmo com um volume limitado de dados anotados.
5. Comparar o desempenho dos modelos treinados com dados aumentados e não aumentados, analisando a melhoria na acurácia e na robustez das classificações de presença cognitiva.
6. Validar a abordagem proposta em um conjunto de dados real de discussões online em ambientes de ensino superior.

1.5 ORGANIZAÇÃO DO TRABALHO

Os Capítulos a seguir estão estruturados da seguinte forma:

Capítulo 2 - Fundamentação Teórica: Este capítulo apresenta as técnicas e conceitos fundamentais para a compreensão total do trabalho. São discutidos os principais *frameworks* teóricos, como o modelo de Comunidade de Investigação, bem como as técnicas de mineração de texto e aprendizagem de máquina utilizadas na análise das interações nos fóruns de discussão.

Capítulo 3 - Trabalhos Relacionados: Este capítulo apresenta uma revisão detalhada dos trabalhos mais relevantes na área, discutindo as abordagens, metodologias e resultados obtidos em pesquisas que, assim como este estudo, buscam automatizar a análise da presença cognitiva em contextos educacionais.

Capítulo 4 - Metodologia: Neste capítulo, detalhamos a metodologia adotada para conduzir os experimentos que buscam responder às perguntas de pesquisa. São explicadas as etapas de coleta e preparação de dados, extração de características, treinamento e avaliação dos modelos de aprendizagem de máquina, bem como as estratégias de aumentação de dados e tratamento do desbalanceamento das categorias da presença cognitiva.

Capítulo 5 - Resultados: Este capítulo apresenta os resultados obtidos nos experimentos realizados. São discutidas as métricas de avaliação dos modelos, como acurácia e coeficiente de Cohen's Kappa, e analisado o desempenho dos diferentes algoritmos e abordagens utilizadas, comparando-os com os objetivos estabelecidos pelas perguntas de pesquisa.

Capítulo 6 - Conclusão: O capítulo final discute as principais conclusões do trabalho, destacando as contribuições para a área de ensino à distância e aprendizagem de máquina. São abordadas também as limitações dos métodos propostos e sugestões para trabalhos futuros, visando aprimorar e expandir as técnicas desenvolvidas.

Cada um desses capítulos é projetado para fornecer uma compreensão abrangente do estudo, desde a fundamentação teórica até as implicações práticas e futuras direções de pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 APRENDIZAGEM DE MÁQUINA

A aprendizagem de máquina, uma subárea da inteligência artificial, desempenha um papel crucial no desenvolvimento de sistemas que aprendem a partir de dados e realizam previsões ou decisões sem a necessidade de serem explicitamente programados (MITCHELL, 1997). Esta área é de particular importância no tratamento de grandes volumes de dados, com aplicações extensas em diversos campos, incluindo a educação.

A capacidade de processar e analisar grandes quantidades de texto de forma automatizada transformou diversas indústrias, incluindo a educacional. A aprendizagem de máquina oferece ferramentas poderosas para classificação textual, o que permite desde a análise de sentimentos até a categorização automática de conteúdos em ambientes virtuais de aprendizagem (AVAs).

Nos AVAs, a classificação textual pode ser utilizada para organizar conteúdos, moderar discussões, personalizar o aprendizado, e até mesmo modelar o perfil do estudante. Por exemplo, postagens de alunos em fóruns podem ser automaticamente categorizadas para identificar dúvidas comuns, sentimentos dos estudantes em relação a tópicos específicos, ou até mesmo para detectar sinais de desengajamento ou confusão.

2.1.1 Tipos de Aprendizagem de Máquina

A aprendizagem de máquina pode ser categorizada em várias formas, dependendo da natureza do "rótulo" ou "*feedback*" disponível para um sistema de aprendizagem:

- **Aprendizagem Supervisionada**: Os modelos aprendem a partir de um conjunto de dados contendo exemplos de entradas e saídas desejadas, fornecidos por um "supervisor". Este método é amplamente utilizado em tarefas como classificação e regressão, onde o objetivo é aprender a mapear entradas para saídas corretas (BISHOP, 2006).
- **Aprendizagem Não Supervisionada**: Em contraste, a aprendizagem não supervisionada lida com dados sem rótulos explícitos, focando em identificar padrões subjacentes ou estruturas a partir dos dados de entrada. Exemplos comuns incluem agrupamento e redução de dimensionalidade (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

- **Aprendizagem Auto-supervisionada**: Uma técnica emergente que gera automaticamente os rótulos a partir dos dados, geralmente modificando os dados de entrada existentes para criar "tarefas" que um modelo tenta resolver. Exemplos incluem prever a próxima palavra em uma sentença ou completar uma imagem parcialmente oculta (DEVLIN et al., 2018).

- **Aprendizagem por Reforço**: Modelos que aprendem a tomar decisões sequenciais, observando recompensas ou punições, sem correção direta. É comumente aplicada em jogos, robótica e navegação, onde um agente aprende a escolher ações que maximizem alguma noção de recompensa cumulativa (SUTTON; BARTO, 2018).

Embora a aprendizagem supervisionada, que depende de dados anotados, geralmente produza modelos que atingem bons resultados (dependendo do conjunto de dados), é comum enfrentar dificuldades devido à escassez de dados anotados em certos problemas. Para mitigar essa limitação, diversos estudos e abordagens foram desenvolvidos ao longo dos anos. Dentre essas abordagens, o *Aprendizado Ativo* (*Active Learning*) se destaca como uma estratégia eficaz para otimizar o uso de dados anotados e melhorar a eficiência dos modelos de aprendizado de máquina.

O Aprendizado Ativo é uma ramificação da aprendizagem de máquina que lida com a escassez de dados anotados. Esta abordagem busca construir um modelo de aprendizado de máquina utilizando um conjunto mínimo necessário de dados anotados (SETTLES, 2009). A ideia central é melhorar a acurácia do modelo por meio de métodos de seleção de instâncias, baseados no grau de informatividade, de um conjunto de dados não anotado que será posteriormente anotado por um especialista, conhecido como oráculo. No caso de um conjunto de dados já anotado, esse oráculo poderia ser o texto previamente anotado.

Existem três abordagens principais para o aprendizado ativo (SHARMA; BILGIC, 2017):

(i) **Síntese de Consulta de Membros**: onde o sistema gera novas instâncias para consulta ao oráculo.

(ii) **Amostragem Seletiva Baseada em Fluxo**: que seleciona instâncias sequencialmente e decide se devem ser anotadas com base em um critério de informatividade.

(iii) **Aprendizado Ativo Baseado em Pool**: É uma abordagem comum em aprendizado de máquina, especialmente útil quando há poucos dados anotados disponíveis. Nesse

método, o modelo escolhe dados de um "pool" de dados não anotados, selecionando, de forma estratégica, as instâncias mais informativas, ou seja, aquelas que ele ainda tem mais incerteza ou dúvidas. Essas instâncias ajudam o modelo a aprender de forma mais eficaz, já que, ao serem anotadas, oferecem as informações de que ele mais precisa para melhorar seu desempenho (SETTLES, 2009).

Várias estratégias podem ser utilizadas para avaliar a informatividade de cada instância no conjunto de dados. Este estudo adotou a estratégia de Amostragem por Incerteza (*Uncertainty Sampling* - US), que é a mais simples e comum na literatura (SHARMA; BILGIC, 2017). Em resumo, a US utiliza a entropia da instância para medir a incerteza sobre como um codificador humano poderia codificar corretamente essa instância (SETTLES, 2009).

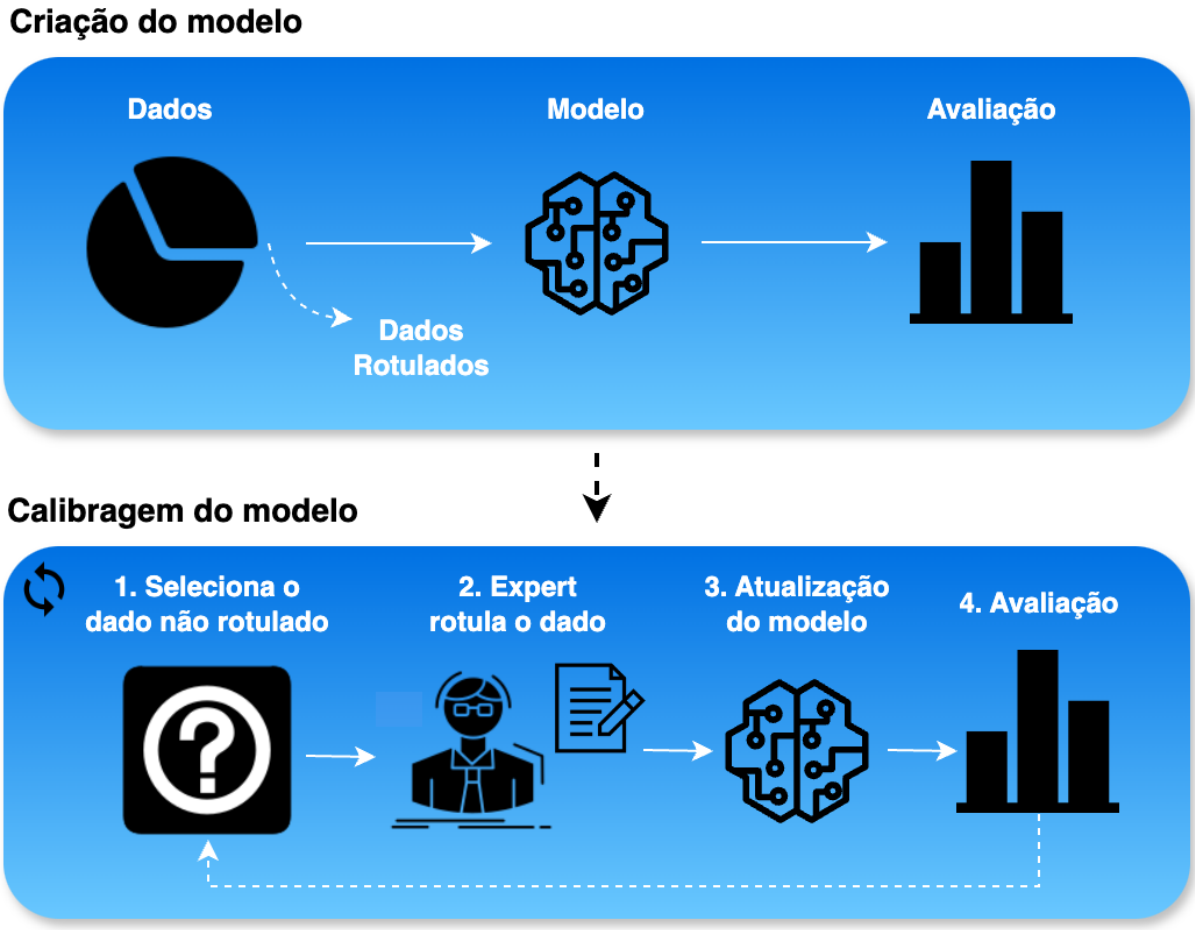

Figura 2 – **Exemplo de uma rede neural.**

A Figura 2 apresenta o processo de treinamento de um classificador incluindo a abordagem de aprendizado ativo. Inicialmente, o modelo de ML é construído usando apenas uma pequena amostra de dados anotados (geralmente entre 10% e 20%), em seguida, cada instância é

consultada uma a uma do conjunto de dados não anotados de acordo com a medida de incerteza para ser anotada por um especialista (oráculo); a nova instância anotada é inserida no modelo de ML para atualizá-lo. Esta etapa de calibração do modelo pode ser executada até que o classificador alcance resultados satisfatórios.

2.1.2 Modelos e Técnicas de Classificação Textual

Nesta seção, discutiremos os modelos e técnicas mais relevantes utilizados para classificar textos, uma tarefa fundamental no processamento de linguagem natural que permite organizar e analisar grandes volumes de dados textuais. Abordaremos desde métodos tradicionais até abordagens modernas baseadas em aprendizagem profunda, destacando suas aplicações, vantagens e desafios no contexto da classificação de textos.

Diversos modelos e técnicas de aprendizagem de máquina são utilizados para a classificação textual, incluindo:

- **Modelos baseados em regras**: Utilizam regras linguísticas pré-definidas para classificar textos.
- **Árvores de Decisão e Máquinas de Vetores de Suporte (SVM)**: Modelos populares em tarefas de classificação que aprendem fronteiras de decisão a partir de dados etiquetados (QUINLAN, 1986; CORTES; VAPNIK, 1995).
- **Modelos de Ensemble**: Combinam as previsões de vários modelos básicos para melhorar a generalização sobre dados novos, como *Random Forests* e *boosting* (FREUND; SCHAPIRE, 1997).
- **Redes Neurais e Redes Neurais Profundas**: Aprendem representações de texto de alta dimensionalidade, ideais para contextos complexos de texto (LECUN; BENGIO; HINTON, 2015).
- **Modelos de linguagem pré-treinados**: Oferecem capacidades avançadas de entendimento de contexto e semântica, úteis para classificação textual sofisticada (DEVLIN et al., 2018).

A aplicação de aprendizagem de máquina na classificação textual em ambientes educacionais online não apenas melhora a eficiência dos processos educativos, mas também enriquece

a experiência de aprendizado dos alunos, proporcionando um ambiente mais interativo e personalizado.

2.1.3 Classificação Textual

A classificação textual é uma área central do processamento de linguagem natural (PLN) que envolve a atribuição automática de etiquetas ou categorias a textos escritos (MANNING; MANNING; SCHÜTZE, 1999). Esta tarefa é fundamental em diversas aplicações, como a filtragem de emails, análise de sentimentos, detecção de spam e categorização automática de documentos (SEBASTIANI, 2002).

2.1.3.1 Processo de Classificação Textual

A classificação textual é uma tarefa complexa que envolve várias etapas, desde o pré-processamento até a avaliação dos modelos de aprendizagem de máquina.

2.1.3.1.1 Pré-processamento Textual

O pré-processamento é uma etapa crítica na classificação textual que envolve a preparação dos textos para uma análise mais eficaz. Técnicas comuns incluem a remoção de palavras de parada, a normalização de texto (como a conversão de todas as letras para minúsculas), a remoção de pontuação e números, e a correção ortográfica. Outras técnicas avançadas podem incluir a lematização e a remoção de palavras raras ou de alta frequência que podem ser irrelevantes para a análise (MANNING; MANNING; SCHÜTZE, 1999).

2.1.3.1.2 Representação Textual

A representação textual é fundamental para transformar texto bruto em um formato que os algoritmos de aprendizagem de máquina possam interpretar:

- **Bag-of-Words (BoW)**: Este método representa textos como vetores onde cada dimensão corresponde à frequência de uma palavra no documento, transformando texto livre em um formato numérico que pode ser facilmente manipulado por algoritmos. O BoW é simples, mas eficaz, apesar de sua principal limitação ser a perda de informação sobre a ordem das palavras e o contexto, o que pode ser crucial para muitas tarefas de PLN.

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: Esta técnica é uma extensão do BoW que ajusta a frequência das palavras levando em conta não apenas sua presença em um único documento, mas em toda a coleção de documentos. Palavras que aparecem frequentemente em um documento, mas raramente em outros, recebem uma maior ponderação, ajudando a destacar termos significativos que são potencialmente mais interessantes que palavras comuns (SALTON; BUCKLEY, 1988).

- **Word Embeddings**: Ao contrário do BoW e TF-IDF, que geram representações esparsas, os *embeddings* de palavras produzem vetores densos, onde cada dimensão representa uma característica latente capturada do corpus de treinamento. Modelos como Word2Vec (MIKOLOV et al., 2013) e GloVe (PENNINGTON; SOCHER; MANNING, 2014) aprendem esses *embeddings* de modo que palavras com significados similares ou relacionados estejam próximas umas das outras no espaço vetorial, capturando nuances semânticas e sintáticas.

- **Encoding com Modelos de Linguagem**: Técnicas mais recentes e avançadas utilizam modelos de linguagem pré-treinados, como BERT (DEVLIN et al., 2018) e GPT (RADFORD et al., 2018), que são capazes de entender o contexto em que as palavras são usadas. Esses modelos geram **embeddings** contextuais, significando que a representação de uma palavra pode mudar com base no seu contexto textual, proporcionando uma compreensão mais profunda do texto.

- **Extração de Características Linguísticas**: Além das representações baseadas puramente em texto, ferramentas como LIWC e Coh-Metrix permitem a análise de características linguísticas, psicológicas e discursivas de textos. Essas ferramentas analisam aspectos como complexidade gramatical, estilo, tom emocional e estruturas cognitivas, que podem ser cruciais para certas aplicações de PLN (TAUSCZIK; PENNEBAKER, 2010; GRAESSER; MCNAMARA; KULIKOWICH, 2011).

2.1.3.1.3 Treinamento, Teste e Avaliação do Modelo

O treinamento de modelos de classificação envolve ajustar os parâmetros dos algoritmos aos dados de treinamento. Após o treinamento, o modelo é testado com um conjunto de dados separado para avaliar sua performance. As métricas comuns de avaliação incluem:

- **Acurácia**: Representa a proporção de previsões corretas realizadas pelo modelo em relação ao total de casos analisados. É uma métrica simples e direta, que oferece uma visão geral da eficiência do modelo em termos de suas previsões corretas. No entanto, deve ser utilizada com cautela em conjuntos de dados desbalanceados, onde uma classe pode dominar sobre as outras, levando a uma interpretação enganosa da eficácia real do modelo.

- **F1-Score**: Funciona como a média harmônica entre precisão e revocação, equilibrando essas duas métricas importantes. A precisão é a proporção de identificações positivas corretas (verdadeiros positivos) em relação a todas as identificações positivas (verdadeiros positivos mais falsos positivos), enquanto a revocação é a proporção de identificações positivas corretas em relação ao número de casos positivos reais (verdadeiros positivos mais falsos negativos). O F1-Score é particularmente útil em situações onde há classes desbalanceadas ou quando a penalidade por diferentes tipos de erros varia, sendo um indicador robusto da qualidade do modelo.

- **Cohen's Kappa**: Esta métrica mede a concordância entre as previsões do modelo e as classificações verdadeiras, ajustada pela probabilidade de concordância ao acaso (COHEN, 1960). O coeficiente de Kappa é especialmente útil quando se deseja avaliar a precisão de um modelo em cenários onde a concordância aleatória pode ser significativa, ajudando a distinguir entre a precisão do modelo e simples coincidências. Ele fornece um escore que vai de -1 a 1, onde valores mais altos indicam uma concordância melhor que o acaso, zero indica concordância ao nível do acaso, e valores negativos indicam discordância.

Cada uma dessas métricas oferece uma visão única sobre diferentes aspectos da performance do modelo. Enquanto a acurácia fornece uma medida geral de desempenho, o F1-Score ajuda a entender melhor o equilíbrio entre a precisão e a revocação, crucial em muitas aplicações práticas onde o custo de falsos negativos é significativo. Por fim, Cohen's Kappa é uma ferramenta essencial para avaliar a confiabilidade do modelo em condições onde as taxas de acerto por chance podem ser altas, garantindo que as previsões do modelo são genuinamente informativas além do que seria esperado por simples aleatoriedade.

A avaliação rigorosa dos modelos é crucial para garantir que eles sejam generalizáveis e eficazes em contextos reais de aplicação.

2.2 ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Esta seção é dedicada à exploração de algoritmos de aprendizagem de máquina que formam a espinha dorsal de muitas aplicações práticas em diversos campos do conhecimento. Após uma introdução geral aos conceitos fundamentais, focaremos agora em três categorias principais de algoritmos que são críticos para o avanço da inteligência artificial e suas aplicações práticas: algoritmos baseados em árvores, redes neurais e, por fim, modelos de linguagem em larga escala (LLM).

2.2.1 Algoritmos Baseados em Árvore

Algoritmos baseados em árvores de decisão são um dos pilares da aprendizagem de máquina, valorizados por sua interpretabilidade, facilidade de implementação e eficácia em uma ampla gama de problemas. A árvore de decisão é um modelo preditivo que segmenta o espaço de dados em regiões menores, onde decisões mais simples são tomadas, baseadas em perguntas sequenciais sobre as características dos dados (BREIMAN et al., 1984).

Uma árvore de decisão é construída a partir de um processo recursivo de divisão do conjunto de dados, começando no nó raiz, que contém o conjunto de dados completo, até os nós folha, que representam as decisões ou previsões finais. Cada divisão é feita de forma a maximizar a homogeneidade dos subconjuntos resultantes em relação à variável alvo. O exemplo da estrutura de uma árvore de decisão pode ser observado na Figura 3

2.2.1.1 Métodos de Ensemble: Bagging e Boosting

Embora eficazes, as árvores de decisão são frequentemente criticadas por sua alta variância e tendência ao *overfitting* (KHOSHGOFTAAR; ALLEN, 2001). Métodos de *ensemble*, como *bagging* e *boosting*, foram desenvolvidos para superar essas limitações, combinando múltiplas árvores para melhorar a robustez e o desempenho do modelo.

Bagging, ou *Bootstrap Aggregating*, envolve treinar múltiplas árvores de decisão de forma independente em diferentes subconjuntos dos dados, criados com reposição, e depois agregá-las para formar a previsão final. Este método reduz a variância sem aumentar o viés, o que é particularmente útil em modelos muito sensíveis à variação dos dados de treinamento, como árvores de decisão (BREIMAN, 1996).

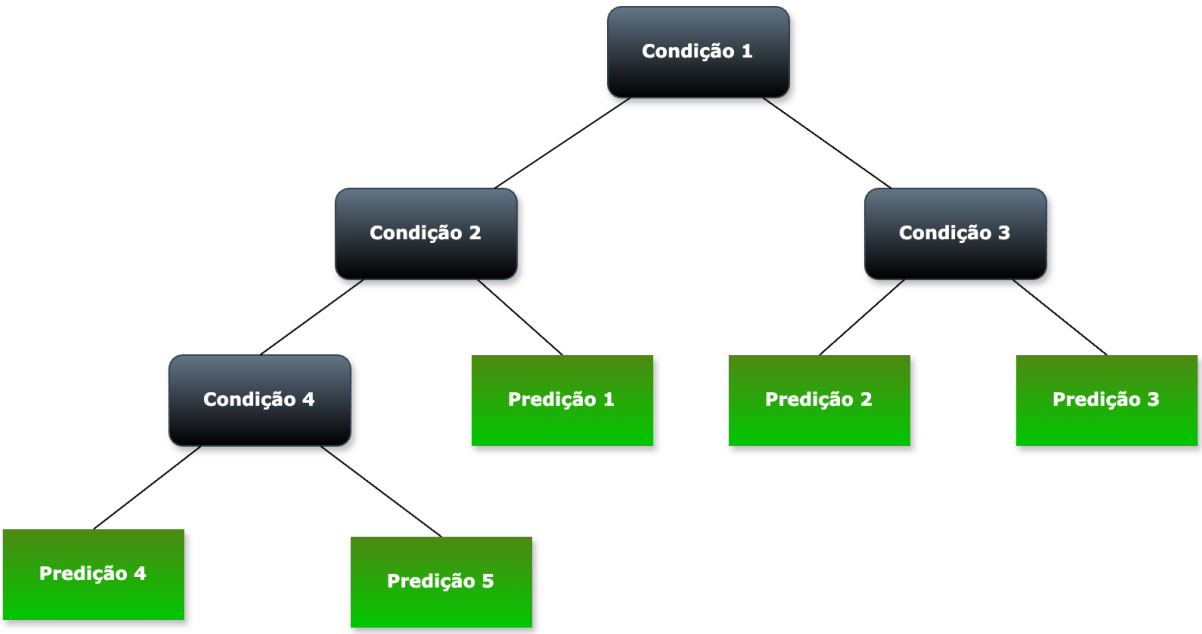

Figura 3 – **Exemplo de Árvore de Decisão.**

Boosting, por outro lado, constrói modelos de forma sequencial, onde cada novo modelo aprende a corrigir os erros cometidos pelos modelos anteriores. Ao contrário do *bagging*, o *boosting* pode ajustar tanto o viés quanto a variância, tornando-se eficaz mesmo quando os modelos individuais são fracos. *AdaBoost* e *Gradient Boosting* são exemplos populares de algoritmos de *boosting* (FREUND; SCHAPIRE, 1997; FRIEDMAN, 2001).

2.2.1.2 Random Forest

O ***Random Forest*** é um dos algoritmos mais poderosos e amplamente utilizados em aprendizagem de máquina para tarefas de classificação e regressão. Baseado no conceito de *bagging*, este algoritmo utiliza múltiplas árvores de decisão para criar um "floresta" que é mais robusta e precisa do que modelos de árvore de decisão individuais (BREIMAN, 2001). O exemplo do seu funcionamento pode ser observado na Figura 4.

O *Random Forest* opera construindo uma coleção de árvores de decisão durante o treinamento. Cada árvore é treinada usando uma amostra aleatória do conjunto de dados de treinamento, escolhida com reposição, conhecida como *bootstrap sample*. Além disso, em cada divisão de um nó durante a construção da árvore, um subconjunto aleatório de características é selecionado. Esse processo aumenta a diversidade entre as árvores na floresta, o que ajuda a reduzir a variância do modelo final e melhora a generalização em dados não vistos.

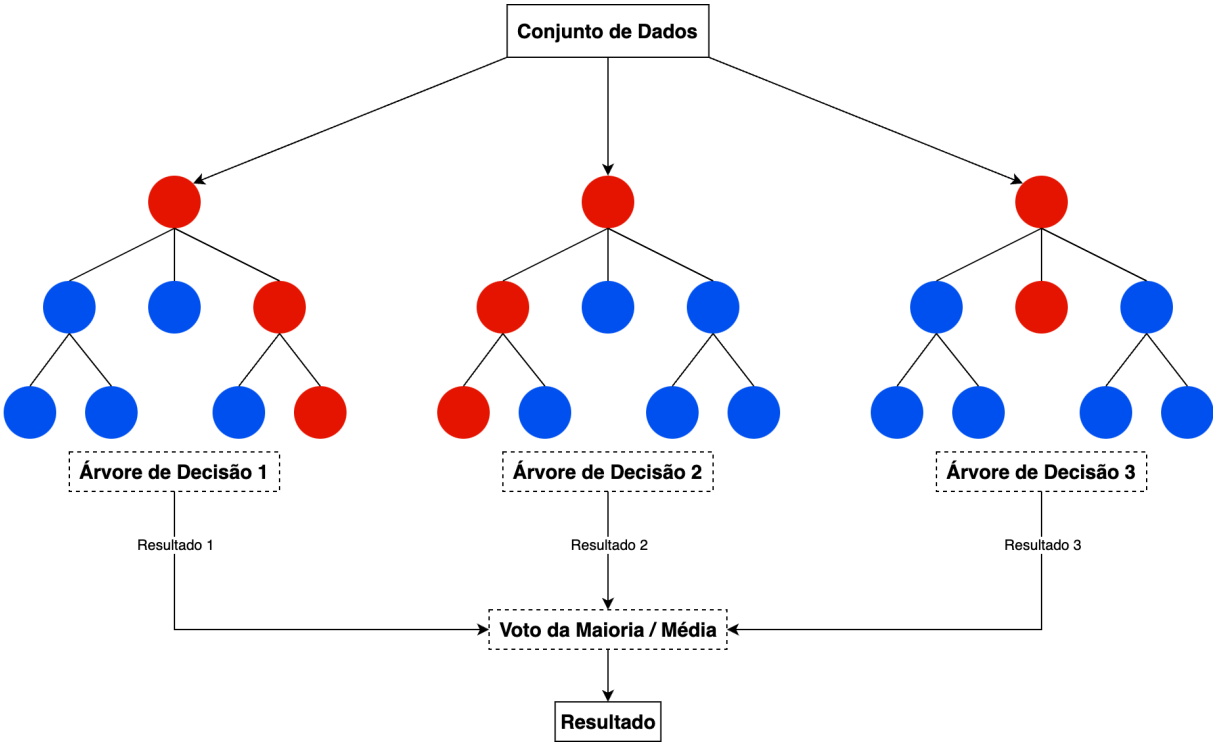

Figura 4 – **Exemplo do Random Forest.**

Uma das principais vantagens do *Random Forest* é sua capacidade de limitar o risco de *overfitting*, um problema comum em modelos de árvore de decisão que aprendem de forma demasiado específica para os dados de treinamento. Ao agregar as previsões de várias árvores, o *Random Forest* tende a cancelar os erros de uma única árvore, especialmente em casos onde esses erros não são sistemáticos (BREIMAN, 2001). Contudo, é importante mencionar que apesar de serem menos propensos a *overfitting*, podem ainda assim sofrer deste problema em situações com ruído em níveis elevados nos dados de treinamento.

Outro benefício do uso do *Random Forest* é sua capacidade de fornecer estimativas da importância de cada característica para a decisão final. Isso é alcançado observando quanto a acurácia das árvores diminui quando os dados para uma determinada característica são aleatorizados. Essa informação é crucial para entender os fatores que mais influenciam o modelo, o que pode ser especialmente útil em aplicações de ciência de dados onde a interpretabilidade é importante.

O *Random Forest* tem sido aplicado com sucesso em uma vasta gama de campos, incluindo, classificação das presenças cognitivas e sociais do modelo CoI (NETO et al., 2018; FERREIRA et al., 2020; KOVANOVIć et al., 2016). Sua flexibilidade e robustez o tornam adequado para problemas onde as relações entre características podem ser complexas e nãolineares. Em

conclusão, o *Random Forest* é uma ferramenta poderosa na caixa de ferramentas de aprendizagem de máquina, oferecendo uma combinação eficaz de precisão, robustez e facilidade de uso que é difícil de superar com modelos mais simples.

2.2.2 XGBoost: Extreme Gradient Boosting

O *Extreme Gradient Boosting* (XGBoost) é um algoritmo de aprendizado de máquina amplamente utilizado para tarefas de classificação e regressão, conhecido por sua eficiência computacional e desempenho superior em uma variedade de aplicações práticas. Desenvolvido por Chen *et al.* (CHEN; GUESTRIN, 2016), o XGBoost é uma implementação aprimorada do *gradient boosting* que otimiza a precisão dos modelos baseados em árvores de decisão.

O XGBoost pertence à classe dos métodos de *ensemble*, que combinam os resultados de múltiplos modelos fracos para formar um modelo forte. Mais especificamente, ele utiliza a técnica de *gradient boosting*, onde modelos são construídos sequencialmente e cada novo modelo é treinado para corrigir os erros do modelo anterior. A intuição por trás do *boosting* é que conjuntos de modelos simples (como árvores de decisão de baixa profundidade) podem ser combinados para formar um modelo forte.

No XGBoost, cada novo modelo tenta minimizar uma função de perda, adicionando novas árvores de decisão ao conjunto. A função de perda típica é baseada no gradiente da função objetivo em relação às predições atuais, daí o nome *gradient boosting* (CHEN; GUESTRIN, 2016).

O XGBoost melhora o *gradient boosting* padrão em vários aspectos. Primeiro, ele implementa uma técnica conhecida como *shrinkage*, onde o peso das árvores adicionadas ao modelo é reduzido para evitar atualizações excessivas. Segundo, ele introduz regularização L1 e L2 para penalizar a complexidade dos modelos, o que é crucial para evitar o *overfitting* (FRIEDMAN, 2001). Além disso, o XGBoost emprega uma técnica de busca aproximada para encontrar os melhores pontos de divisão nas árvores de decisão, o que reduz substancialmente o tempo de computação (CHEN; GUESTRIN, 2016).

O XGBoost tem sido amplamente adotado em competições de ciência de dados e na indústria devido à sua habilidade em produzir modelos de alta precisão com tempo de treinamento relativamente curto. Por exemplo, em desafios de classificação de dados tabulares, como os hospedados na plataforma *Kaggle*, o XGBoost frequentemente aparece entre os melhores algoritmos (NIELSEN, 2016).

Os benefícios do XGBoost são especialmente evidentes em cenários com dados desbalanceados ou onde as relações entre as variáveis independentes e a variável dependente são não-lineares. A capacidade de ajustar finamente parâmetros como a taxa de aprendizado (*learning rate*), o número de árvores, e a profundidade máxima das árvores permite que os modelos XGBoost se adaptem a uma ampla gama de problemas de aprendizado supervisionado.

Embora o XGBoost seja um algoritmo poderoso, ele não é isento de limitações. Por exemplo, em casos onde os conjuntos de dados são maiores, os requisitos de memória podem se tornar um obstáculo, apesar das otimizações computacionais (HE et al., 2014). Além disso, enquanto o XGBoost tende a se destacar em dados tabulares, seu desempenho pode não ser tão competitivo em outras modalidades de dados, como os dados não estruturados (áudio, imagem, texto), onde redes neurais profundas são mais eficazes. Em resumo, o XGBoost continua a ser uma ferramenta valiosa no arsenal de técnicas de aprendizado de máquina, combinando eficiência computacional com flexibilidade para lidar com uma ampla variedade de problemas.

2.2.3 Redes Neurais

Redes neurais são modelos computacionais que mimetizam a estrutura e funcionamento do sistema nervoso dos seres vivos, em especial o cérebro humano, com o objetivo de processar informações de maneira similar. Elas são compostas por unidades de processamento chamadas neurônios artificiais, organizados em camadas, que trabalham conjuntamente para realizar tarefas específicas, como reconhecimento de padrões ou processamento de linguagem natural.

Conforme pode-se observar na Figura 5, uma rede neural típica consiste em três componentes principais:

- **Camada de entrada**: Responsável por receber as entradas de dados.
- **Camadas ocultas**: Camadas intermediárias onde ocorrem a maior parte dos cálculos através de conexões ponderadas.
- **Camada de saída**: Produz o resultado do processamento da rede.

Cada neurônio em uma camada está conectado a vários outros na próxima camada, e essas conexões são chamadas de sinapses em uma analogia direta ao cérebro humano. Cada sinapse tem um peso associado que é ajustado durante o treinamento do modelo, permitindo que a rede aprenda a realizar tarefas específicas.

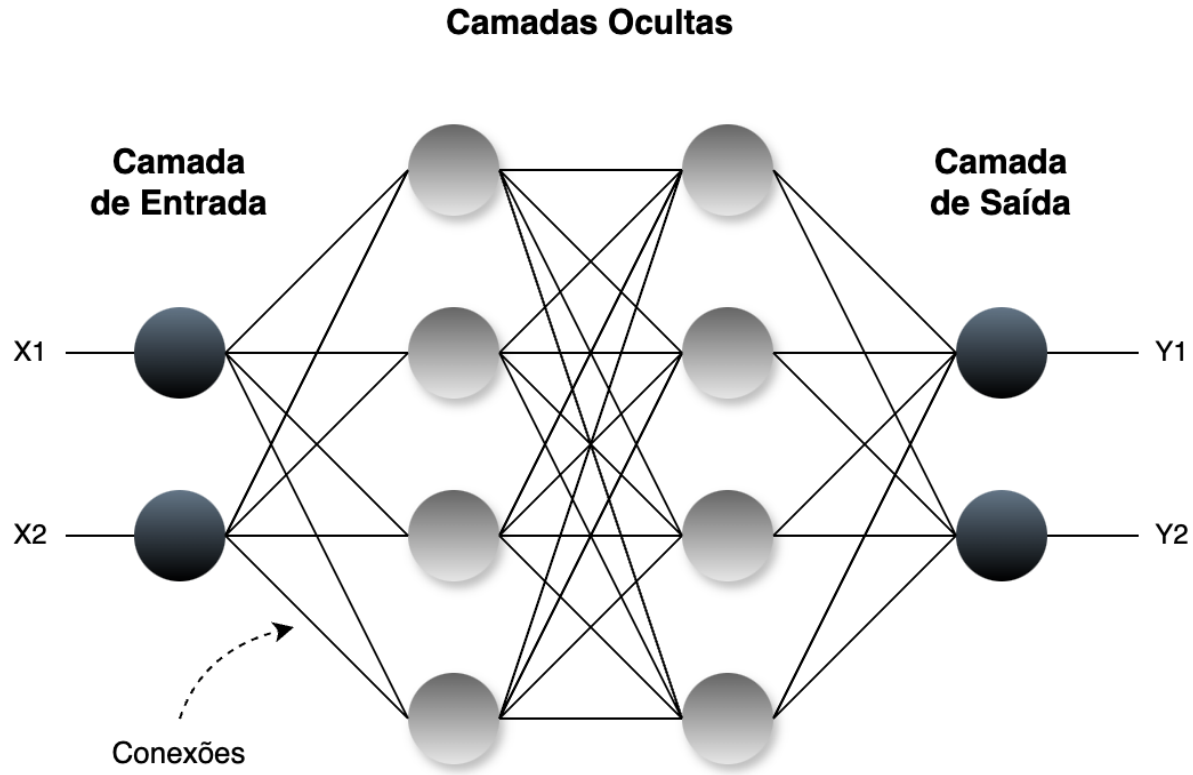

Figura 5 – **Exemplo de uma rede neural.**

No campo do processamento de texto, as redes neurais profundas, ou *deep learning* (redes neurais com muitas camadas intermediárias), desempenham um papel crucial. Elas são capazes de capturar sequências longas de palavras e suas dependências sintáticas e semânticas (*word embeddings*), o que as torna particularmente eficazes para tarefas como tradução automática, reconhecimento de fala, análise de sentimentos, classificação textual.

As redes neurais têm transformado o campo do processamento de texto. Atualmente os modelos de linguagem são baseados na arquitetura *Transformer*, oferecendo melhorias significativas na precisão das classificações e na capacidade de compreender e gerar texto natural de maneira coerente e relevante. Elas são fundamentais em sistemas de resposta automática, assistentes virtuais e outras aplicações interativas que dependem do entendimento profundo da linguagem humana.

2.2.3.1 Arquitetura Transformer

Os modelos de linguagem neurais, como GPT-3 e GPT-4 (*Generative Pre-Trained Transformer*), são conhecidos como modelos de linguagem em larga escala, ou LLMs, devido a

quantidade de parâmetros na arquitetura da rede neural utilizada (Bilhões de parâmetros). Esses modelos são treinados com grandes quantidades de texto e são capazes de realizar uma variedade de tarefas de linguagem, desde tradução automática até geração de conteúdo, com pouca ou nenhuma modificação específica para uma tarefa. Sua capacidade de generalizar a partir de exemplos de treinamento os torna particularmente úteis em educação, onde podem ser adaptados para aplicações como a análise automática de textos de estudantes. Contudo, é necessário o entendimento correto dessa arquitetura tão popular, para a escolha adequada do modelo para o problema a ser resolvido.

A arquitetura *Transformer* (Figura 6), introduzida por (VASWANI et al., 2017), revolucionou o campo do processamento de linguagem natural (PLN) ao introduzir o mecanismo de atenção, que permite que o modelo pondere diferentes partes de uma sequência de entrada de maneira adaptativa. Esta arquitetura é caracterizada pela sua capacidade de lidar com sequências de dados sem a necessidade de recorrência, o que facilita o treinamento em paralelo e melhora a eficiência computacional.

2.2.3.1.1 Modelos Codificadores

Os modelos baseados apenas em encoder (Codificador), como BERT (Bidirectional Encoder Representations from Transformers) e DeBERTa (Decoding-enhanced BERT with Disentangled Attention), utilizam a parte do encoder do Transformer para processar e entender todo o contexto da entrada de uma só vez. Esses modelos são treinados tipicamente com tarefas como preenchimento de lacunas e predições de próximas sentenças, fazendo-os excelentes para tarefas que requerem uma compreensão profunda do contexto da entrada, como classificação de texto, análise de sentimentos e perguntas e respostas (DEVLIN et al., 2018; HE et al., 2021).

O BERT é um modelo de linguagem desenvolvido pela Google que revolucionou o campo do processamento de linguagem natural (PLN). Utilizando uma arquitetura baseada apenas no encoder dos Transformers, o BERT é capaz de capturar o contexto bidirecional de uma frase, o que significa que ele considera tanto a esquerda quanto a direita do token-alvo para entender seu significado. Este modelo é pré-treinado em duas tarefas principais: preenchimento de lacunas (*masked language model*) e predição da próxima sentença (*next sentence prediction*). O treinamento em grande escala em vastos corpora de texto permite ao BERT gerar representações contextualmente ricas que são finamente ajustadas para diversas tarefas específicas. Sua capacidade de compreender o contexto bidirecionalmente o torna um modelo poderoso

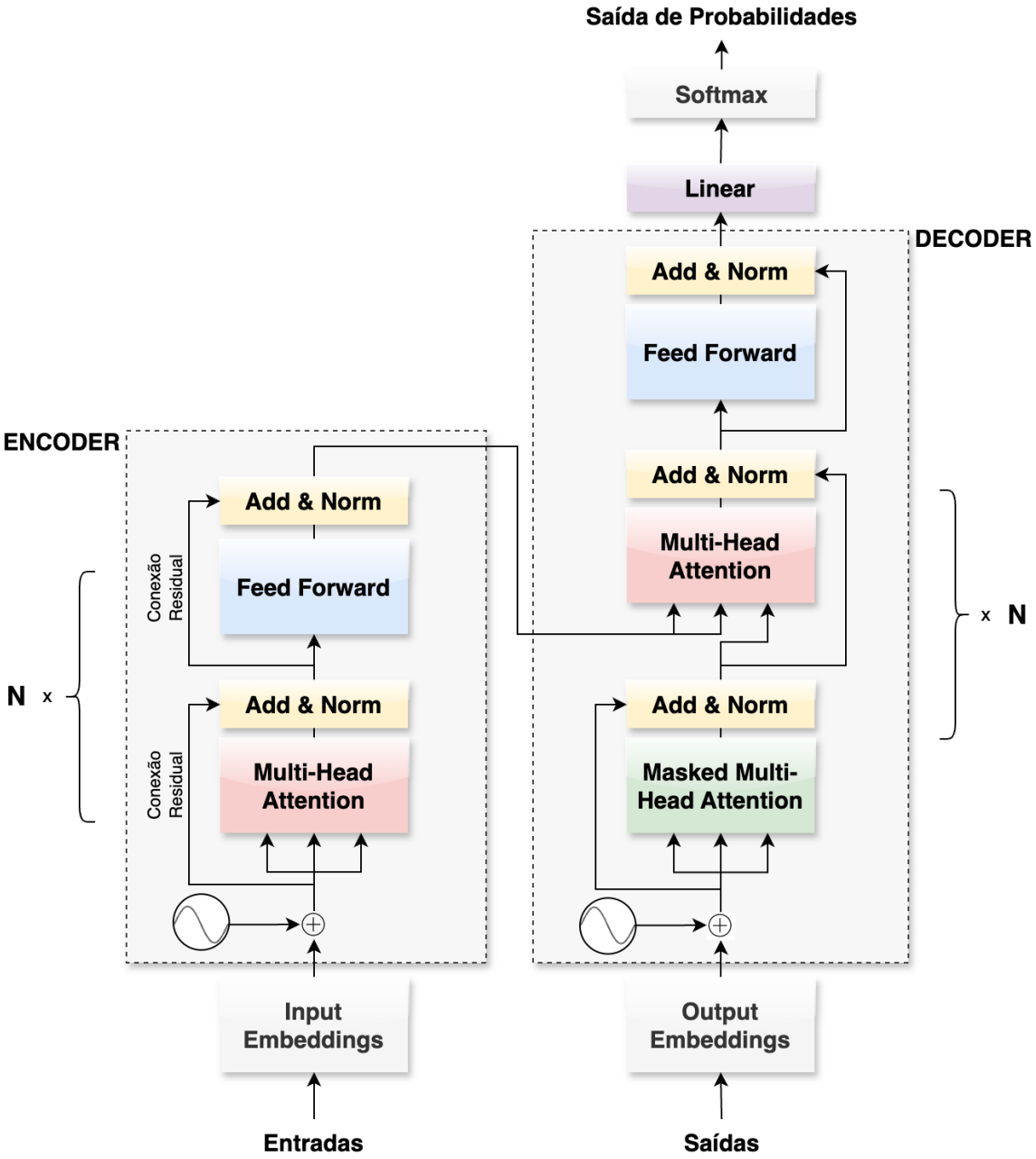

Figura 6 – **Arquitetura Transformer Básica.**

para a análise e interpretação de textos complexos, proporcionando uma base robusta para a construção de sistemas de PLN avançados (DEVLIN et al., 2018).

O DeBERTa, em particular, introduz melhorias significativas em relação ao BERT, principalmente através de sua arquitetura de atenção desmembrada e a introdução de uma máscara de decodificação. A atenção desmembrada permite que o modelo capture melhor as relações entre palavras ao tratar separadamente as contribuições de posição e conteúdo. Isso resulta em representações mais ricas e precisas, especialmente em tarefas que envolvem longos contextos

textuais (HE et al., 2021).

Além disso, o DeBERTa utiliza uma técnica de máscara de decodificação que, durante o pré-treinamento, ajuda o modelo a entender melhor a estrutura da linguagem ao forçar a prever *tokens* em uma sequência específica, em vez de em uma ordem aleatória. Isso aumenta a capacidade do modelo de capturar dependências de longo alcance e contextos complexos que são críticos para a compreensão profunda do texto.

A capacidade do DeBERTa de gerar *embeddings* contextualmente ricos e detalhados torna-o especialmente adequado para tarefas de classificação de texto em ambientes educacionais. Em nosso estudo, o DeBERTa será utilizado para extrair características textuais dos fóruns de discussão, permitindo uma análise detalhada das interações dos alunos. Sua robustez em capturar nuances semânticas e contextuais é crucial para identificar as diferentes categorias da presença cognitiva.

Essas características tornam o DeBERTa uma escolha excelente para nossa abordagem de análise automatizada, proporcionando uma base sólida para a construção de modelos de aprendizagem de máquina que possam classificar com precisão as interações educacionais.

Estas variantes da arquitetura Transformer demonstram a flexibilidade e poder dos modelos baseados em mecanismos de atenção, que continuam a ser uma área de intensa pesquisa e desenvolvimento no campo da inteligência artificial. Trabalhos recentes na área educacional tem empregado esses modelos para classificação de presenças do modelo CoI (BA et al., 2023; LEE et al., 2022)

2.2.3.1.2 Modelos Decodificadores

Por outro lado, as arquiteturas baseadas apenas em decoder (Decodificadores), como GPT-3 e GPT-4, são treinadas para prever a próxima palavra em uma sequência, dada as palavras anteriores. Esta abordagem é conhecida como modelagem autoregressiva de linguagem. Na modelagem autoregressiva, o modelo processa a sequência de entrada de forma sequencial, aprendendo a dependência entre as palavras ao prever iterativamente a próxima palavra, baseada no contexto fornecido pelas palavras anteriores.

Esses modelos são altamente eficazes na geração de texto coerente e contextualmente apropriado. Eles são capazes de produzir longos trechos de texto que mantêm a coesão e a coerência, imitando o estilo e o conteúdo dos dados de treinamento. Tal capacidade de geração

textual é impulsionada pela arquitetura Transformer do decoder, que utiliza mecanismos de autoatenção para capturar dependências de longo alcance e nuances contextuais.

A capacidade de geração textual dos modelos decodificadores como GPT-3 e GPT-4 torna-os particularmente úteis para diversas aplicações práticas. Por exemplo, eles podem ser utilizados para escrita criativa, composição de e-mails, geração de código, criação de diálogos em chatbots, entre outras tarefas que requerem a produção de texto natural e fluido. Além dessas aplicações, uma área de destaque é a aumentação de dados (KUMAR; CHOUDHARY; CHO, 2020).

A aumentação de dados refere-se à prática de expandir o conjunto de dados de treinamento através da geração de novos exemplos sintéticos, o que tem potencial para melhorar o desempenho dos modelos de aprendizagem de máquina, especialmente em situações onde os dados anotados são escassos ou desbalanceados. Modelos como GPT-3 e GPT-4 são particularmente eficazes nessa tarefa devido à sua capacidade de gerar exemplos diversificados e realistas que preservam as características semânticas e sintáticas do conjunto de dados original.

Ao aplicar a aumentação de dados com GPT-3 e GPT-4, podemos criar instâncias adicionais que ajudam a equilibrar as categorias da presença cognitiva nos fóruns de discussão. Isso não apenas melhora a representatividade dos dados de treinamento, mas também permite que os modelos de classificação aprendam melhor as características distintivas de cada categoria. Dessa forma, a geração de texto por modelos decodificadores contribui significativamente para a robustez e a eficácia dos sistemas de análise automatizada, ampliando o alcance e a aplicabilidade das técnicas de aprendizagem de máquina em contextos educacionais e outros domínios (BROWN et al., 2020).

2.2.3.1.3 Modelos Codificadores-Decodificadores

Finalmente, modelos que combinam tanto o encoder quanto o decoder, como o T5 (*Text-to-Text Transfer Transformer*), são projetados para transformar um texto de entrada em um novo texto de saída. Esta configuração é particularmente versátil, permitindo que o modelo seja treinado de maneira unificada para realizar uma ampla variedade de tarefas de PLN, convertendo todas elas em um problema de transformação de texto para texto (RAFFEL et al., 2020). Estes modelos são comumente utilizados em tarefas como tradução automática, sumarização de texto e outras tarefas de PLN.

2.3 CARACTERÍSTICAS LINGUÍSTICAS NA CLASSIFICAÇÃO TEXTUAL

A extração de características linguísticas é um componente fundamental no processo de classificação textual. Ferramentas como LIWC e Coh-Metrix permitem uma análise profunda das propriedades linguísticas, psicológicas e estruturais dos textos, fornecendo informações que vão além dos métodos tradicionais de representação de texto.

2.3.1 LIWC

O *Linguistic Inquiry and Word Count* (LIWC) é uma ferramenta computacional projetada para realizar análises textuais através da categorização de palavras em diversas classes psicolinguísticas, linguísticas, cognitivas e sociais (TAUSCZIK; PENNEBAKER, 2010). O sistema funciona contando palavras em categorias predeterminadas que refletem aspectos importantes do comportamento humano e interações sociais.

2.3.1.1 Funcionamento e Categorias do LIWC

O LIWC analisa textos ao comparar cada palavra do texto com um dicionário interno que associa palavras a categorias específicas. Este dicionário foi desenvolvido e refinado através de múltiplas iterações e revisões por especialistas em linguagem, psicologia e análise textual. As categorias do LIWC abrangem uma ampla gama de características psicológicas e sociais (Tabela 1, incluindo:

- **Funções de Palavras**: Artigos, preposições, auxiliares, etc., que podem dar informações sobre a estrutura gramatical e complexidade do texto.
- **Processos Cognitivos**: Categorias que refletem os processos de pensamento, como causalidade, diferenciação e certeza.
- **Processos Sociais**: Palavras que se referem a interações sociais, incluindo família, amigos e comunicação interpessoal.
- **Afetos**: Dividido em emoções positivas e negativas, ajudando a captar o tom emocional do texto.
- **Biologia**: Termos relacionados à saúde, corpo e sexualidade.

2.3.2 Dimensões do LIWC

A tabela a seguir detalha as principais dimensões do LIWC, proporcionando exemplos de palavras para cada categoria:

Tabela 1 – Dimensões do LIWC e exemplos de palavras

Dimensão	**Exemplos**
I. Dimensões Linguísticas Padrão	
Pronomes	eu, eles, si
Artigos	um, uma, os
Tempo passado	andou, foram, tinha
Tempo presente	é, faz, ouve
Tempo futuro	irá, vai
Preposições	com, acima
Negações	não, nunca, nada
Números	um, trinta, milhão
Palavras de baixo calão	*****
II. Processos Psicológicos	
Processos Sociais	falar, nós, amigo
Amigos	parceiro, colega, companheiro de trabalho
Família	mãe, irmão, primo
Humanos	menino, mulher, grupo
Processos Afetivos	feliz, feio, amargo
Emoções Positivas	feliz, bonito, bom
Emoções Negativas	ódio, inútil, inimigo
Ansiedade	nervoso, medo, tenso
Raiva	ódio, matar, irritado
Tristeza	tristeza, chorar, triste
Processos Cognitivos	causa, saber, dever

Perspicácia	pensar, saber, considerar
Causalidade	porque, efeito, portanto
Discrepância	deveria, poderia
Tentativa	talvez, talvez, suponho
Certeza	sempre, nunca
Inibição	bloquear, restringir
Inclusivo	com, e, incluir
Exclusivo	mas, exceto, sem
Processos Perceptuais	ver, tocar, ouvir
Visão	ver, viu, olhar
Audição	ouviu, ouvir, som
Tato	tocar, segurar, sentiu
Processos Biológicos	comer, sangue, dor
Corpo	dor, coração, tossir
Sexualidade	excitado, amor, incesto
Relatividade	área, curvar, sair, parar
Movimento	andar, mover, ir
Espaço	baixo, dentro, fino
Tempo	hora, dia, hora certa
III. Preocupações Pessoais	
Trabalho	trabalho, aula, chefe
Conquista	tentar, objetivo, ganhar
Lazer	casa, TV, música
Casa	casa, cozinha, gramado
Dinheiro	auditoria, dinheiro, deve
Religião	altar, igreja, mesquita
Morte	enterrar, caixão, matar

IV. Categorias Faladas	
Concordância	concordar, OK, sim
Não fluências	uh, rr*
Enchimentos	blá, você sabe, quero dizer

2.3.2.1 Aplicações do LIWC na Classificação Textual

Na classificação textual, o LIWC é utilizado para extrair features que são particularmente valiosas em contextos onde o subtexto psicológico ou emocional é relevante para a tarefa de classificação. Exemplos de aplicações incluem:

- **Análise de Sentimentos**: Identificar e classificar a valência emocional dos textos, seja como positiva, negativa ou neutra.
- **Detecção de Decepção**: Analisar padrões de fala que podem sugerir tentativas de engano ou omissão.
- **Perfil Psicológico**: Determinar traços de personalidade ou estados mentais dos autores com base no uso de determinadas categorias de palavras.

Essas características tornam o LIWC uma ferramenta poderosa para pesquisadores e profissionais interessados em entender mais profundamente as nuances por trás das palavras usadas em diferentes tipos de comunicações textuais.

Apesar de suas fortes capacidades a interpretação dos resultados exige uma compreensão clara do contexto e das limitações das categorias do dicionário. A eficácia do LIWC pode ser comprometida em textos com uso intensivo de jargão ou linguagem altamente técnica, que podem não estar adequadamente representados no dicionário existente.

Em suma, o LIWC oferece uma perspectiva única sobre a análise de texto, combinando informações linguísticas e psicológicas para aprimorar a classificação e análise textual.

2.3.3 Coh-Metrix

Coh-Metrix é uma ferramenta sofisticada de análise textual desenvolvida para oferecer uma ampla gama de medidas linguísticas e discursivas. Essas medidas proporcionam uma compre-

ensão detalhada da coesão, compreensibilidade, complexidade gramatical e outros aspectos textuais que influenciam como os textos são processados e compreendidos pelos leitores (GRAESSER et al., 2004).

2.3.3.1 Funcionalidades do Coh-Metrix

A ferramenta Coh-Metrix permite aos pesquisadores acessar informações detalhadas sobre a estrutura linguística e o conteúdo de textos escritos (Table 2), incluindo:

- **Índices de Coesão**: Medem como os textos facilitam a compreensão do leitor por meio da ligação lógica e estrutural entre palavras, frases e passagens.
- **Complexidade Gramatical**: Analisa a complexidade da construção de frases e o uso de diferentes estruturas gramaticais.
- **Níveis de Legibilidade**: Utiliza fórmulas específicas para determinar a dificuldade de leitura de um texto, com base em aspectos como comprimento das palavras e das frases.
- **Frequência de Palavras e Densidade de Concretude**: Avalia quão frequentemente termos concretos e abstratos são usados.

Tabela 2 – Características do Coh-Metrix

ID	Dimensão	Descrição
Descritivas		
1	DESPC, READNP	Contagem de parágrafos, número de parágrafos
2	DESSC, READNS	Contagem de sentenças, número de sentenças
3	DESWC, READNW	Contagem de palavras, número de palavras
4	DESPL, READAPL	Comprimento do parágrafo, número de sentenças, média
5	DESPLd	Comprimento do parágrafo, número de sentenças, desvio padrão
6	DESSL, READASL	Comprimento da sentença, número de palavras, média

ID	Dimensão	Descrição
7	DESSLd	Comprimento da sentença, número de palavras, desvio padrão
8	DESWLsy, READASW	Comprimento da palavra, número de sílabas, média
9	DESWLsyd	Comprimento da palavra, número de sílabas, desvio padrão
10	DESWLlt	Comprimento da palavra, número de letras, média
11	DESWLltd	Comprimento da palavra, número de letras, desvio padrão
Facilidade do texto		
12	PCNARz	Pontuação de narratividade, z-score
13	PCNARp	Pontuação de narratividade, percentil
14	PCSYNz	Simplicidade sintática, z-score
15	PCSYNp	Simplicidade sintática, percentil
16	PCCNCz	Concretude de palavra, z-score
17	PCCNCp	Concretude de palavra, percentil
18	PCREFz	Coesão referencial, z-score
19	PCREFp	Coesão referencial, percentil
20	PCDCz	Coesão profunda, z-score
21	PCDCp	Coesão profunda, percentil
22	PCVERBz	Coesão de verbo, z-score
23	PCVERBp	Coesão de verbo, percentil
24	PCCONNz	Conectividade, z-score
25	PCCONNp	Conectividade, percentil
26	PCTEMPz	Temporalidade, z-score
27	PCTEMPp	Temporalidade, percentil
Coesão referencial		

ID	Dimensão	Descrição
28	CRFNO1, CRFBN1um	Sobreposição de substantivos, sentenças adjacentes, binário, média
29	CRFAO1, CRFBA1um	Sobreposição de argumentos, sentenças adjacentes, binário, média
30	CRFSO1, CRFBS1um	Sobreposição de raízes, sentenças adjacentes, binário, média
31	CRFNOa, CRFBNaum	Sobreposição de substantivos, todas as sentenças, binário, média
32	CRFAOa, CRFBAaum	Sobreposição de argumentos, todas as sentenças, binário, média
33	CRFSOa, CRFBSaum	Sobreposição de raízes, todas as sentenças, binário, média
34	CRFCWO1, CRFPC1um	Sobreposição de palavras de conteúdo, sentenças adjacentes, proporcional, média
35	CRFCWO1d	Sobreposição de palavras de conteúdo, sentenças adjacentes, proporcional, desvio padrão
36	CRFCWOa, CRFPCaum	Sobreposição de palavras de conteúdo, todas as sentenças, proporcional, média
37	CRFCWOad	Sobreposição de palavras de conteúdo, todas as sentenças, proporcional, desvio padrão
38	CRFANP1, CREFP1u	Sobreposição de anáforas, sentenças adjacentes
39	CRFANPa, CREFPau	Sobreposição de anáforas, todas as sentenças
LSA		
40	LSASS1, LSAassa	Sobreposição LSA, sentenças adjacentes, média
41	LSASS1d, LSAassd	Sobreposição LSA, sentenças adjacentes, desvio padrão
42	LSASSp, LSAPssa	Sobreposição LSA, todas as sentenças no parágrafo, média

ID	Dimensão	Descrição
43	LSASSpd, LSAPssd	Sobreposição LSA, todas as sentenças no parágrafo, desvio padrão
44	LSAPP1, LSAppa	Sobreposição LSA, parágrafos adjacentes, média
45	LSAPP1d, LSAppd	Sobreposição LSA, parágrafos adjacentes, desvio padrão
46	LSAGN, LSAGN	LSA dado/novo, sentenças, média
47	LSAGNd	LSA dado/novo, sentenças, desvio padrão
Diversidade lexical		
48	LDTTRc, TYPTOKc	Diversidade lexical, taxa de tipo-token, lemas de palavras de conteúdo
49	LDTTRa	Diversidade lexical, taxa de tipo-token, todas as palavras
50	LDMTLDa, LEXDIVTD	Diversidade lexical, MTLD, todas as palavras
51	LDVOCDa, LEXDIVVD	Diversidade lexical, VOCD, todas as palavras
Conectividade		
52	CNCAll, CONi	Incidência de todos os conectivos
53	CNCCaus, CONCAUSi	Incidência de conectivos causais
54	CNCLogic, CONLOGi	Incidência de conectivos lógicos
55	CNCADC, CONADVCONi	Incidência de conectivos adversativos e contrastivos
56	CNCTemp, CONTEMPi	Incidência de conectivos temporais
57	CNCTempx, CONTEMPEXi	Incidência expandida de conectivos temporais
58	CNCAdd, CONADDi	Incidência de conectivos aditivos
59	CNCPos	Incidência de conectivos positivos

ID	Dimensão	Descrição
60	CNCNeg	Incidência de conectivos negativos
Modelo de situação		
61	SMCAUSv, CAUSV	Incidência de verbos causais
62	SMCAUSvp, CAUSVP	Incidência de verbos causais e partículas causais
63	SMINTEp, INTEi	Incidência de verbos intencionais
64	SMCAUSr, CAUSC	Razão de partículas causais para verbos causais
65	SMINTEr, INTEC	Razão de partículas intencionais para verbos intencionais
66	SMCAUSlsa, CAUSLSA	Sobreposição LSA de verbos
67	SMCAUSwn, CAUSWN	Sobreposição WordNet de verbos
68	SMTEMP, TEMPta	Coesão temporal, repetição de tempo e aspecto, média
Complexidade sintática		
69	SYNLE, SYNLE	Embutimento à esquerda, palavras antes do verbo principal, média
70	SYNNP, SYNNP	Número de modificadores por frase nominal, média
71	SYNMEDpos, MEDwtm	Distância de Edição Mínima, parte do discurso
72	SYNMEDwrd, MEDawm	Distância de Edição Mínima, todas as palavras
73	SYNMEDlem, MEDalm	Distância de Edição Mínima, lemas
74	SYNSTRUTa, STRUTa	Similaridade da sintaxe de sentenças, sentenças adjacentes, média
75	SYNSTRUTt, STRUTt	Similaridade da sintaxe de sentenças, todas as combinações, entre parágrafos, média

ID	Dimensão	Descrição
Padrão de densidade sintática		
76	DRNP	Densidade de frase nominal, incidência
77	DRVP	Densidade de frase verbal, incidência
78	DRAP	Densidade de frase adverbial, incidência
79	DRPP	Densidade de frase preposicional, incidência
80	DRPVAL, AGLSPSVi	Densidade de voz passiva sem agente, incidência
81	DRNEG, DENNEGi	Densidade de negação, incidência
82	DRGERUND, GERUNDi	Densidade de gerúndios, incidência
83	DRINF, INFi	Densidade de infinitivos, incidência
Informação sobre as palavras		
84	WRDNOUN, NOUNi	Incidência de substantivos
85	WRDVERB, VERBi	Incidência de verbos
86	WRDADJ, ADJi	Incidência de adjetivos
87	WRDADV, ADVi	Incidência de advérbios
88	WRDPRO, DENPRPi	Incidência de pronomes
89	WRDPRP1s	Incidência de pronomes da primeira pessoa do singular
90	WRDPRP1p	Incidência de pronomes da primeira pessoa do plural
91	WRDPRP2, PRO2i	Incidência de pronomes da segunda pessoa
92	WRDPRP3s	Incidência de pronomes da terceira pessoa do singular
93	WRDPRP3p	Incidência de pronomes da terceira pessoa do plural
94	WRDFRQc, FRCLacwm	Frequência CELEX de palavras de conteúdo, média
95	WRDFRQa, FRCLaewm	Logaritmo da frequência CELEX para todas as palavras, média

ID	Dimensão	Descrição
96	WRDFRQmc, FRCLmcsm	Logaritmo da frequência mínima CELEX para palavras de conteúdo, média
97	WRDAOAc, WRDAacwm	Idade de aquisição para palavras de conteúdo, média
98	WRDFAMc, WRDFacwm	Familiaridade para palavras de conteúdo, média
99	WRDCNCc, WRDCacwm	Concretude para palavras de conteúdo, média
100	WRDIMGc, WRDIacwm	Imaginabilidade para palavras de conteúdo, média
101	WRDMEAc, WRDMacwm	Significância, normas de Colorado, palavras de conteúdo, média
102	WRDPOLc, POLm	Polissemia para palavras de conteúdo, média
103	WRDHYPn, HYNOUNaw	Hiperonímia para substantivos, média
104	WRDHYPv, HYVERBaw	Hiperonímia para verbos, média
105	WRDHYPnv, HYPm	Hiperonímia para substantivos e verbos, média
Legibilidade		
106	RDFRE, READFRE	Facilidade de Leitura Flesch
107	RDFKGL, READFKGL	Nível de Grau Flesch-Kincaid
108	RDL2, L2	Legibilidade L2 Coh-Metrix

2.3.3.2 Aplicações em Educação

Em contextos educacionais, o Coh-Metrix revela-se uma ferramenta valiosa para:

- **Avaliação de Materiais Didáticos**: Professores e designers educacionais podem utilizar o Coh-Metrix para avaliar a adequação dos textos aos níveis de compreensão dos alunos,

garantindo que os materiais de leitura sejam acessíveis e estejam alinhados com as capacidades linguísticas dos estudantes.

- **Desenvolvimento Curricular**: A análise detalhada oferecida pelo Coh-Metrix ajuda na criação de programas de estudos que melhor atendam às necessidades linguísticas e cognitivas dos alunos.

2.3.3.3 Classificação Textual com Coh-Metrix

Na classificação textual, o Coh-Metrix pode ser utilizado para:

- **Diferenciação de Textos**: Identificar diferenças em níveis de formalidade, complexidade ou clareza entre textos, o que é útil para categorizar e filtrar conteúdo em grandes bases de dados textuais.
- **Análise de Consistência Estilística**: Verificar a consistência no uso da linguagem em documentos de um mesmo autor ou de uma mesma organização.

O Coh-Metrix, portanto, oferece um conjunto robusto de ferramentas analíticas que são essenciais para educadores, pesquisadores e profissionais envolvidos com a elaboração, análise e classificação de textos.

2.3.4 Utilização do LIWC e Coh-Metrix na Classificação Textual

Ambas as ferramentas, LIWC e Coh-Metrix, oferecem uma rica fonte de características linguísticas e psicológicas que podem ser exploradas para aprimorar os modelos de classificação textual em diversos contextos. Estas ferramentas analisam diferentes aspectos dos textos, permitindo aplicações variadas e complementares em tarefas de Processamento de Linguagem Natural (PLN).

2.3.4.1 Aplicações do LIWC e Coh-Metrix na Classificação Textual

O LIWC, por ser focado em categorias psicolinguísticas, é bastante útil em tarefas como análise de sentimentos. As categorias do LIWC permitem identificar e quantificar a presença de emoções nos textos, como alegria, tristeza ou raiva. Essas informações são cruciais para

classificar textos com base em sua carga emocional, permitindo, por exemplo, diferenciar automaticamente comentários positivos de negativos em plataformas de redes sociais ou fóruns educacionais (GENG et al., 2020). Além disso, o LIWC pode ser usado para detectar traços de personalidade ou estados psicológicos dos autores com base em suas escolhas lexicais, o que é valioso para a área educacional.

Coh-Metrix, por outro lado, oferece medidas detalhadas de coesão textual e complexidade linguística. Essas medidas podem ser usadas para avaliar a legibilidade de textos ou a adequação de materiais didáticos ao nível de compreensão de diferentes públicos. Em termos de classificação textual, o Coh-Metrix permite distinguir textos com base em seu estilo e formalidade. Por exemplo, pode-se usar a ferramenta para diferenciar entre textos acadêmicos e posts de blog, que geralmente apresentam níveis distintos de complexidade e formalidade. Essa capacidade é especialmente útil em sistemas de recomendação de conteúdo educacional, onde é importante alinhar os recursos de aprendizagem com as capacidades e necessidades dos estudantes.

2.3.5 Utilização em Classificação Textual para Análise do Desenvolvimento Cognitivo

A integração das características extraídas pelo LIWC e pelo Coh-Metrix em modelos de classificação textual pode oferecer uma abordagem sofisticada e precisa na análise automática do desenvolvimento cognitivo do aluno (CoI). O LIWC, por exemplo, permite a identificação e quantificação das emoções expressas nos textos, bem como traços de personalidade ou estados psicológicos dos autores. Essas informações são fundamentais para entender o engajamento emocional e o bem-estar dos alunos ao longo do tempo (KOVANOVIĆ et al., 2016).

Por outro lado, o Coh-Metrix fornece medidas detalhadas de coesão textual e complexidade linguística, que podem ser indicativas do progresso cognitivo e da proficiência linguística dos alunos. Essas medidas podem ser usadas para avaliar a clareza e a compreensibilidade dos textos produzidos pelos alunos, bem como para identificar padrões de desenvolvimento em habilidades de escrita, como vocabulário, estrutura de sentenças e coerência textual (FARROW; MOORE; GAŠEVIĆ, 2019).

Ao integrar as informações fornecidas pelo LIWC e pelo Coh-Metrix em modelos de classificação textual, podemos desenvolver sistemas automatizados capazes de monitorar de forma contínua e não invasiva o desenvolvimento cognitivo dos alunos ao longo do tempo. Esses sis-

temas podem identificar tendências e padrões de comportamento, destacar áreas de melhoria para o planejamento de intervenções pedagógicas personalizadas (FERREIRA et al., 2020).

Por exemplo, um modelo de classificação textual incorporando características do LIWC e do Coh-Metrix pode ser usado para identificar alunos em risco de desengajamento ou desmotivação com base em padrões de linguagem associados a emoções negativas ou à falta de clareza e coesão nos textos produzidos. Essa identificação precoce de alunos em situação de vulnerabilidade pode permitir a implementação de estratégias de apoio e intervenções preventivas, visando promover o engajamento e o sucesso acadêmico (NETO et al., 2018).

Em resumo, a utilização de características linguísticas e psicológicas extraídas pelo LIWC e pelo Coh-Metrix em modelos de classificação textual pode ser valiosa para modelos de aprendizagem de máquina que têm como objetivo a classificação automática do CoI, proporcionando uma análise automática do desenvolvimento cognitivo do aluno. Essa abordagem permite uma compreensão mais holística e individualizada do progresso dos alunos, contribuindo para a promoção de um ambiente de aprendizado mais inclusivo e eficaz.

2.3.6 Social Network Analysis (SNA)

Além das características linguísticas, *features* derivadas da Análise de Redes Sociais (SNA) também são valiosas para a classificação textual, especialmente em ambientes como fóruns online ou redes sociais, onde a interação entre usuários gera estruturas de dados complexas. SNA pode ajudar a identificar influenciadores, comunidades ou padrões de comunicação que são importantes para a classificação de conteúdos, como identificar líderes de opinião ou detectar disseminação de informações (SCOTT, 2017).

Combinando características linguísticas com dados de SNA (grau de centralidade, proximidade, centralidade de intermediação, etc), os pesquisadores podem criar modelos mais robustos e precisos para a classificação textual, proporcionando análises mais profundas e significativas em diversas aplicações.

2.4 MODELO COMUNIDADE DE INVESTIGAÇÃO (COI)

2.4.1 Modelo de Comunidade de Investigação

O modelo de Comunidade de Investigação (CoI), proposto inicialmente por (GARRISON; ANDERSON; ARCHER, 1999), representa um dos frameworks teóricos mais influentes para a compreensão e o desenvolvimento da educação online. Esse modelo descreve a experiência educacional como um processo dinâmico construído sobre três dimensões fundamentais: a presença social, a presença cognitiva e a presença de ensino. Cada uma dessas dimensões aborda aspectos vitais da aprendizagem em ambientes virtuais, oferecendo um framework robusto para analisar e melhorar a interação e o engajamento em comunidades virtuais.

Conforme observado na Figura 7, a interação entre esses três componentes demonstra como eles se sobrepõem para criar uma experiência educacional completa e integrada. Esse modelo integra os componentes de tal forma que: a **presença social** permite aos participantes projetarem suas personalidades e estabelecerem relações pessoais em um ambiente online, criando um clima acolhedor e encorajador para a interação e colaboração; a **presença cognitiva** refere-se à capacidade dos alunos de construir e confirmar significados através da reflexão e do discurso, sendo fundamental para o pensamento crítico e a construção do conhecimento; e a **presença de ensino** envolve o design, a facilitação e a direção de processos cognitivos e sociais, essenciais para orientar os alunos no alcance dos objetivos de aprendizagem desejados, proporcionando suporte e feedback contínuos.

Este modelo não apenas facilita uma análise profunda das experiências educacionais online, mas também serve como uma base para a implementação de práticas pedagógicas que promovam uma aprendizagem eficaz e significativa. Através de sua aplicação, educadores e pesquisadores podem avaliar e aprimorar as interações educativas, garantindo que todos os elementos essenciais da experiência de aprendizado estejam presentes e ativos.

2.4.2 Presença Social

A presença social, conforme definida por (GARRISON; ANDERSON; ARCHER, 1999), é a capacidade dos participantes de se projetarem social e emocionalmente como pessoas reais, através do meio de comunicação utilizado. A presença social é fundamental para a criação de um ambiente de aprendizagem acolhedor e interativo, onde os alunos se sentem confortáveis

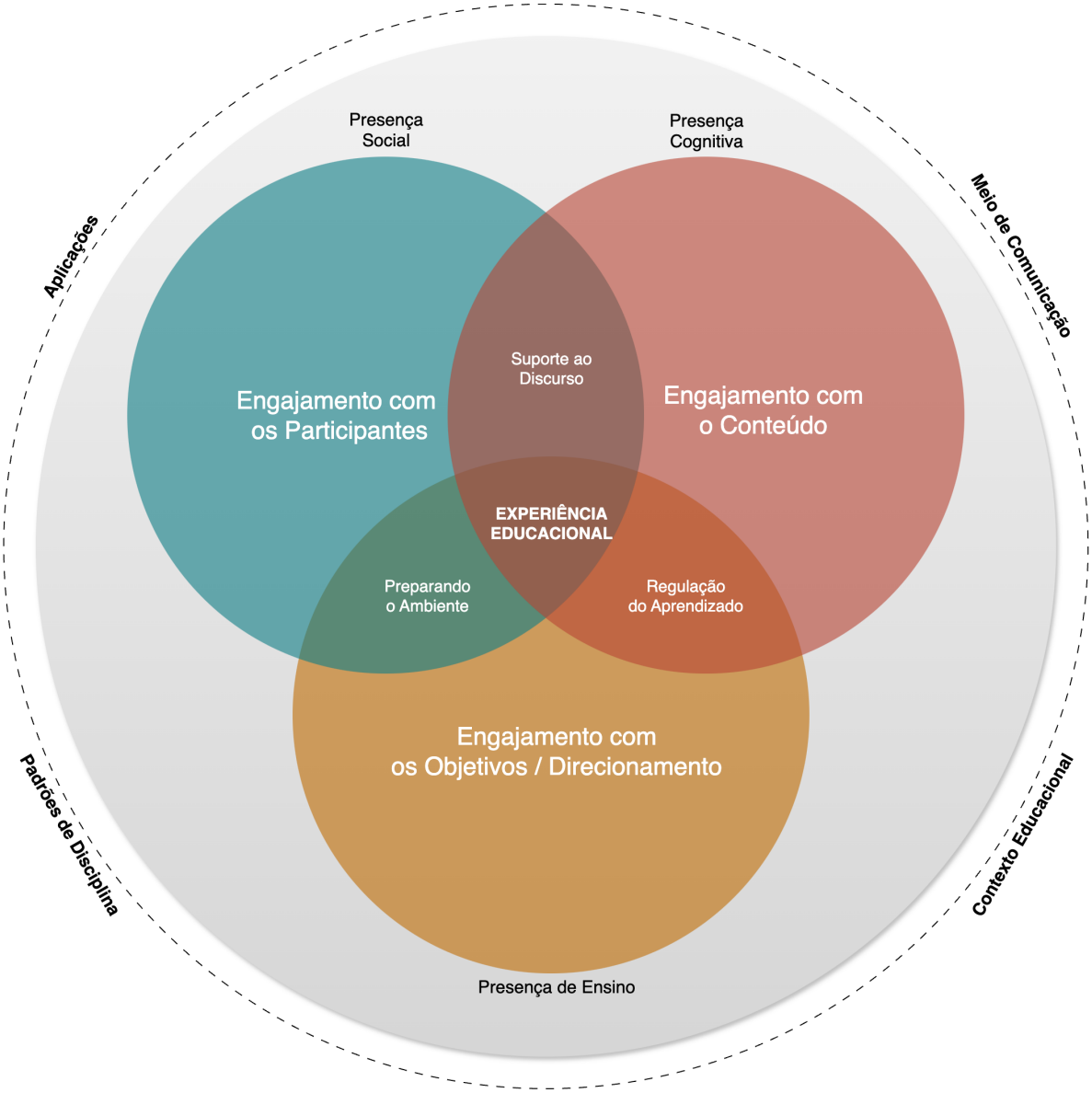

Figura 7 – **Modelo de Comunidade de Investigação.**

para expressar suas ideias e sentimentos. Isso envolve três categorias principais:

- **Afetiva:** Esta categoria analisa como emoções reais são traduzidas em texto, incluindo expressões de emoção, sentimentos e humor. A expressão afetiva é essencial para a humanização das interações online, permitindo que os alunos se conectem emocionalmente uns com os outros. Exemplos incluem o uso de *emoticons*, exclamações e linguagem informal que refletem o estado emocional dos participantes. Estudos mostram que a presença afetiva pode aumentar a motivação e o engajamento dos alunos, contribuindo para um ambiente de aprendizagem mais positivo (ROURKE et al., 1999).

- **Interativa:** Foca na interatividade das mensagens trocadas, com o objetivo de promover uma comunicação aberta entre os estudantes. A presença interativa é evidenciada por atos de comunicação como perguntas, respostas e comentários que demonstram a participação ativa dos alunos nas discussões. Isso inclui a reciprocidade das interações, onde os alunos reconhecem e respondem às contribuições dos colegas, promovendo um diálogo contínuo e construtivo. A interação de alta qualidade está correlacionada com um maior aprendizado percebido e satisfação com o curso (SWAN, 2003).
- **Coesão de Grupo:** Explora o senso de união e compromisso do grupo, fortalecendo a comunidade de aprendizagem. A coesão de grupo é refletida no uso de vocativos, saudações, referências inclusivas (como "nós"e "nosso") e em ações que promovem o espírito de equipe e a solidariedade entre os membros do grupo. Uma forte coesão de grupo pode levar a uma maior colaboração e apoio mútuo, essenciais para a construção de um ambiente de aprendizagem cooperativo e eficaz (GARRISON; ANDERSON; ARCHER, 2001).

A presença social desempenha um papel importante na mediação do aprendizado e no suporte ao desenvolvimento cognitivo em ambientes de educação online. Ao garantir que os alunos se sintam conectados e parte de uma comunidade de aprendizagem, a presença social pode aumentar a motivação, o engajamento e a retenção dos alunos. Além disso, a presença social é um facilitador importante para a presença cognitiva e de ensino, completando o framework CoI e proporcionando uma base sólida para uma experiência educacional rica e integrada (GARRISON; ARBAUGH, 2007).

2.4.3 Presença Cognitiva

A presença cognitiva está relacionada ao desenvolvimento de resultados de aprendizagem desejáveis, como pensamento crítico, resolução de problemas e construção de conhecimento. O modelo de investigação prática, que operacionaliza a presença cognitiva, é composto por quatro fases (GARRISON; ANDERSON; ARCHER, 2001):

1. *Evento Desencadeador:* Representa o início das discussões, onde um problema ou dilema é identificado. Nesta fase, a curiosidade é estimulada e os alunos são motivados a participar da discussão. A identificação de um problema relevante e desafiador é crucial

para engajar os alunos no processo de investigação (GARRISON; ANDERSON; ARCHER, 2001).

2. *Exploração:* Os estudantes exploram soluções possíveis, debatendo ideias e trocando informações. Esta fase é caracterizada pela busca de informações, *brainstorming* e compartilhamento de diferentes perspectivas. A exploração pode envolver a leitura de materiais adicionais, a realização de pesquisas e a discussão colaborativa com colegas. É uma fase que é vital para a construção de uma base sólida de conhecimento sobre o problema identificado (GARRISON, 2003).

3. *Integração:* Os estudantes sintetizam novas ideias e conhecimentos através da construção social. Nesta fase, os alunos começam a conectar as informações coletadas durante a exploração, integrando-as em conceitos e soluções mais coesos. A integração envolve a aplicação de pensamento crítico para analisar e combinar diferentes ideias, resultando em um entendimento mais profundo do problema e das possíveis soluções (ROURKE et al., 2001).

4. *Resolução:* Os estudantes resolvem o dilema inicial, avaliando o conhecimento criado. Esta fase final do modelo de investigação envolve a aplicação prática do conhecimento adquirido para resolver o problema ou dilema identificado na fase inicial. Os alunos avaliam a eficácia das soluções propostas e refletem sobre o processo de aprendizagem. A resolução bem-sucedida promove a confiança dos alunos em suas habilidades de resolução de problemas e pensamento crítico, e reforça a aplicação do conhecimento em contextos reais (GARRISON; ARBAUGH, 2007).

A presença cognitiva é essencial para garantir que os alunos não apenas absorvam informações passivamente, mas se envolvam ativamente na construção e aplicação de conhecimento. Esse engajamento ativo é fundamental para o desenvolvimento de habilidades cognitivas superiores, como a análise crítica e a síntese de informações complexas. Ao focar na presença cognitiva, os educadores podem criar ambientes de aprendizagem que promovem o desenvolvimento intelectual profundo e sustentado dos alunos, preparando-os para enfrentar desafios acadêmicos e profissionais com eficácia (GARRISON; CLEVELAND-INNES; FUNG, 2010).

2.4.4 Presença de Ensino

A presença de ensino abrange o papel dos instrutores antes e durante o curso, incluindo o design do curso, facilitação e instrução direta. Esta presença é crucial para estabelecer um ambiente de aprendizagem estruturado e eficiente, garantindo que os alunos recebam orientações claras e suporte contínuo. Um bom design de curso envolve a definição de objetivos de aprendizagem, a organização do conteúdo de maneira lógica e a criação de atividades que promovam a interação e a colaboração entre os alunos (ANDERSON et al., 2001).

Além do design e da organização, a facilitação do discurso é um aspecto central da presença de ensino. Os instrutores desempenham um papel ativo na promoção de discussões significativas, com o objetivo de aumentar o engajamento social e a coesão de grupo, além de estimular o desenvolvimento cognitivo. Para alcançar esses objetivos, os instrutores incentivam a participação dos alunos e orientam as interações, garantindo que sejam produtivas e relevantes. A facilitação inclui não apenas a moderação das discussões, mas também intervenções estratégicas para manter o foco e a profundidade das conversas, promovendo um ambiente de aprendizado colaborativo e interativo (GARRISON, 2003).

A instrução direta, outro componente essencial da presença de ensino, envolve a entrega de conteúdo especializado, feedback detalhado e suporte individualizado aos alunos. Isso pode incluir a explicação de conceitos complexos, a correção de mal-entendidos e a orientação dos alunos na aplicação prática do conhecimento adquirido. A presença de ensino eficaz assegura que os alunos não apenas adquiram conhecimentos teóricos, mas também desenvolvam habilidades práticas e críticas, essenciais para seu sucesso acadêmico (HOSLER; AREND, 2012).

2.4.5 Análise Automática do CoI

A análise de conteúdo de mensagens em discussões online é amplamente utilizada para examinar as presenças da CoI. Garrison *et al.* (KOVANOVIĆ et al., 2016) desenvolveram esquemas de codificação para avaliar as Presenças Social e Cognitiva, usados tanto na análise manual quanto automática do conteúdo da CoI. No entanto, a codificação manual é trabalhosa e requer codificadores experientes, o que limita sua viabilidade. Estudos iniciais sobre análise automática combinaram contagem de palavras e frases com algoritmos tradicionais de aprendizado de máquina, como SVM e redes neurais (MCKLIN, 2004; CORICH; HUNT; HUNT, 2006), alcançando uma precisão justa, mas com limitações na explicabilidade e generalização

dos modelos (KOVANOVIĆ et al., 2016).

Estudos mais recentes, como os de Kovanović *et al.*(KOVANOVIĆ et al., 2016), utilizaram diferentes conjuntos de características e classificadores, como Coh-Metrix, LIWC, LSA, entidades nomeadas e contexto da discussão, aplicando o algoritmo Random Forest. Essa abordagem obteve 70,3% de acurácia e 0.63 de Cohen's kappa. Neto *et al.*(NETO et al., 2018) seguiram abordagem semelhante para identificar fases da Presença Cognitiva em português, alcançando 83% de precisão e 0.72 de Cohen's kappa. Também exploraram a generalização do modelo em diferentes disciplinas, como Biologia e Tecnologia, com resultados variando de 0.20 a 0.55 de Cohen's kappa, dependendo do conjunto de dados e do processo de treinamento.

Além disso, uma abordagem automática para análise da Presença Social foi proposta recentemente (FERREIRA et al., 2020), combinando Coh-Metrix, LIWC e características tradicionais de mineração de texto. Estudos também têm utilizado métodos de deep learning, como BERT, para identificar a Presença Social, com BERT alcançando os melhores resultados em precisão e F1 (ZOU et al., 2021b; ZOU et al., 2021a).

Estudos demonstraram o valor do Random Forest como algoritmo de "caixa branca" e o uso de diferentes conjuntos de características, como Coh-Metrix e LIWC, na predição da Presença Cognitiva e Social (FERREIRA et al., 2020; BARBOSA et al., 2020; BARBOSA et al., 2021). Entretanto, abordagens mais recentes, como AdaBoost e XGBoost, apresentaram desempenho superior em comparação ao Random Forest (CHEN; GUESTRIN, 2016).

Decidimos, neste trabalho, ter como referência o estudo de Farrow et al. (FARROW; MOORE; GAŠEVIĆ, 2019), que alcançou uma acurácia de 0.60 e um coeficiente de Cohen's Kappa de 0.43, utilizando a mesma base de dados, as mesmas características textuais extraídas e o mesmo conjunto de dados para a análise da presença cognitiva no contexto do modelo de Comunidade de Investigação. A escolha dessa referência se justifica pela metodologia similar empregada, permitindo uma comparação direta entre os resultados obtidos. Essa comparação nos proporciona um *benchmark* sólido para avaliar as melhorias introduzidas pelas técnicas de aumento de dados e pela aplicação de modelos de aprendizagem de máquina mais avançados, como o DeBERTa, que foram explorados nesta pesquisa. Além disso, essa referência é fundamental para validar a eficácia das abordagens propostas e para demonstrar a relevância dos avanços alcançados no contexto de análise automatizada de interações educacionais.

2.5 RESUMO DO CAPÍTULO

Concluindo, este capítulo explorou a Presença de Ensino e a Análise Automática da Comunidade de Investigação em ambientes educacionais online. Destacamos a importância dos instrutores no *design* do curso, facilitação do discurso e instrução direta para criar um ambiente de aprendizagem eficaz. Abordamos a evolução da análise automática da CoI, desde métodos manuais até técnicas modernas de aprendizagem de máquina e *deep learning*. Com base no estudo de Farrow et al. (FARROW; MOORE; GAŠEVIĆ, 2019), estabelecemos um *benchmark* para aprimorar a análise da Presença Cognitiva utilizando técnicas avançadas como o DeBERTa e estratégias de aumento de dados. No próximo capítulo, discutiremos os trabalhos relacionados que fundamentam e contextualizam nossa pesquisa.

3 TRABALHOS RELACIONADOS

O campo da análise automatizada de discussões online, especialmente no contexto educacional, tem recebido crescente atenção nos últimos anos devido à expansão do ensino a distância e ao aumento da utilização dos AVAs. Diversos estudos têm explorado a aplicação de técnicas de processamento de linguagem natural e aprendizagem de máquina para a análise e classificação das interações em fóruns de discussão, visando compreender melhor o desenvolvimento cognitivo dos estudantes e melhorar a eficácia do ensino online.

O estudo de (KOVANOVIć et al., 2014) apresenta uma investigação exploratória sobre a aplicação de técnicas de mineração de texto e classificação de texto para a automação da análise de conteúdo de transcrições de discussões no contexto da educação a distância. A pesquisa utiliza o framework de Comunidade de Investigação, com foco específico no construto de presença cognitiva, devido à sua importância central dentro do modelo CoI. O estudo desenvolveu um classificador que atingiu uma acurácia de 58.4% e um coeficiente Cohen's Kappa de 0.41 em uma tarefa de classificação de 5 categorias, demonstrando o potencial da abordagem proposta. Os autores analisam várias características de classificação, incluindo a marcação de partes do discurso (POS), reconhecimento de entidades nomeadas (NER) e características estruturais, como se uma mensagem é a primeira em um tópico ou uma resposta. Para o treinamento do classificador foi utilizado um conjunto de dados de 1747 mensagens de um curso de pós-gradução em engenharia de software (o mesmo conjunto de dados utilizado neste trabalho, ver Seção 4.1). Apesar das métricas de desempenho modestas, os achados ressaltam a viabilidade de métodos automatizados para análise de conteúdo em contextos educacionais, particularmente em cursos de engenharia de software totalmente online de pós-graduação. A pesquisa utilizou Máquinas de Vetores de Suporte (SVM) para a classificação e empregou validação cruzada para avaliação do modelo.

Neto *et al*. (NETO et al., 2018), apresenta um método para automatizar a análise de conteúdo das mensagens dos alunos em discussões online assíncronas realizadas em português. Esta pesquisa aborda o desafio de codificar transcrições de discussões para a presença cognitiva. Embora existam métodos estabelecidos para codificação da presença cognitiva em inglês, há uma carência de técnicas semelhantes para outros idiomas, particularmente o português. O método proposto utiliza um conjunto de 87 características para desenvolver um classificador Random Forest com o objetivo de extrair automaticamente as diferentes fases da presença

cognitiva. O modelo alcançou um coeficiente Cohen's Kappa de 0.72, indicando um nível substancial de concordância, que supera o limiar de 0.70 comumente aceito para análises de conteúdo quantitativas confiáveis. O estudo utilizou várias características, incluindo características estruturais e *embeddings* de palavras, juntamente com ferramentas como LIWC, Coh-metrix e SpaCy, todas para o idioma português. Amostragem estratificada e *oversampling* com SMOTE foram empregados para lidar com o desbalanceamento de dados, e o modelo foi validado usando validação cruzada, alcançando uma acurácia de 83%. Este trabalho representa uma contribuição significativa para a análise automatizada da presença cognitiva em idiomas diferentes do inglês, oferecendo um método robusto que pode ser aplicado a contextos educacionais semelhantes.

O trabalho (FARROW; MOORE; GAŠEVIĆ, 2019) investiga a robustez e a generalização de métodos automatizados de classificação para a detecção de presença cognitiva em fóruns de discussão online, um componente chave dentro do framework de Comunidade de Investigação. Dada a ampla utilização de fóruns online em ambientes educacionais, esta pesquisa é significativa para compreender como a interação e colaboração dos estudantes podem aprimorar o aprendizado. Os autores focam especificamente na replicação e avaliação crítica de um modelo de ponta para a classificação da presença cognitiva, abordando uma lacuna notável na literatura, onde estudos de replicação são escassos. Utilizando um conjunto de dados de 1747 mensagens de um curso de pós-graduação em engenharia de software (o mesmo conjunto de dados utilizado neste trabalho, ver Seção 4.1), o estudo compara diferentes abordagens para gerenciar o desbalanceamento de classes, particularmente através da aplicação do SMOTE com o classificador Random Forest. Foram utilizadas características linguísticas como do LIWC e Coh-Metrix, entidades nomeadas, e outras características estruturais. Quando os dados não foram rebalanceados, o estudo alcançou uma acurácia de 60% e Cohen's Kappa de 0.43, demonstrando os desafios de lidar com classes desbalanceadas. Com o rebalanceamento com SMOTE, os melhores resultados relatados foram uma acurácia de 62% e Cohen's Kappa de 0.43. A pesquisa destaca os possíveis riscos das práticas comuns de pré-processamento de dados, revelando como elas podem levar a resultados de desempenho excessivamente otimistas. Os autores enfatizam a importância de uma metodologia cuidadosa na análise automatizada de conteúdo para garantir resultados confiáveis e válidos, contribuindo para o discurso mais amplo sobre a confiabilidade das aplicações de aprendizado de máquina na mineração de dados educacionais.

Dando continuidade ao estudo (FARROW; MOORE; GAŠEVIĆ, 2019), o trabalho (FARROW;

MOORE; GAŠEVIĆ, 2020) explora a identificação e modelagem da profundidade e qualidade da participação dos estudantes em fóruns de discussão online, utilizando o conteúdo das mensagens. A pesquisa foca na comparação de dois *frameworks* amplamente estudados para avaliar o discurso crítico e o engajamento cognitivo: os *frameworks* ICAP (Interativo, Construtivo, Ativo, Passivo) e Comunidade de Investigação (presença cognitiva). O objetivo do estudo é descobrir onde esses *frameworks* se alinham e onde oferecem perspectivas complementares sobre o aprendizado. Para alcançar esse objetivo, os autores treinaram classificadores preditivos em um conjunto de dados de 1151 mensagens de um curso de pós-graduação em engenharia de software, totalmente online (o mesmo conjunto de dados utilizado neste trabalho, ver Seção 4.1, contudo apenas as 4 primeiras ofertas do curso). Os classificadores foram utilizados para determinar quais atributos do diálogo são mais preditivos da profundidade e qualidade da participação e como esses atributos se correlacionam com as categorias dos *frameworks*. Os resultados indicam que uma maior profundidade e qualidade de participação estão associadas a mensagens mais longas e complexas em ambos os *frameworks*, sendo que a estrutura das respostas encadeadas é mais significativa do que a ordem temporal das mensagens. No entanto, o estudo também identifica diferenças notáveis, particularmente em como as mensagens de afirmação são tratadas nos *frameworks*. A pesquisa empregou um classificador Random Forest, utilizando características estruturais (LIWC) como o tamanho da discussão, a posição dentro do tópico, e o número de respostas diretas ou indiretas. Foi aplicado oversampling com SMOTE para lidar com o desbalanceamento das classes, e a validação cruzada foi utilizada para garantir a robustez dos resultados. A métrica de desempenho mais alta relatada foi um coeficiente Cohen's Kappa de 0.42 (com rebalanceamento) e 0.41 (sem rebalanceamento), demonstrando uma concordância moderada alcançada pelo classificador na predição da presença cognitiva com base nos atributos do diálogo identificados.

O trabalho de (BARBOSA et al., 2020) explora a viabilidade da classificação automática de presença cognitiva em diferentes idiomas, um componente chave do modelo CoI. Especificamente, a pesquisa foca na classificação de 1500 mensagens de discussão em português utilizando um classificador Random Forest treinado em um corpus de 1747 mensagens de discussão em inglês (o mesmo conjunto de dados utilizado neste trabalho, ver Seção 4.1). O classificador foi construído utilizando um conjunto conciso de 108 indicadores validados que capturam processos psicológicos, coerência linguística e a estrutura das discussões online. O classificador alcançou uma acurácia de 67% e um coeficiente Cohen's Kappa de 0.53, indicando um nível moderado de concordância entre avaliadores. Isso sugere que certos elementos

da presença cognitiva são transferíveis entre idiomas, destacando o potencial de generalização em contextos educacionais multilíngues. O estudo também introduz uma abordagem para lidar com o desbalanceamento de classes por meio de um algoritmo heurístico genético, que demonstrou melhorias em comparação com métodos tradicionais de conjuntos de dados desbalanceados. A pesquisa destaca a viabilidade do uso de características linguísticas e estruturais, como aquelas derivadas do LIWC, Coh-Metrix e indicadores manuais como profundidade da mensagem e contagem de respostas, para classificar a presença cognitiva em diferentes idiomas. Os resultados têm implicações significativas para o desenvolvimento de sistemas automatizados capazes de apoiar a análise educacional multilíngue, particularmente em cursos totalmente online em domínios diversos como engenharia de software e biologia.

O estudo (BARBOSA et al., 2021), investiga a eficácia do uso de tradução automática de textos na classificação de mensagens de discussão online segundo as presenças social e cognitiva. A pesquisa foca na classificação automatizada de 1.500 mensagens em português e 1747 mensagens em inglês (o mesmo conjunto de dados utilizado neste trabalho, ver Seção 4.1), utilizando classificadores treinados em conjuntos de dados tanto antes quanto após a aplicação da tradução dos textos. O estudo encontrou que os classificadores treinados nos textos originais e traduzidos para o inglês alcançaram métricas de desempenho semelhantes às relatadas em estudos anteriores. No entanto, a tradução dos textos para o português resultou em uma diminuição notável no desempenho, indicando que, enquanto a tradução para o inglês é geralmente viável, a tradução para o português apresenta desafios. Essa discrepância foi atribuída à disponibilidade e qualidade dos recursos linguísticos para o português em comparação com o inglês. O estudo ressalta a importância da seleção de características e da qualidade dos recursos, particularmente o uso de ferramentas como LIWC, Coh-Metrix e características estruturais como a similaridade cosseno de LSA, para alcançar uma classificação precisa. Além disso, a pesquisa destaca as limitações dos recursos existentes para o português e enfatiza a necessidade de ferramentas linguísticas mais robustas para línguas que não sejam o inglês. O estudo utilizou um classificador de Random Forest e empregou técnicas como SMOTE para lidar com conjuntos de dados desbalanceados, com validação cruzada utilizada para garantir a confiabilidade dos resultados. O estudo relata o desempenho para a classificação de presença social em textos traduzidos, com uma acurácia de 95% e Cohen's Kappa de 0.92, e classificação de presença cognitiva com uma acurácia de 83% e Cohen's Kappa de 0.69, traduzindo os dados para o inglês. Treinando o classificador com o conjunto de dados em inglês e testando com uma parte do mesmo conjunto, o classificador apresentou acurácia de 58% e Cohen's

Kappa de 0.38.

O trabalho (HU; MELLO; GAŠEVIĆ, 2021) explora o uso de técnicas de deep learning para automatizar a categorização de mensagens de discussão online de acordo com as fases da presença cognitiva. Esta pesquisa investiga não apenas o desempenho de um classificador de Rede Neural Convolucional (CNN), mas também sua generalização e interpretabilidade por meio da aplicação de algoritmos de inteligência artificial explicável (XAI). O estudo compara o desempenho do modelo CNN com abordagens anteriores que utilizaram classificadores de Random Forest e características linguísticas relacionadas a processos psicológicos e coesão (LIWC e Coh-metrix). O classificador CNN, quando treinado e testado em conjuntos de dados formado por 1747 mensagens de discussão em inglês (o mesmo conjunto de dados utilizado neste trabalho, ver Seção 4.1), atingiu uma acurácia de 63% e Cohen's Kappa de 0.48 quando o conjunto de dados em inglês foi dividido em 90%/10% para treinamento e teste, respectivamente. Em comparação, o classificador de Random Forest, com ajuste de parâmetros, alcançou uma acurácia ligeiramente maior de 64% e o mesmo Cohen's Kappa de 0.48 para a mesma divisão de dados. O estudo também destaca os novos *insights* proporcionados pelas visualizações de XAI, que ajudam a identificar as fases da presença cognitiva por meio de indicadores relevantes a nível de palavras, complementando a análise de importância das características dos modelos de Random Forest. Os autores propõem que a combinação de métodos de deep learning, como as CNNs, com algoritmos tradicionais de aprendizado de máquina, como Random Forest, pode servir como abordagens complementares para melhorar a classificação das fases da presença cognitiva em discussões online.

(NETO et al., 2021) examina o impacto de diferentes contextos educacionais na classificação automática de mensagens de discussão online com base na presença cognitiva e evolui o estudo de (NETO et al., 2018). Esta pesquisa analisa especificamente mensagens de discussão online escritas em português brasileiro de dois cursos distintos, um em biologia e outro em tecnologia, cada um com diferentes níveis de presença de ensino nas discussões online. O estudo utilizou um classificador Random Forest e um conjunto de 127 características (LIWC, Coh-metrix, etc), incluindo características estruturais e orientadas por dados, para identificar automaticamente as fases da presença cognitiva nessas mensagens. Os resultados revelaram que o classificador teve um desempenho melhor quando aplicado a todo o conjunto de dados, sugerindo que um modelo treinado em dados de um curso específico pode não ter a generalização necessária para ser aplicado de maneira eficaz a cursos de diferentes áreas. Os resultados também destacaram a importância de certas características preditivas no reconhecimento da presença cognitiva em

diferentes contextos educacionais. O classificador alcançou uma acurácia de 76% e um coeficiente Cohen's Kappa de 0.55, demonstrando uma confiabilidade moderada. Com base nesses achados, o estudo sugere que trabalhos futuros devem utilizar um conjunto de características semelhante, mas treinar o classificador em conjuntos de dados de áreas de conhecimento intimamente relacionadas aos tópicos de discussão, para aprimorar a generalização e a eficácia do modelo em diferentes domínios educacionais.

A pesquisa de Ba *et al.* (BA et al., 2023) utilizou uma metodologia abrangente em duas partes, envolvendo análise de conteúdo automatizada e análise de rede epistêmica (ENA), para avaliar a presença cognitiva em discussões online em três cursos distintos. O processo de análise automática começou com a coleta e codificação manual das transcrições das discussões de dois cursos (13995 sentenças em inglês), que foram então usadas para desenvolver e avaliar modelos de classificação de textos. Foram exploradas várias técnicas de representação de texto, incluindo TF-IDF, Word2Vec e BERT, com os modelos baseados em BERT sendo ainda aprimorados pela integração de características linguísticas como contagem de palavras, diversidade lexical e entidades nomeadas. Os classificadores foram treinados e testados em uma divisão de dados de 90%/10%, e múltiplos modelos foram comparados, incluindo Regressão Logística, Naive Bayes, Random Forest e uma Rede Neural Profunda. Entre eles, os classificadores baseados em BERT superaram significativamente os outros, especialmente quando combinados com características linguísticas, alcançando um Cohen's Kappa de 0.76, indicando um alto nível de concordância com os codificadores humanos. Esse desempenho foi atribuído ao pré-treinamento sensível ao contexto do BERT, que permitiu capturar melhor as nuances das fases de presença cognitiva em comparação com modelos mais tradicionais. Em seguida, o melhor classificador baseado em BERT foi empregado para codificar os dados de discussão de um terceiro curso (5110 sentenças em inglês), que foram então analisados usando ENA. A ENA permitiu modelar a presença cognitiva mapeando as coocorrências dos elementos codificados nas discussões, tanto a nível individual quanto de grupo.

Escolhemos o trabalho de Farrow *et al.* (FARROW; MOORE; GAŠEVIĆ, 2019) como referência central para este estudo devido à sua relevância metodológica e à similaridade com nossa abordagem em termos de conjunto de dados, divisão de treinamento, características extraídas e uso do classificador Random Forest. Esse estudo utilizou o mesmo conjunto de dados, proveniente de um curso de pós-graduação em engenharia de software, o que permite uma comparação direta e robusta dos resultados. Além disso, a divisão do conjunto de treinamento foi realizada de maneira idêntica, garantindo consistência nas condições experimentais. A

escolha das características extraídas, bem como a utilização do Random Forest, reflete um alinhamento metodológico que facilita a replicação e a validação dos resultados. Ademais, o estudo (FARROW; MOORE; GAŠEVIĆ, 2019) apresenta resultados reportados tanto com quanto sem o rebalanceamento dos dados, proporcionando uma visão abrangente do impacto do desbalanceamento na performance do classificador. Essa abordagem detalhada e comparável fortalece a justificativa para a escolha desse trabalho como uma referência fundamental em nossa pesquisa. Além disso, o trabalho (BA et al., 2023) sugere que a utilização de modelos Transformers como o BERT pode melhorar a performance na classificação das categorias da presença cognitiva. Portanto, conduzimos experimentos que demonstram a eficácia dessa abordagem no nosso cenário.

Nosso trabalho se destaca ao explorar abordagens inovadoras de aprendizado ativo para lidar com a escassez de dados anotados, um dos principais desafios na classificação das categorias da presença cognitiva. Diferente de estudos anteriores que se concentraram em métodos tradicionais de classificação, como o Random Forest e o XGBoost, adotamos uma estratégia que integra o aprendizado ativo, permitindo a seleção das instâncias mais informativas para anotação adicional. Essa abordagem não apenas otimiza o uso dos dados disponíveis, mas também reduz significativamente a necessidade de grandes volumes de dados anotados, viabilizando a aplicação prática desses modelos em ambientes reais de ensino à distância.

Além disso, nosso trabalho avança ao treinar um modelo MLP utilizando DeBERTa, um dos modelos de codificação mais avançados disponíveis atualmente. Essa escolha reflete nosso compromisso em explorar o estado da arte em representação textual, garantindo que a análise das categorias da presença cognitiva seja precisa e detalhada. Complementarmente, investigamos cenários com e sem aumentação de dados utilizando o GPT, proporcionando uma compreensão mais ampla dos impactos dessas técnicas na acurácia dos modelos, e oferecendo uma base sólida para futuras implementações em contextos educacionais diversos. A Tabela 3 apresenta um comparativo de trabalhos que analisam a presença cognitiva.

Tabela 3 – Comparação de Trabalhos Relacionados

Trabalho	**Objetivo**	**Técnicas Utilizadas**	**Conjunto de Dados**	**Métricas**
Kovanović et al. (2014)	Analisar a presença cognitiva em discussões online.	SVM com características linguísticas e estruturais.	1747 mensagens de um curso online.	Acurácia: 58.4%, Kappa: 0.41.
Neto et al. (2018)	Codificar presença cognitiva em português.	Random Forest com 87 características, LIWC, SpaCy.	Mensagens de um curso online em português.	Acurácia: 83%, Kappa: 0.72.
Farrow et al. (2019)	Avaliar métodos de classificação da presença cognitiva.	Random Forest, SMOTE, características de LIWC e Coh-Metrix.	1747 mensagens de um curso online.	Acurácia: 62%, Kappa: 0.43.
Barbosa et al. (2020)	Classificar presença cognitiva em diferentes idiomas.	Random Forest com 108 indicadores linguísticos.	1500 mensagens em português, 1747 em inglês.	Acurácia: 67%, Kappa: 0.53.
Neto et al. (2021)	Analisar impacto do contexto educacional na classificação	Random Forest, LIWC	M127 mensagens de umdois cursos online em português	Acurácia: 76%, Kappa: 0,55
Hu et al. (2021)	Aplicar deep learning para classificar presença cognitiva.	CNN, Random Forest, XAI.	1747 mensagens de um curso online.	Acurácia: 64%, Kappa: 0.48.
Ba et al. (2023)	Avaliar presença cognitiva usando BERT e ENA.	BERT, TF-IDF, Word2Vec, Regressão Logística, ENA.	13995 sentenças de discussões online.	Kappa: 0.76 (BERT).
Nossa Proposta	Avaliar métodos de classificação da presença cognitiva.	MLP com DeBERTa, aumento de dados com GPT.	1747 mensagens de um curso online.	Resultados experimentais variados.

3.1 RESUMO DO CAPÍTULO

Neste capítulo, revisamos os principais trabalhos relacionados à classificação automatizada da presença cognitiva em discussões online educacionais. Estudos anteriores aplicaram técnicas de processamento de linguagem natural e aprendizado de máquina, como SVM, Random Forest e modelos de *deep learning*, obtendo diferentes níveis de acurácia e coeficientes Kappa. Observamos que, apesar dos avanços, persistem desafios na generalização dos modelos para diferentes contextos e idiomas.

Nosso trabalho diferencia-se ao explorar o aprendizado ativo para lidar com a escassez de dados anotados e ao utilizar modelos avançados de codificação textual, como o DeBERTa, treinando um modelo MLP. Além disso, investigamos o impacto da aumentação de dados com o GPT-4 na acurácia dos modelos, buscando contribuir para o desenvolvimento de sistemas automatizados mais eficazes na análise da presença cognitiva em ambientes de aprendizagem online.

No próximo capítulo, apresentaremos a metodologia e os experimentos realizados.

4 METODOLOGIA

Neste capítulo, detalhamos a metodologia adotada para investigar a viabilidade e eficácia de técnicas automáticas de processamento de linguagem natural e análise de redes no contexto de ambientes virtuais de aprendizagem para a classificação automática da presença cognitiva do CoI. A metodologia proposta é descrita em etapas sequenciais integrando diversas abordagens, começando com a coleta e preparação dos dados, seguida pela extração de características e o treinamento de modelos de classificação. Após a implementação dos modelos, procedemos com a avaliação dos resultados de diversos experimentos que empregam diferentes arquiteturas e conjuntos de características. Essa fase de avaliação é importante para entender como as variáveis e as técnicas selecionadas influenciam a precisão, a robustez e a aplicabilidade dos modelos em contextos educacionais reais.

4.1 COLETA E PREPARAÇÃO DOS DADOS

O conjunto de dados utilizado neste trabalho foi escolhido especificamente pela sua robustez teórica e pela ampla utilização em diversos trabalhos de referência no campo de *learning analytics*, conforme mencionado em alguns desses estudos no Capítulo 3.

O conjunto de dados foi extraído de seis ofertas de um curso intensivo de mestrado em engenharia de software oferecido totalmente online por uma universidade pública canadense. Durante as seis ofertas do curso (inverno de 2008, outono de 2008, verão de 2009, outono de 2009, inverno de 2010, inverno de 2011), 81 alunos produziram 1747 mensagens. Os alunos gravaram e participaram de discussões online assíncronas sobre apresentações em vídeo relacionadas a artigos de pesquisa sobre um dos tópicos do curso (ou seja, requisitos de software, design e manutenção). Tal atividade contabilizou 15% da nota final do curso.

O conjunto de dados foi anotado de acordo com os indicadores da presença social e as fases da presença cognitiva. Utilizando o esquema de codificação proposto por (GARRISON; ANDERSON; ARCHER, 2001), dois codificadores experientes anotaram todas as mensagens de discussão online de acordo com as fases da presença cognitiva. Na anotação o acordo interavaliador entre os codificadores foi excelente (Percentual de acordo=98.1% e κ=0.974), com apenas 32 discordâncias. As diferenças foram resolvidas através de discussões entre os dois codificadores. A Tabela 4 mostra a distribuição final das mensagens de acordo com as fases

Tabela 4 – Distribuição das fases da presença cognitiva.

ID	Fase	Mensagens	%
0	*Outros*	140	8.01%
1	Evento Desencadeador	308	17.63%
2	Exploração	684	39.15%
3	Integração	508	29.08%
4	Resolução	107	6.13%
	Total	1747	100.00%

da presença cognitiva.

O conjunto de dados foi dividido em conjuntos de treinamento e teste (escolhidos aleatoriamente) seguindo o trabalho de (FARROW; MOORE; GAŠEVIĆ, 2019).

4.2 EXTRAÇÃO DE CARACTERÍSTICAS

A extração de características é uma etapa importante na análise de dados coletados de fóruns de discussão, sendo essencial para a classificação e avaliação da presença cognitiva. Com uma abordagem integrada que combina métodos de diferentes disciplinas, conseguimos capturar e quantificar os padrões de interação que são fundamentais para entender os processos de aprendizagem em ambientes virtuais.

Este trabalho integra características tradicionais de mineração de texto, como a contagem de entidades nomeadas e métricas de características da mensagem (incluindo similaridade com mensagens anteriores e subsequentes, contagem de respostas, identificação de mensagem inicial, entre outras), com as ferramentas linguísticas LIWC e Coh-Metrix para extrair indicadores de presença cognitiva a partir de contribuições textuais. Essas ferramentas são amplamente reconhecidas na literatura como extratoras eficazes de aspectos coesivos, psicológicos e sociais dos textos (PENNEBAKER et al., 2015)

Além das características linguísticas, este trabalho propõe a utilização da representação textual baseada em *embeddings* extraídos com a utilização de modelos baseados na arquitetura *Transformer*, como BERT e GPT.

Em resumo, este trabalho utilizou duas abordagens distintas para a representação de dados textuais em fóruns de discussão online. A abordagem tradicional, baseada em ferramentas como LIWC e Coh-Metrix, focou na extração de 209 características coesivas, psicológicas e sociais, entre outras (ver Tabela 5. A outra abordagem utilizou *embeddings* de modelos *Transformers*

como representação textual; esses *embeddings* geraram vetores de 768 dimensões.

Tabela 5 – Características extraídas e organizadas por categoria.

Categoria	Características	Total
LIWC	Analytic, Clout, Authentic, Tone, Sixltr, Dic, function, pronoun, ppron, i, we, you, shehe, they, ipron, article, prep, auxverb, adverb, conj, negate, verb, adj, compare, interrog, number, quant, affect, posemo, negemo, anx, anger, sad, social, family, friend, female, male, cogproc, insight, cause, discrep, tentat, certain, differ, percept, see, hear, feel, bio, body, health, sexual, ingest, drives, affiliation, achieve, power, reward, risk, focuspast, focuspresent, focusfuture, relativ, motion, space, time, work, leisure, home, money, relig, death, informal, swear, netspeak, assent, nonflu, filler, AllPunc, Period, Comma, Colon, SemiC, QMark, Exclam, Dash, Quote, Apostro, Parenth, OtherP	93
Coh-Metrix	DESWLsy, PCCNCz, PCVERBz, DESWLltd, WRDPOLc, WRDPRO, LSASSp, PCCNCp, CRFAO1, CNCPos, SMCAUSlsa, CRFSO1, SMCAUSwn, WRDADJ, DESWLsyd, DESSL, DESPLd, SYNMEDlem, SYNLE, DESSC, WRDHYPn, CRFCWO1, DRVP, CNCTempx, PCREFp, PCSYNz, SYNMEDwrd, DRGERUND, CNCADC, WRDMEAc, WRDPRP3s, RDFKGL, PCNARp, SYNNP, LDTTRa, WRDPRP1p, CNCAll, WRDPRP3p, PCTEMPz, LDVOCD, CNCNeg, DRPVAL, WRDPRP1s, CNCTemp, SMTEMP, LSASS1, PCCONNz, WRDPRP2, DESSLd, DRINF, WRDAOAc, RDFRE, DESPC, PCVERBp, DESWLlt, SMINTEr, DRNP, CNCCaus, CRFCWOa, CRFNOa, DRPP, LSASS1d, WRDNOUN, LSASSpd, CNCAdd, WRDCNCc, CNCLogic, DESWC, WRDFRQc, DRNEG, WRDVERB, LDMTLD, PCREFz, DRAP, PCNARz, SMCAUSvp, SMINTEp, SYNMEDpos, RDL2, WRDHYPv, LDTTRc, PCSYNp, WRDADV, WRDFRQmc, LSAPP1, WRDFAMc, LSAGN, WRDHYPnv, WRDIMGc, PCCONNp, CRFNO1, DESPL, PCDCz, LSAPP1d, SMCAUSv, SMCAUSr, CRFAOa, LSAGNd, CRFCWOad, WRDFRQa, SYNSTRUTt, CRFSOa, PCDCp, SYNSTRUTa, PCTEMPp, CRFCWO1d	108
NER	ner.entity.cnt	1
Outras	message.reply.cnt, message.depth, message.sim.prev, message.sim.next, message.is.first, message.is.last, lsa.similarity	7

4.3 ALGORITMOS UTILIZADOS

Dentre os algoritmos e modelos de linguagem empregados nos experimentos deste trabalho, destacam-se aqueles que foram fundamentais para a tarefa de classificação das categorias da presença cognitiva em fóruns de discussão online.

Os algoritmos baseados em árvore foram amplamente utilizados nesta pesquisa devido à sua capacidade de modelar relações complexas nos dados e lidar com o desbalanceamento de classes. Entre esses algoritmos, destacam-se:

- **Random Forest**[1]: Um método de *ensemble* que constrói múltiplas árvores de decisão durante o treinamento e as combina para produzir um modelo mais robusto e preciso. O Random Forest foi utilizado em vários experimentos devido à sua eficácia em lidar com dados ruidosos e desbalanceados, além de apresentar bom desempenho na classificação das presenças do modelo CoI (veja o Capítulo 3).
- **XGBoost (Extreme Gradient Boosting)**[2]: Um algoritmo de *boosting* que combina várias árvores de decisão em sequência, onde cada árvore corrige os erros das anteriores. O XGBoost é conhecido por seu desempenho superior em competições de ciência de dados, sendo particularmente eficaz em cenários com dados estruturados.

Além destes, outros algoritmos também foram explorados, como ExtraTrees (*Extremely Randomized Trees*), LightGBM (*Light Gradient Boosting Machine*), CatBoost e KNeighbors.

Além dos algoritmos baseados em árvore e do KNeighbors, destacamos a utilização de redes neurais. O **MLP (Perceptron Multicamadas)** é uma rede neural *feedforward* que consiste em várias camadas de nós, permitindo capturar complexidades não-lineares nos dados. O MLP foi utilizado para investigar as capacidades de redes neurais na classificação de texto, especialmente quando combinado com representações textuais avançadas (*embeddings*).

Por fim, modelos de linguagem foram utilizados para melhorar a representação dos textos e mitigar o problema da escassez de dados:

- **BERT**[3]: Utilizado para aumentação de dados por meio da substituição de palavras no texto. O BERT compreende o contexto das palavras em ambas as direções e gera *embeddings* contextuais profundos, melhorando a representação dos dados textuais.

[1] https://scikit-learn.org/stable/
[2] https://xgboost.readthedocs.io/en/stable/
[3] https://nlpaug.readthedocs.io/en/latest/overview/overview.html

- **GPT-4**[4]: Um modelo de linguagem autoregressivo utilizado para gerar textos sintéticos que preservam o contexto e o estilo dos dados originais, auxiliando na mitigação do desbalanceamento e da escassez de dados anotados.

- **DeBERTa**[5]: Um modelo de linguagem avançado que melhora o BERT com uma atenção descentralizada e uma codificação de posição relativa. Foi utilizado como *encoder* nos experimentos, capturando informações contextuais e semânticas profundas.

Esses algoritmos e modelos foram escolhidos para explorar diferentes abordagens na classificação automática das categorias da presença cognitiva.

4.4 MÉTRICAS DE AVALIAÇÃO

Para avaliar a eficácia dos modelos de classificação das presenças social e cognitiva no contexto do Community of Inquiry (CoI), utilizamos duas métricas principais: acurácia e o coeficiente de Cohen's Kappa. A acurácia é uma das métricas mais diretas para medir o desempenho de um modelo de classificação, representando a proporção de classificações corretas feitas pelo modelo em relação ao total de casos testados. Em outros termos, é o total de acertos sobre o total de casos (Equação 4.1). Essa métrica fornece uma visão geral rápida da eficiência do modelo em classificar corretamente as presenças nas discussões online. No entanto, em conjuntos de dados desbalanceados, onde as classes podem ter representações desiguais, a acurácia pode não refletir completamente o desempenho do modelo.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.1}$$

Para complementar a avaliação da acurácia, empregamos o coeficiente de *Cohen's Kappa*, que mede a concordância entre duas avaliações que classificam itens em categorias mutuamente exclusivas, ajustando-se pela concordância que ocorreria por acaso (COHEN, 1960). Na Equação 4.2, o valor P_o representa a proporção de concordância observada entre os avaliadores, e o P_e é a proporção de concordância que seria esperada por acaso. O *Kappa* é particularmente útil em contextos educacionais onde as classificações podem ser subjetivas e onde é crucial diferenciar a concordância real daquela que poderia ocorrer aleatoriamente (MCHUGH, 2012).

[4] https://platform.openai.com/docs/api-reference?lang=python
[5] https://huggingface.co/

Este coeficiente oferece uma medida robusta da consistência interna dos modelos ao classificar as presenças definidas pelo CoI, considerando tanto a presença social quanto a cognitiva.

$$\textit{Cohen's Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{4.2}$$

Utilizando essas métricas, podemos avaliar não apenas a precisão dos modelos em termos de classificação correta, mas também sua capacidade de realizar essas classificações de maneira consistente em diferentes conjuntos de dados e configurações de teste. Além disso, a utilização dessas métricas nos permitirá comparar os resultados com outros trabalhos relevantes da área que utilizaram o mesmo conjunto de dados (FARROW; MOORE; GAŠEVIĆ, 2019; FERREIRA et al., 2020; KOVANOVIć et al., 2014).

É importante destacar que a escolha dessas métricas para avaliar os modelos está alinhada com trabalhos de referência na área. A utilização das mesmas métricas adotadas em estudos anteriores permite a comparação dos resultados e facilita a verificação de avanços no estado da arte (veja o Capítulo 3).

4.5 DESCRIÇÃO DOS EXPERIMENTOS

A sequência dos experimentos foi cuidadosamente planejada para explorar, de forma incremental, diferentes abordagens na classificação das categorias de presença cognitiva em fóruns de discussão online, visando aprimorar o desempenho dos modelos e superar desafios específicos dos dados.

Iniciamos com o Experimento 1, utilizando características textuais tradicionais extraídas por ferramentas como LIWC, Coh-Metrix e SNA, e empregamos o AutoGluon para selecionar automaticamente o melhor modelo de classificação. Essa etapa estabeleceu uma linha de base com métodos convencionais e automatizou a seleção de modelos, permitindo uma avaliação inicial do desempenho sem intervenção manual intensiva.

No Experimento 2, direcionamos o foco para modelos baseados em árvores (Random Forest e XGBoost) incorporando aprendizado ativo, o que nos permitiu simular cenários com dados anotados limitados e otimizar o modelo adicionando iterativamente novas instâncias com base na incerteza das previsões.

Reconhecendo o desequilíbrio das classes no conjunto de dados, o Experimento 3 introduziu técnicas de aumento de dados usando nlpaug e BERT para equilibrar as categorias menos

representadas, mantendo o uso de modelos baseados em árvores para avaliar o impacto do balanceamento no desempenho.

Avançando para abordagens de aprendizado profundo, o Experimento 4 empregou embeddings gerados pelo modelo DeBERTa, eliminando a necessidade de engenharia de características manuais e explorando representações contextuais mais ricas por meio de um MLP. Por fim, o Experimento 5 combinou a capacidade avançada do GPT-4 para aumento de dados com os embeddings do DeBERTa, buscando enriquecer ainda mais o conjunto de dados e potencializar o desempenho do modelo.

Essa progressão lógica, do tradicional ao avançado, e a abordagem sistemática para enfrentar desafios como desequilíbrio e escassez de dados, permitiram uma avaliação abrangente das estratégias e seu impacto na acurácia da classificação, conforme será detalhado a seguir.

4.5.1 Experimento 1: Classificação das Categorias da Presença Cognitiva com AutoML

4.5.1.1 Características Extraídas

O LIWC é um recurso de análise textual linguística que extrai 93 características divididas nas seguintes categorias: resumo de variáveis linguísticas, dimensões linguísticas, gramática e processos psicológicos. A última categoria inclui palavras que expressam Processos Afetivos, Emoção Positiva, Emoção Negativa, Raiva, Tristeza, Processos Sociais, entre outros. Relacionado a definição de presença cognitiva, hipotetizamos que o uso deste recurso linguístico pode contribuir para a construção de classificadores capazes de discriminar corretamente mensagens com ou sem evidência de presença cognitiva. O LIWC já foi utilizado em um estudo anterior focado na automação da identificação de presença cognitiva, alcançando altos níveis de acurácia (KOVANOVIć et al., 2016; NETO et al., 2018). Assim, neste estudo, a versão LIWC 2015 foi usada para extrair características das mensagens em nosso conjunto de dados.

De acordo com (GRAESSER et al., 2004), o Coh-Metrix utiliza léxicos, POS Tagger, LSA, entre outras técnicas de processamento de linguagem natural (PLN) para analisar coesão e coerência textual. Vários estudos relataram bons resultados ao usar o Coh-Metrix para gerar indicadores de coesão textual (MCCARTHY et al., 2006; GRAESSER; MCNAMARA; KULIKOWICH, 2011). Hipotetizamos que a existência/ausência de indicadores da presença cognitiva poderia estar relacionada ao índice de coesão e complexidade textual proposto no Coh-Metrix. Ao con-

siderar a utilização dessa ferramenta para identificação automática das categorias da presença cognitiva, características como diversidade léxica, número de tokens, entre outras, podem ser um bom indicativo do estágio do desenvolvimento cognitivo do aluno (KOVANOVIć et al., 2016; FARROW; MOORE; GAŠEVIĆ, 2019).

A análise de redes sociais (SNA, do inglês *Social Network Analysis*) tem sido largamente utilizada em pesquisas para medir o nível de interação social entre participantes em redes formadas por atividades como discussões online (LIU. et al., 2018; GAŠEVIĆ et al., 2018; OLIVARES et al., 2019). A SNA permite identificar quais participantes são mais influentes e/ou intermediários dentro de uma rede. Portanto, medidas amplamente aceitas de SNA são utilizadas aqui: i) *Proximidade* avalia quão próximo um estudante na rede está de todos os outros estudantes. Conforme (YUSOF; RAHMAN et al., 2009), estudantes com um maior grau de proximidade tendem a ser eficazes na disseminação de informações pela rede; ii) *Centralidade de intermediação* investiga a importância da intervenção de um estudante nas interações entre outros estudantes. Assim, um alto nível de intermediação indica a liderança do estudante entre seus pares; e (iii) *Grau de centralidade* busca analisar a interação do estudante com outros estudantes presentes na rede. Portanto, um alto grau de centralidade do estudante significa que o estudante é mais influente na rede.

4.5.1.2 Detalhes do Experimento

Inicialmente, o conjunto de dados foi dividido em conjuntos de treinamento e teste, de forma semelhante aos experimentos anteriores (FERREIRA et al., 2020; FARROW; MOORE; GAŠEVIĆ, 2019). As mensagens dos fóruns de discussão foram processadas para extrair características textuais relevantes, como as características mencionadas anteriormente (LIWC, Coh-Metrix, SNA, etc).

Neste experimento, utilizamos a ferramenta AutoGluon[6] para selecionar o melhor modelo de classificação para as categorias da presença cognitiva. AutoGluon é uma biblioteca de AutoML (Aprendizagem de Máquina Automático) que facilita a construção e a otimização de modelos de aprendizagem de máquina de alto desempenho com intervenção mínima do usuário (ERICKSON et al., 2020). Esse tipo de ferramenta é projetada para realizar automaticamente a seleção de modelos, ajuste de hiperparâmetros, e empilhamento de modelos, visando maximizar a acurácia preditiva. A AutoGluon foi escolhida especificamente pela flexibilidade de modelos

[6] https://auto.gluon.ai/stable/index.html

e estratégias de treinamento de modelo, além de ter apresentado resultados significativamente melhores que ferramentas similares (ERICKSON et al., 2020).

Com os dados preparados, alimentamos o AutoGluon com as características extraídas. A AutoGluon então realiza uma busca extensiva de modelos, utilizando os hiperparâmetros padrão de cada modelo/biblioteca e avaliando múltiplos algoritmos de classificação, como: Random Forest, Gradient Boosting Machines, e redes neurais profundas.

Durante o treinamento, o AutoGluon monitorou o desempenho dos modelos em termos de acurácia e utilizou técnicas de empilhamento para combinar os melhores modelos individuais em uma metapredição, que busca otimizar a performance preditiva (ERICKSON et al., 2020). O modelo final selecionado foi então avaliado no conjunto de teste para medir sua eficácia na classificação das categorias da presença cognitiva.

4.5.2 Experimento 2: Classificação com Modelos Baseados em Árvore e Aprendizado Ativo

4.5.2.1 Características Extraídas

Neste experimento, utilizamos as mesmas características extraídas do Experimento 1 (Subseção 4.5.1) para realizar a classificação das categorias da presença cognitiva. As características foram obtidas a partir das ferramentas LIWC, Coh-Metrix e Análise de Redes Sociais (SNA), conforme descrito anteriormente.

Essas características fornecem uma visão abrangente das interações textuais e sociais nos fóruns de discussão, permitindo a construção de modelos de classificação robustos. O LIWC analisa aspectos psicolinguísticos, o Coh-Metrix avalia a coesão e coerência textual, e a SNA mede a estrutura e dinâmica das interações entre os participantes.

4.5.2.2 Detalhes do Experimento

A Figura 2 no Capítulo 2 apresenta o processo de treinamento de um classificador incluindo a abordagem de aprendizado ativo. Inicialmente, o modelo de ML é construído usando apenas uma pequena amostra de dados anotados (entre 10% e 20%), em seguida, uma instância por vez é consultada a partir dos dados não anotados de acordo com a medida de incerteza para ser anotada por um especialista (oráculo); a nova instância anotada é inserida no modelo de

ML a fim de atualizá-lo. Essa etapa de calibração do modelo pode ser executada até que o classificador alcance resultados satisfatórios.

Na primeira etapa, o conjunto de dados foi dividido em conjuntos de treinamento e teste (escolhidos aleatoriamente), para ambas as tarefas de classificação (ou seja, presenças social e cognitiva) seguindo o trabalho de (FERREIRA et al., 2020) e (FARROW; MOORE; GAŠEVIĆ, 2019). O algoritmo *Random Forest*, que é amplamente aplicado na literatura para identificar presenças sociais e cognitivas (KOVANOVIć et al., 2016; FARROW; MOORE; GAŠEVIĆ, 2019; FERREIRA et al., 2020), além disso mostrou boa performance no trabalho de (ROLIM et al., 2021), foi usado aqui para construir o classificador. O outro modelo utilizado foi o XGBoost (CHEN et al., 2015) que é também é um modelo baseado em árvore e possui boa performance para diversos conjuntos de dados nos mais variados domínios.

Inicialmente, 15% do conjunto de treinamento (conjunto de dados descrito na Seção 4.1) foi utilizado como amostra inicial de dados (requerida pelo processo de aprendizado ativo). Como o conjunto de dados usado já está anotado, a abordagem proposta considerou os dados anotados como uma entrada fornecida por um oráculo. A cada execução do pipeline de treinamento novas amostras do conjunto de dados eram escolhidas com base na medida de incerteza, até atingir valores mais altos nas métricas de avaliação.

4.5.3 Experimento 3: Classificação com Modelos Baseados em Árvore e Dados Aumentados

4.5.3.1 Características Extraídas

Assim como os experimentos anteriores (Subseções 4.5.1 e 4.5.2), foram extraídas as mesmas características textuais, a fim de manter a padronização da experimentações e obter resultados comparáveis.

4.5.3.2 Detalhes do Experimento

Para abordar o problema de desbalanceamento nas categorias da presença cognitiva no conjunto de dados, empregamos técnicas de aumentação de dados. Nossa base de dados apresenta um desbalanceamento significativo, com as categorias "Evento Desencadeador" e "Resolução" possuindo menos instâncias comparadas às outras categorias. A categoria domi-

nante, "Exploração", possui 608 instâncias no conjunto de treinamento.

Utilizamos a biblioteca `nlpaug`[7] para realizar a aumentação de dados, especificamente a função de aumentação por palavras contextuais de modelos de linguagem, empregando o modelo BERT. Para cada categoria da presença cognitiva, geramos instâncias adicionais suficientes para igualar a quantidade da categoria dominante, "Exploração". A aumentação foi realizada sem alterar as stopwords e com um máximo de 100 palavras (aleatórias) por instância. Na Tabela 6, podemos observar a quantidade de instâncias que tiveram que ser aumentadas para cada categoria. As instâncias de referência de cada categoria são selecionadas aleatoriamente.

Tabela 6 – Número de Instâncias Aumentadas para Cada Categoria da Presença Cognitiva

Categoria	**Número de Instâncias Aumentadas**
Outros	496
Evento Desencadeador	328
Exploração	0
Integração	183
Resolução	563

Nota: A categoria "Exploração" é a categoria predominante com 608 instâncias no conjunto de treinamento original (após a separação de treinamento e teste), portanto não necessitou de aumentação de dados.

Com o conjunto de dados balanceado através da aumentação de dados, treinamos o modelo de Random Forest e o XGBoost para a classificação das categorias da presença cognitiva. Utilizamos a abordagem de aprendizado ativo, onde o modelo é inicialmente treinado com uma pequena amostra de dados anotados e iterativamente refinado através da adição de novas instâncias anotadas com base na incerteza das previsões.

A avaliação do modelo foi conduzida utilizando as métricas de acurácia e o coeficiente de Cohen's Kappa, permitindo uma comparação direta com os resultados obtidos em experimentos anteriores. Esta abordagem visa melhorar a robustez e a precisão das classificações, garantindo que todas as categorias da presença cognitiva sejam representadas de forma equitativa no modelo final.

[7] https://nlpaug.readthedocs.io/en/latest/index.html

4.5.4 Experimento 4: Classificação com DeBERTa

4.5.4.1 Características Extraídas

Para o experimento de classificação utilizando o DeBERTa, as características textuais foram exclusivamente baseadas nos *embeddings* gerados pelo modelo *DeBERTa*(V3 Base)[8]. Esses *embeddings* fornecem representações contextuais profundas de cada mensagem no fórum de discussão, capturando tanto o significado das palavras quanto as relações semânticas entre elas em diferentes contextos. A literatura recente sugere que, ao utilizar *embeddings* gerados por modelos baseados na arquitetura Transformer, como o DeBERTa, não é necessário incorporar outras características textuais tradicionais, como frequências de palavras ou métricas de coesão, uma vez que os *embeddings* encapsulam de forma abrangente a informação relevante para a tarefa de classificação (HE; GAO; CHEN, 2021). Esta abordagem simplifica o processo de extração de características, concentrando-se na poderosa capacidade de representação do DeBERTa, que já provou ser eficaz em uma variedade de tarefas de processamento de linguagem natural (WOLF et al., 2020).

Os *embeddings* finais são gerados pela última camada do DeBERTa após passar por todas as camadas de atenção. Cada camada do DeBERTa refina a representação do texto, considerando as interdependências contextuais de todas as palavras na sequência. A saída dessa última camada, que contém os *embeddings* mais ricos em contexto e semântica, é então passada através de uma camada linear final que infere as características necessárias para a classificação. Esta última camada linear atua como um classificador que utiliza as informações codificadas nos *embeddings* para distinguir entre as diferentes categorias da presença cognitiva.

Com essa representação gerada, é possível realizar o treinamento de um modelo de Perceptron Multicamadas (MLP). Os *embeddings* extraídos pelo DeBERTa são utilizados como entradas para o MLP, que então aprende a mapear essas representações contextuais para as respectivas categorias da presença cognitiva. O MLP, com sua capacidade de capturar não-linearidades complexas, pode aproveitar a riqueza dos *embeddings* do DeBERTa.

[8] https://huggingface.co/microsoft/deberta-v3-base

4.5.4.2 Detalhes do Experimento

Neste experimento, utilizamos o mesmo conjunto de dados descrito na Seção 4.1, que foi dividido em conjuntos de treinamento e teste, conforme descrito na seção anterior e seguindo os trabalhos de referência na literatura (FERREIRA et al., 2020; FARROW; MOORE; GAŠEVIĆ, 2019). Treinamos um Perceptron Multicamadas (MLP) especificamente para a tarefa de classificação das categorias da presença cognitiva, utilizando os *embeddings* gerados pelo modelo DeBERTa. O MLP foi configurado com uma camada de entrada correspondente ao tamanho dos *embeddings* do DeBERTa, seguida por uma ou mais camadas ocultas com unidades neuronais totalmente conectadas, e uma camada de saída ajustada para refletir as diferentes categorias da presença cognitiva.

4.5.5 Experimento 5: Classificação com DeBERTa com Dados Aumentados pelo GPT-4

4.5.5.1 Características Extraídas

Assim como no Experimento 4.5.4, empregamos a utilização do modelo de linguagem, DeBERTa, para realizar a representação textual do conjunto de dados e assim extrair os *embeddings* para poder utilizarmos na classificação.

4.5.5.2 Detalhes do Experimento

Neste experimento, utilizamos a técnica de aumento de dados com o modelo GPT-4 para gerar novas instâncias e balancear as categorias da presença cognitiva no conjunto de dados. Optamos pelo GPT-4 devido às suas habilidades avançadas de geração de linguagem natural, permitindo-nos expandir o conjunto de dados com instâncias mais diversas e contextualmente ricas, o que é fundamental para melhorar o desempenho dos modelos de classificação. Em particular, o modelo DeBERTa, que utilizamos para a etapa de classificação, requer dados diversificados para alcançar um desempenho otimizado, beneficiando-se significativamente de um conjunto de treinamento amplo e variado, diferentemente do Experimento 3, em que o BERT foi utilizado para aumento dos dados por meio da substituição de palavras por sinônimos. Foram utilizados três tipos diferentes de prompts para gerar essas novas instâncias:

- ***Zero-shot* sem contexto:** Este prompt solicitou ao GPT-4 a geração de postagens de fórum sem fornecer contexto adicional sobre o CoI, apenas especificando a categoria da presença cognitiva a ser representada.
- ***Zero-shot* com contexto:** Este prompt incluiu uma breve explicação sobre o modelo de CoI e as categorias de presença cognitiva, antes de solicitar a geração das postagens.
- ***Few-shot* com contexto:** Além de fornecer o contexto sobre o CoI e suas categorias, este prompt incluiu exemplos específicos de postagens para guiar a geração de novas instâncias pelo GPT-4.

Para a geração de novas instâncias utilizando o modelo GPT-4, empregamos três tipos diferentes de prompts, descritos a seguir:

- **Zero-shot sem contexto:**

  System: You are a student of a software engineering post-graduation. CONTEXT: - Take into consideration the forum discussion analysis and Community of Inquiry (CoI) framework and cognitive presence categories.

  User: Write 1 post of a discussion forum for a master level research-intensive course in software engineering offered entirely online. All posts must represent the 'category' category of the cognitive presence from CoI.

- **Zero-shot com contexto:**

  System: You are a student of a software engineering post-graduation. CONTEXT: - Take into consideration the forum discussion analysis and Community of Inquiry (CoI) framework and cognitive presence categories. - As you already know, Community of Inquiry (CoI) is a well-known framework that aims to outline how asynchronous online communication shapes student learning and their cognitive development. - The CoI model defines three dimensions that mold the learning experience (i.e., social presence, cognitive presence, and teaching presence). - Cognitive presence is highly related to the concept of knowledge construction and problem-solving. - Cognitive presence has 4 different categories. (1) Triggering event: A problem or dilemma is identified and conceptualized. In an educational context, discussions are usually triggered

by instructors; however, they can also be initiated by any participant in the discussion. (2) Exploration: The students explore the potential solutions to a given problem, typically by information seeking and brainstorming different ideas. (3) Integration: The students synthesize new ideas and knowledge by employing social (co-)construction. (4) Resolution: Finally, students solve the original dilemma or problem triggered at the beginning of the learning cycle. Here, students evaluate the newly created knowledge through hypothesis testing, vicarious application, or consensus building.

User: Write 1 post of a discussion forum for a master level research-intensive course in software engineering offered entirely online. All posts must represent the 'category' category of the cognitive presence from CoI.

- **Few-shot com contexto:**

  System: You are a student of a software engineering post-graduation. CONTEXT: - Take into consideration the forum discussion analysis and Community of Inquiry (CoI) framework and cognitive presence categories. - As you already know, Community of Inquiry (CoI) is a well-known framework that aims to outline how asynchronous online communication shapes student learning and their cognitive development. - The CoI model defines three dimensions that mold the learning experience (i.e., social presence, cognitive presence, and teaching presence). - Cognitive presence is highly related to the concept of knowledge construction and problem-solving. - Cognitive presence has 4 different categories. (1) Triggering event: A problem or dilemma is identified and conceptualized. In an educational context, discussions are usually triggered by instructors; however, they can also be initiated by any participant in the discussion. (2) Exploration: The students explore the potential solutions to a given problem, typically by information seeking and brainstorming different ideas. (3) Integration: The students synthesize new ideas and knowledge by employing social (co-)construction. (4) Resolution: Finally, students solve the original dilemma or problem triggered at the beginning of the learning cycle. Here, students evaluate the newly created knowledge through hypothesis testing, vicarious application, or consensus building.

User: Pretend to be a student and write 1 post simulating a discussion forum for a master level research-intensive course in software engineering offered entirely online. All posts must represent the 'category' category of the cognitive presence from CoI.

Assistant: ...

User: Pretend to be a student and write 1 post simulating a discussion forum for a master level research-intensive course in software engineering offered entirely online. All posts must represent the 'category' category of the cognitive presence from CoI.

Assistant: ...

Tabela 7 – Quantidade de Dados Gerados para Cada Prompt

Categoria	**Prompt 1**	**Prompt 2**	**Prompt 3**
Evento Desencadeador	750	552	626
Exploração	750	550	603
Integração	750	550	597
Resolução	742	550	617

A Tabela 7 apresenta a quantidade de dados gerados para cada categoria de presença cognitiva utilizando três diferentes tipos de prompts. No Prompt 1, foram geradas aproximadamente 750 instâncias para cada categoria. O Prompt 2 gerou cerca de 550 instâncias por categoria, enquanto o Prompt 3 produziu um número variável de instâncias, com a categoria "Evento Desencadeador" recebendo 626 instâncias e "Exploração" recebendo 603 instâncias.

Treinamos um MLP para a tarefa de classificação, utilizando os *embeddings* gerados pelo DeBERTa. O MLP foi configurado com uma camada de entrada correspondente ao tamanho dos *embeddings*, seguido por varias camadas ocultas com unidades neuronais totalmente conectadas, e uma camada de saída ajustada para refletir as diferentes categorias da presença cognitiva. Um modelo foi treinado para cada estratégia de aumentação.

4.6 RESUMO DO CAPÍTULO

Neste capítulo, detalhamos a metodologia empregada para investigar a viabilidade e a eficácia de técnicas automáticas de processamento de linguagem natural e análise de redes na classificação da presença cognitiva em ambientes virtuais de aprendizagem. Utilizamos

dados coletados de fóruns de discussão de um curso online de mestrado em engenharia de software, extraindo características textuais por meio de ferramentas como LIWC, Coh-Metrix e modelos de linguagem baseados em transformadores, como BERT, GPT-4 e DeBERTa. Diversos algoritmos de classificação foram aplicados, incluindo Random Forest, XGBoost e redes neurais MLP, avaliados por métricas como acurácia e o coeficiente Kappa de Cohen. Os experimentos foram planejados progressivamente para aprimorar o desempenho dos modelos, abordando desafios como o desequilíbrio de classes e a escassez de dados, culminando no uso de técnicas avançadas de aumento de dados e representações textuais enriquecidas. No próximo capítulo, discutiremos os resultados obtidos a partir desses experimentos.

5 RESULTADOS

Neste capítulo, apresentamos os resultados obtidos a partir dos experimentos conduzidos para a classificação das categorias da presença cognitiva em fóruns de discussão online. Os experimentos foram realizados utilizando diferentes abordagens e modelos de aprendizado de máquina, incluindo técnicas de aprendizado ativo, aumentação de dados e o uso de modelos baseados na arquitetura Transformer, como o DeBERTa. Avaliamos o desempenho dos modelos utilizando métricas de acurácia e o coeficiente de Cohen's Kappa para medir a concordância entre as classificações previstas e as categorias reais. A seguir, detalhamos os resultados obtidos em cada experimento e discutimos as implicações e observações importantes derivadas dos dados.

5.1 RESULTADOS DO EXPERIMENTO 1

A Tabela 8 apresenta os resultados do Experimento 1, no qual utilizamos a ferramenta AutoGluon para executar diversos treinamentos de modelos em dados tabulares, com o objetivo de identificar os melhores modelos para o conjunto de dados. Os modelos foram avaliados com base na acurácia, com o XGBoost alcançando a maior acurácia de 0.56, seguido pelo RandomForest com 0.55 e o MLP com 0.55. Modelos como CatBoost, LightGBM e ExtraTrees apresentaram acurácias ligeiramente menores, enquanto o KNeighbors teve desempenho inferior, com acurácias de 0.34, respectivamente. Estes resultados destacam a eficácia relativa dos diferentes modelos de aprendizado de máquina para a tarefa de classificação das categorias de presença cognitiva em fóruns de discussão online. É importante mencionar que a utilização da acurácia para avaliação deste experimento está de acordo com outros trabalhos que utilizaram o mesmo conjunto de dados, e essa faixa de valores atingidos também está alinhada com a literatura, como pode ser observado no Capítulo 3.

Embora os resultados obtidos com os três modelos de melhor desempenho - XGBoost, RandomForest e MLP - não tenham alcançado os níveis do estado da arte para este problema, eles mostram um caminho promissor a ser seguido. O XGBoost apresentou resultados razoáveis, porém sua eficácia é limitada pelo pequeno tamanho e pelo desbalanceamento do conjunto de dados. O RandomForest também mostrou um desempenho moderado, mas enfrenta desafios similares devido à quantidade limitada de dados anotados. O MLP, com sua arquitetura baseada

#	Modelo	Acurácia
1	XGBoost	0.560641
2	RandomForest	0.558352
3	MLP	0.551487
4	CatBoost	0.535469
5	LightGBM	0.530892
6	ExtraTrees	0.519451
7	KNeighbors	0.338673

Tabela 8 – Modelos e suas respectivas acurácias

em redes neurais, conseguiu capturar algumas complexidades não-lineares nos dados, mas ainda assim, os resultados indicam que há espaço significativo para melhorias.

Com base nesses resultados, todos os outros experimentos desta tese focarão na utilização desses três tipos de modelos. Implementaremos técnicas de aprendizado ativo com XGBoost e RandomForest para explorar e maximizar o potencial dessas abordagens em conjuntos de dados anotados de forma limitada. Além disso, treinaremos um MLP utilizando o DeBERTa como modelo de linguagem para a extração de *embeddings*. A escolha do DeBERTa se deve à sua capacidade de fornecer representações contextuais profundas, o que, combinado com o MLP, tem o potencial de melhorar significativamente a acurácia das classificações. Esta estratégia integrada visa aproveitar ao máximo as vantagens de cada modelo, garantindo uma análise robusta e precisa das presenças cognitivas em fóruns de discussão online, mesmo diante dos desafios impostos pelos dados desbalanceados e de tamanho reduzido.

5.2 RESULTADOS DO EXPERIMENTO 2

Como mencionado no Capítulo 4 os experimentos com base no aprendizado ativo inicial o seu treinamento com 15% do conjunto de dados (Seção 4.1). Logo, a comparação a seguir apresenta o resultado do treinamento inicial do modelo sem levar em consideração novas consultas na base de dados, a fim de encontrar o melhor modelo para o problema.

A Figura 8 apresenta uma comparação entre os modelos XGBoost e Random Forest utilizando aprendizado ativo. Analisando as métricas de acurácia e Cohen's Kappa para ambos os modelos, podemos observar que:

Acurácia: O XGBoost apresenta uma acurácia ligeiramente inferior, em torno de 0.55, enquanto o Random Forest apresenta uma acurácia muito próxima, também em torno de 0.58.

Esta proximidade indica que ambos os modelos são eficazes na classificação das categorias de presença cognitiva, embora a diferença seja mínima. **Cohen's Kappa**: O coeficiente de Cohen's Kappa para ambos os modelos está em torno de 0.40 (XGBoost = 0.38 e Random Forest = 0.41). Este coeficiente, que ajusta a acurácia pela concordância esperada ao acaso, mostra que ambos os modelos ainda enfrentam desafios significativos na classificação precisa das diferentes categorias de presença cognitiva. Para o conjunto de dados, essa faixa de valores atingidos pelo Cohen's Kappa está alinhada com a literatura (ver o Capítulo 3).

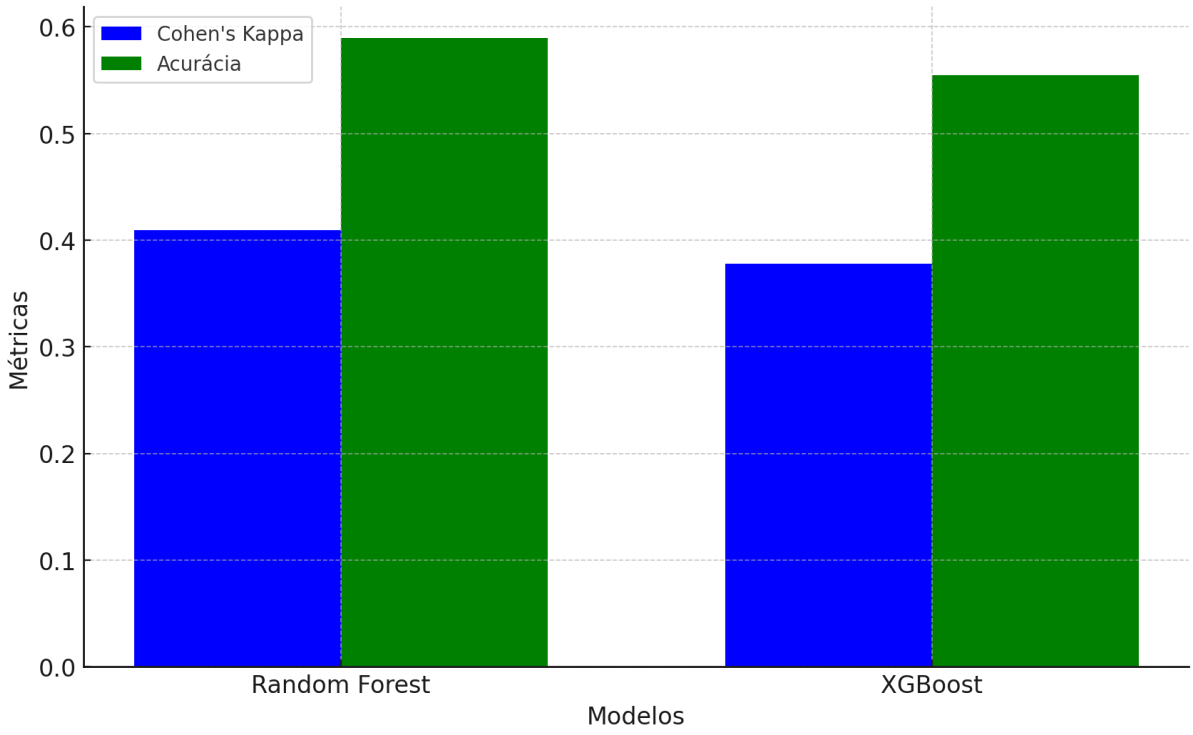

Figura 8 – Comparação do Random Forest e XGBoost na abordagem de aprendizagem ativa.

A Figura 9 mostra a evolução do desempenho do modelo Random Forest ao longo das iterações de aprendizado ativo, destacando as métricas de acurácia (linha laranja) e Cohen's Kappa (linha amarela). A cada nova consulta na base de dados o modelo é avaliado novamente. Foram executadas 50 consultas nesse experimento, o que acabou representando 18% do total do conjunto de dados. Analisando essa figura, podemos observar que:

Acurácia: A acurácia do Random Forest começa em torno de 0.58 e sobe lentamente até 0.60 ao longo das iterações. Este aumento gradual indica que o modelo se beneficia do aprendizado ativo, mas a melhoria é limitada pelo tamanho e desbalanceamento do conjunto de dados. **Cohen's Kappa**: O coeficiente de Cohen's Kappa mostra uma tendência de melhoria mais gradual, começando em torno de 0.41 e subindo lentamente até um aproximadamente

0.43. Esta melhoria sugere que, embora o aprendizado ativo ajude a melhorar a concordância ajustada, ainda existem desafios na classificação precisa das categorias.

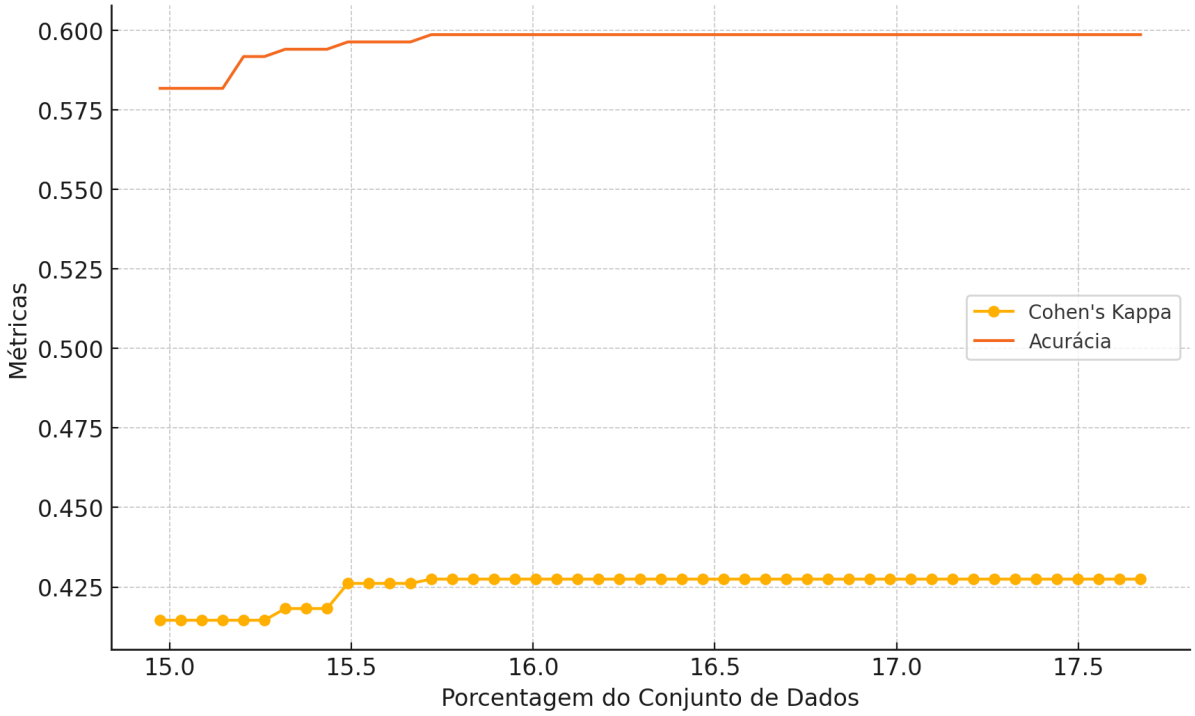

Figura 9 – Análise da performance do Random Forest após novas consultas no conjunto de dados.

Esses resultados indicam que estratégias adicionais, como aumentação de dados e refinamento de modelos, possuem o potencial de melhorar os resultados para alcançar um desempenho mais próximo do estado da arte. Embora já seja um resultado superior ao de (FARROW; MOORE; GAŠEVIĆ, 2019) com a mesma base de dados, contudo utilizando apenas 18% dos dados. A combinação de aprendizado ativo com outras técnicas avançadas de processamento de linguagem natural e aprendizado de máquina pode ser uma direção promissora para melhorar a classificação das presenças cognitivas em fóruns de discussão online.

5.3 RESULTADOS DO EXPERIMENTO 3

A Figura 10 apresenta uma comparação entre os modelos XGBoost e Random Forest utilizando aprendizado ativo com aumentação de dados gerados pelo BERT. Analisando as métricas de acurácia e Cohen's Kappa para ambos os modelos, podemos observar que:

Acurácia: Tanto o XGBoost quanto o Random Forest apresentam uma acurácia próxima a 0.60, com uma ligeira vantagem para o XGBoost. Esta proximidade indica que ambos os

modelos performance semelhante para a classificação das categorias de presença cognitiva, com diferença mínima. **Cohen's Kappa**: O coeficiente de Cohen's Kappa para ambos os modelos está em torno de 0.40 (Random Forest = 0.36 e XGBoost = 0.41), o que é consistente com os resultados sem aumentação de dados.

Comparando com os resultados sem aumentação de dados, a acurácia e o Cohen's Kappa são similares, com ligeira piora do Random Forest e melhora do XGBoost em ambas as métricas.

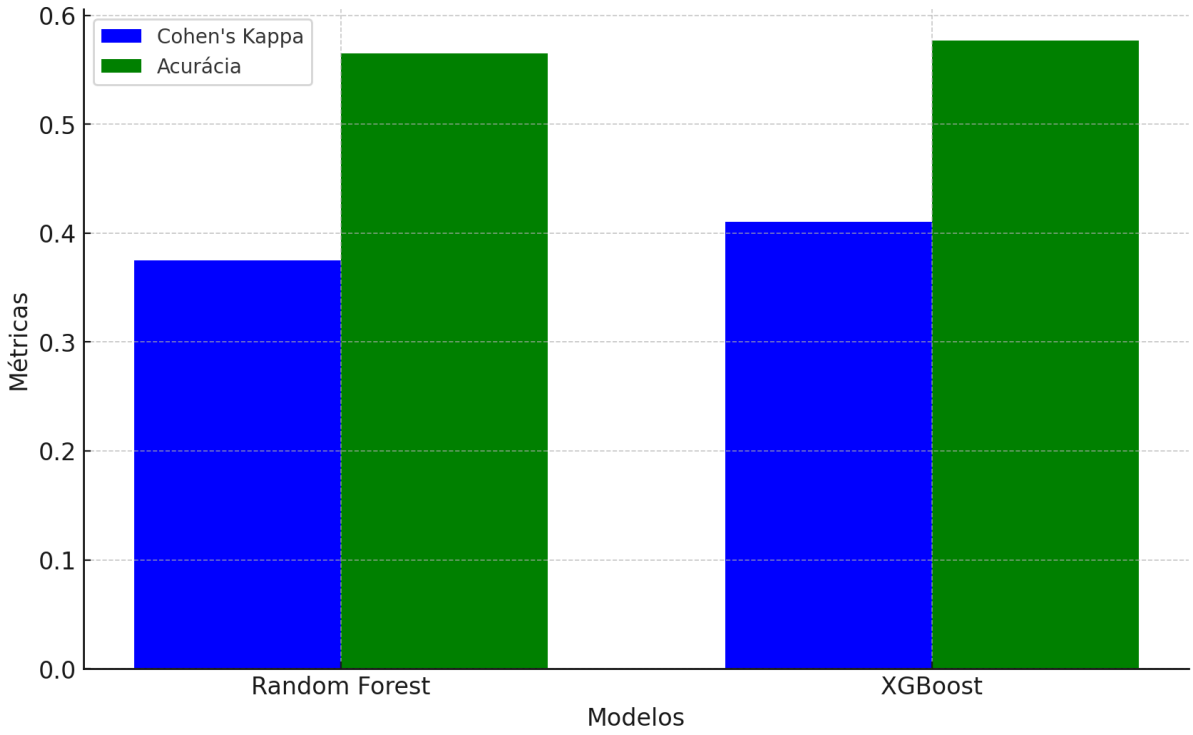

Figura 10 – Comparação do Random Forest e XGBoost na abordagem de aprendizagem ativa com dados aumentados.

A Figura 11 mostra a evolução do desempenho do modelo XGBoost ao longo das iterações de aprendizado ativo com aumentação de dados, destacando as métricas de acurácia (linha laranja) e Cohen's Kappa (linha amarela). A cada nova consulta na base de dados o modelo é avaliado novamente. Foram executadas 50 consultas nesse experimento, o que acabou representando 18% do total do conjunto de dados. Analisando essa figura, podemos observar que:

Acurácia: A acurácia do XGBoost permanece estável ao longo das iterações em torno de 0.58. Isso indica que a aumentação de dados com BERT não trouxe melhorias significativas na acurácia do modelo. **Cohen's Kappa**: O coeficiente de Cohen's Kappa mostra uma tendência estável, começando em torno de 0.41 e permanecendo próximo a esse valor ao longo das

iterações. Isso sugere que, embora a aumentação de dados tenha gerado mais instâncias e aumentado a diversidade do conjunto de dados e o conjunto inicial de treinamento ser maior (15% = 361 instâncias), ela não melhorou substancialmente a concordância ajustada do modelo.

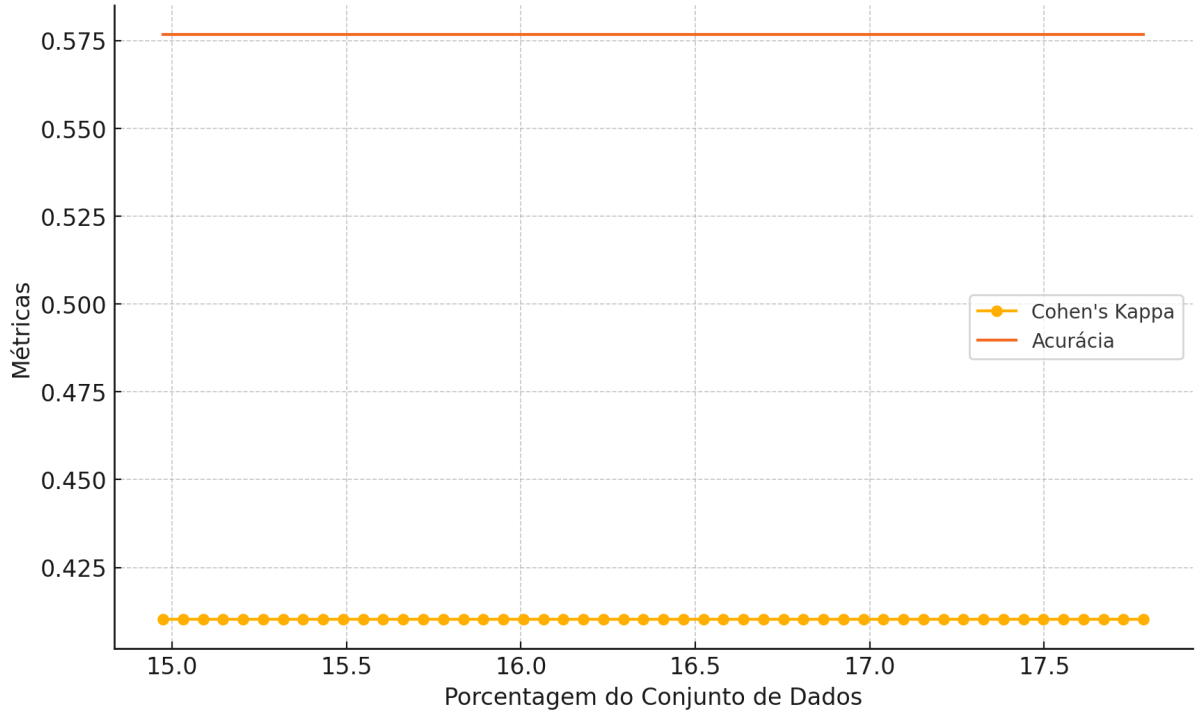

Figura 11 – Análise da performance do XGBoost após novas consultas no conjunto de dados aumentados.

A análise conjunta das Figuras 10 e 11 revela algumas conclusões importantes. Primeiramente, os resultados do aprendizado ativo com aumentação de dados utilizando o BERT não diferem significativamente dos resultados sem aumentação de dados. Tanto a acurácia quanto o coeficiente de Cohen's Kappa permaneceram em níveis semelhantes, indicando que a aumentação de dados não proporcionou uma melhoria notável na classificação das presenças cognitivas.

Esses resultados sugerem que, embora a aumentação de dados seja uma técnica promissora, a qualidade e a relevância dos dados aumentados são cruciais. Estratégias adicionais, como a utilização de técnicas mais avançadas de aumentação de dados ou a integração de outras fontes de dados, podem ser necessárias para alcançar melhorias significativas. Para tanto, nos próximos experimentos iremos utilizar técnicas mais avançadas de processamento de linguagem natural para tentar alcançar melhores resultados.

5.4 RESULTADOS DO EXPERIMENTO 4

Neste experimento, utilizamos o modelo DeBERTa v3-base da Microsoft para realizar a classificação das categorias da presença cognitiva. A arquitetura DeBERTa (Decoding-enhanced BERT with Disentangled Attention) é conhecida por sua capacidade de capturar representações contextuais profundas e precisas, o que a torna ideal para tarefas de classificação de texto.

O modelo utilizado foi o `AutoModelForSequenceClassification` da biblioteca Hugging Face Transformers, configurado com a seguinte arquitetura e parâmetros:

- **Modelo**: `microsoft/deberta-v3-base`
- **Número de Labels**: 4 (correspondente às categorias da presença cognitiva)
- **Tamanho do Embedding**: 768
- **Número de Camadas Escondidas**: 12
- **Número de Cabeças de Atenção**: 12
- **Tamanho do Feedforward**: 3072
- **Função de Ativação**: GELU
- **Dropout na Atenção**: 0.1
- **Dropout nas Camadas Escondidas**: 0.1
- **Learning Rate**: 0.00002
- **Weight Decay**: 0.01
- **Batch size**: 8
- **Passos de Acumulação de Gradiente**: 4
- **Número de Épocas de Treinamento**: 3
- **Scheduler de Learning Rate**: Linear
- **Precisão Mista (FP16)**: Ativado

O modelo DeBERTa foi treinado utilizando os embeddings gerados pelas camadas finais do modelo para capturar as representações contextuais das mensagens nos fóruns de discussão. Esses embeddings foram então utilizados para classificar as categorias da presença cognitiva, usando uma camada linear (MLP) final que infere as classes a partir dessas representações.

Os dados de entrada foram processados e divididos em conjuntos de treinamento e teste (75% e 25% respectivamente), estratificados. A configuração do treinamento incluiu técnicas como acumulação de gradientes e precisão mista para otimizar o uso de memória e acelerar o treinamento. O aprendizado foi monitorado e avaliado em intervalos regulares para garantir a convergência do modelo.

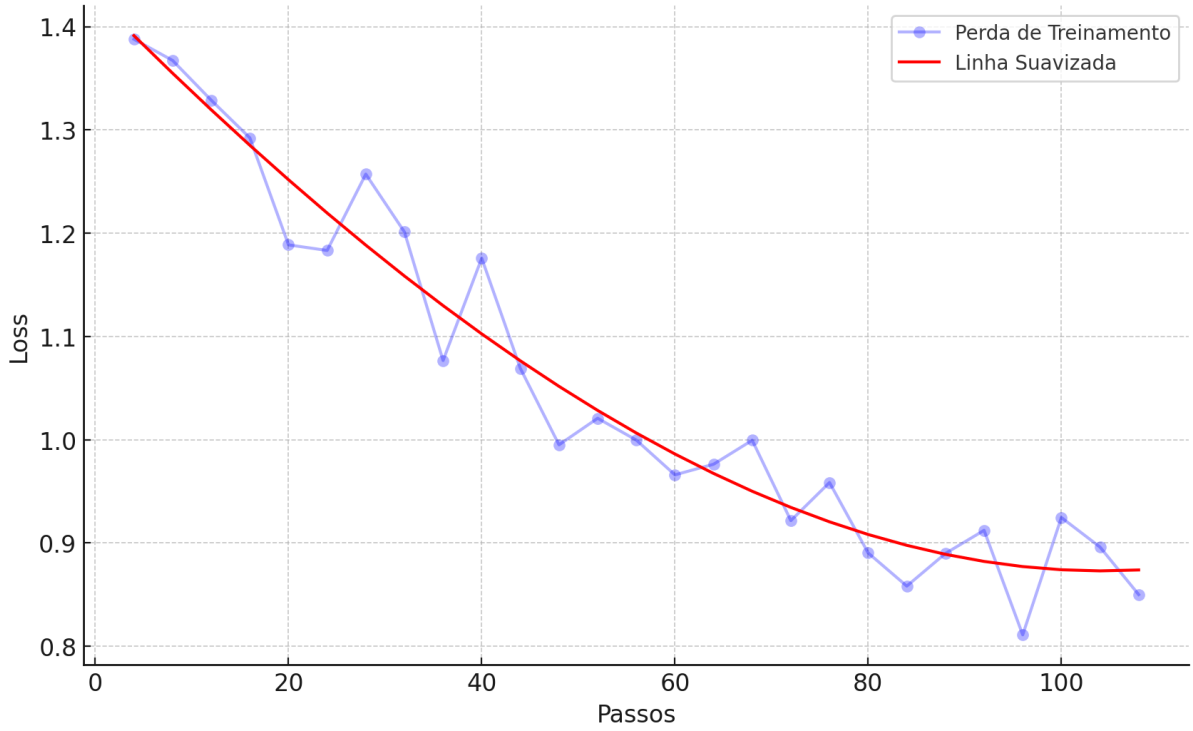

Figura 12 – Curva Loss do treinamento do modelo.

A Figura 12 apresenta a perda de treinamento ao longo dos passos no Experimento 4, onde o modelo DeBERTa foi utilizado. Observa-se uma redução consistente na perda de treinamento, indicando que o modelo estava aprendendo de forma durante o processo de treinamento. A perda final de treinamento foi de 1.048, demonstrando uma boa convergência do modelo após 111 passos de treinamento. Ao final do treinamento, o modelo atingiu o valor de 0.58 e 0.38 de acurácia e Cohen's Kappa respectivamente.

5.5 RESULTADOS DO EXPERIMENTO 5

Neste experimento, utilizamos o modelo DeBERTa v3-base da Microsoft, assim como no Experimento 4, para realizar a classificação das categorias da presença cognitiva. A diferença principal deste experimento é a utilização de dados aumentados pelo GPT-4 para melhorar o desempenho do modelo.

Para este experimento, os dados foram aumentados utilizando o GPT-4, que gerou dados adicionais baseados nos prompts fornecidos, conforme descrito anteriormente (Seção 4.5.5. Esses dados aumentados foram incorporados ao conjunto de dados original, visando lidar com o desbalanceamento e aumentar a quantidade de instâncias disponíveis para treinamento.

O modelo DeBERTa foi então treinado utilizando este conjunto de dados aumentado. Os embeddings gerados pelas camadas finais do modelo foram usados para capturar as representações contextuais das mensagens nos fóruns de discussão, da mesma forma que no Experimento 4. A camada linear (MLP) final foi utilizada para classificar as categorias da presença cognitiva a partir dessas representações.

A Figura 13 apresenta a perda de treinamento ao longo dos passos de treinamento. A linha azul mostra a perda de treinamento em cada passo, enquanto a linha vermelha representa uma suavização dos dados, proporcionando uma visão mais clara da tendência de convergência. Observa-se uma redução consistente na perda de treinamento, indicando que o modelo estava aprendendo de forma eficaz durante o processo de treinamento. A perda final de treinamento foi de 0.6229, demonstrando uma boa convergência do modelo após 104 passos de treinamento. Ao final do treinamento, o modelo atingiu o valor de 0.59 e 0.37 de acurácia e Cohen's Kappa respectivamente.

A Figura 14 apresenta a perda de treinamento ao longo dos passos de treinamento. Observa-se uma redução consistente na perda de treinamento, indicando que o modelo estava aprendendo de forma eficaz durante o processo de treinamento. Ao final do treinamento, o modelo atingiu o valor de 0.59 e 0.40 de acurácia e Cohen's Kappa respectivamente.

A Figura 15 apresenta a perda de treinamento ao longo dos passos de treinamento. Observa-se uma redução consistente na perda de treinamento, indicando que o modelo estava aprendendo de forma eficaz durante o processo de treinamento. Ao final do treinamento, o modelo atingiu o valor de 0.59 e 0.39 de acurácia e Cohen's Kappa respectivamente.

Como o modelo que foi treinado com o conjunto de dados aumentado pelo prompt 2 obteve o melhor valor de Kappa, decidimos aumentar a quantidade de épocas para verificar se a ten-

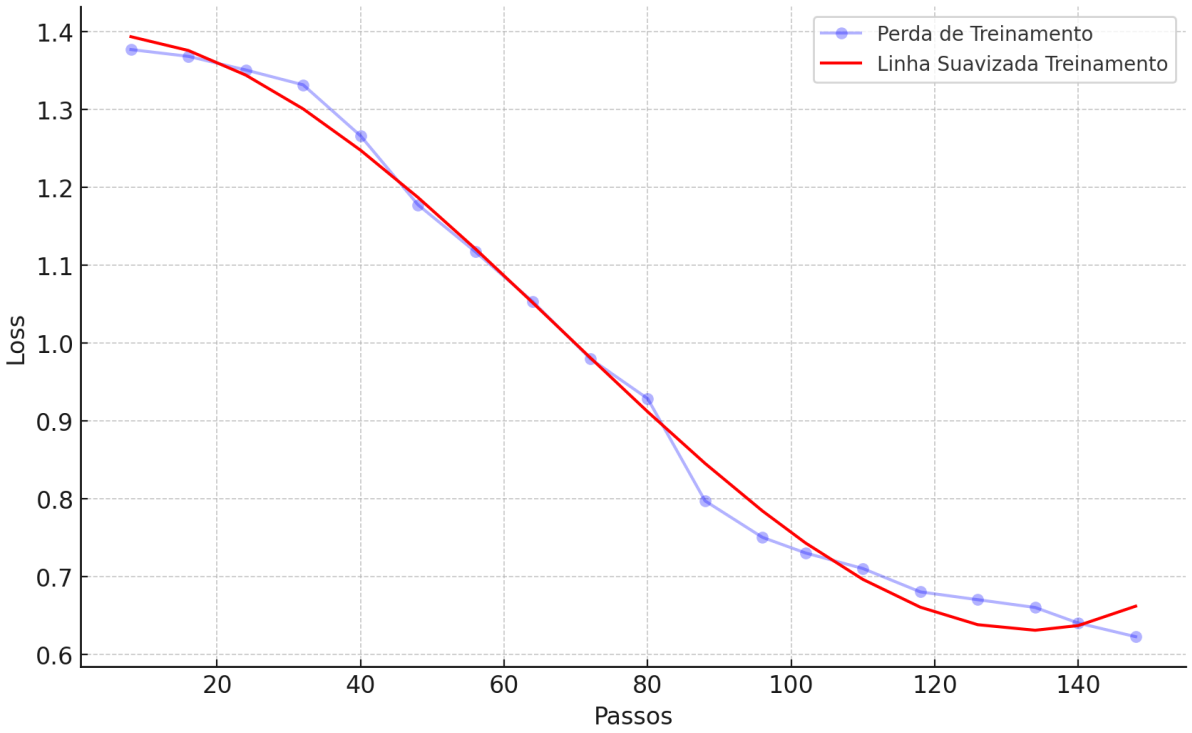

Figura 13 – Curva Loss do treinamento do modelo com dados do prompt 1.

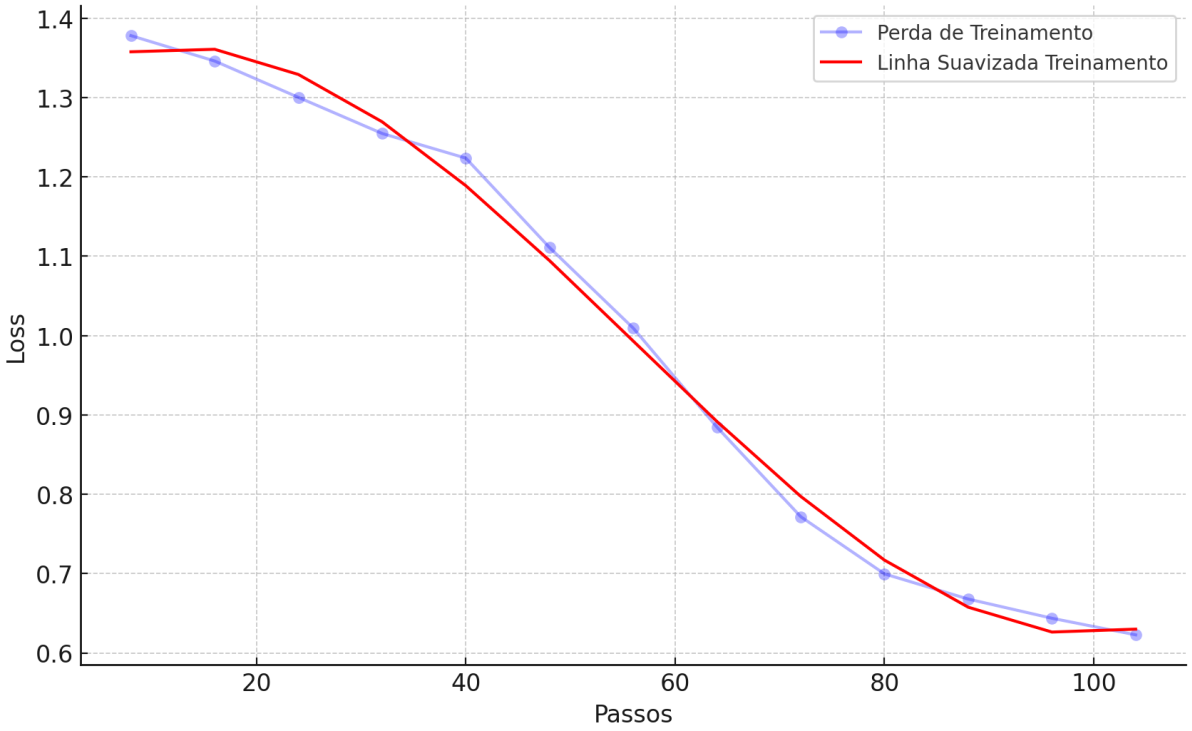

Figura 14 – Curva Loss do treinamento do modelo com dados do prompt 2.

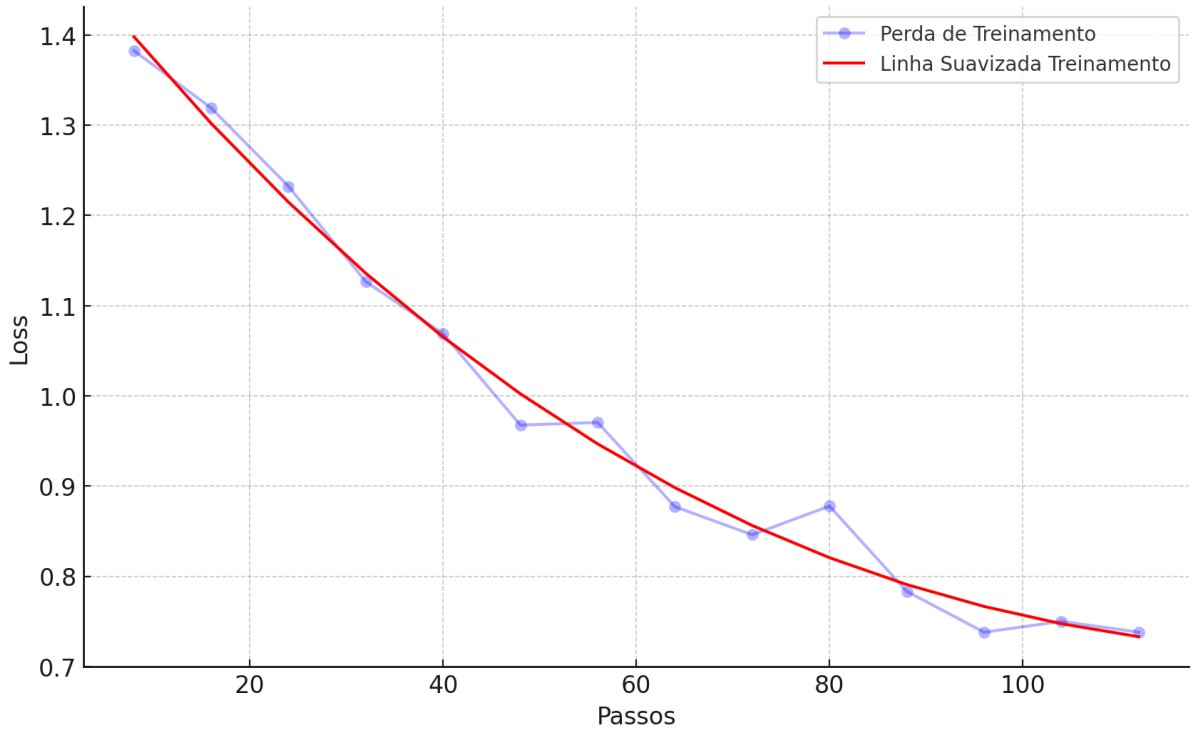

Figura 15 – Curva Loss do treinamento do modelo com dados do prompt 3.

dência de convergência se mantém e o modelo melhora a sua performance. A Figura 16 mostra que o modelo convergiu bem e atingiu valores baixos de perda, 0.28. Ao final do treinamento o modelo atingiu o valor de 0.62 e 0.42 de acurácia e Cohen's Kappa respectivamente.

Conseguimos verificar que o modelo apresenta uma tendência de melhoria à medida que a quantidade de dados aumenta e o tempo de treinamento é prolongado (mais épocas). Contudo, é crucial considerar a relação custo-benefício dessa abordagem, visto que o treinamento dessa arquitetura possui um custo elevado, muitas vezes resultando em ganhos marginais. Portanto, é necessário avaliar cuidadosamente se os benefícios adicionais justificam o investimento de recursos.

Os dados de entrada foram processados e divididos em conjuntos de treinamento e teste (preservamos o conjunto de teste que foi utilizado no Experimento 4). Os resultados deste experimento foram comparados com os modelos anteriores, utilizando métricas como acurácia e coeficiente de Cohen's Kappa, para avaliar a eficácia da abordagem baseada em DeBERTa e aumentação de dados com GPT-4 na classificação das presenças cognitivas e esse comparativo geral pode ser observado na Tabela 9.

A Tabela 9 apresenta os resultados dos experimentos conduzidos para a classificação das categorias da presença cognitiva do CoI. No Experimento 1, os modelos XGBoost, RandomFo-

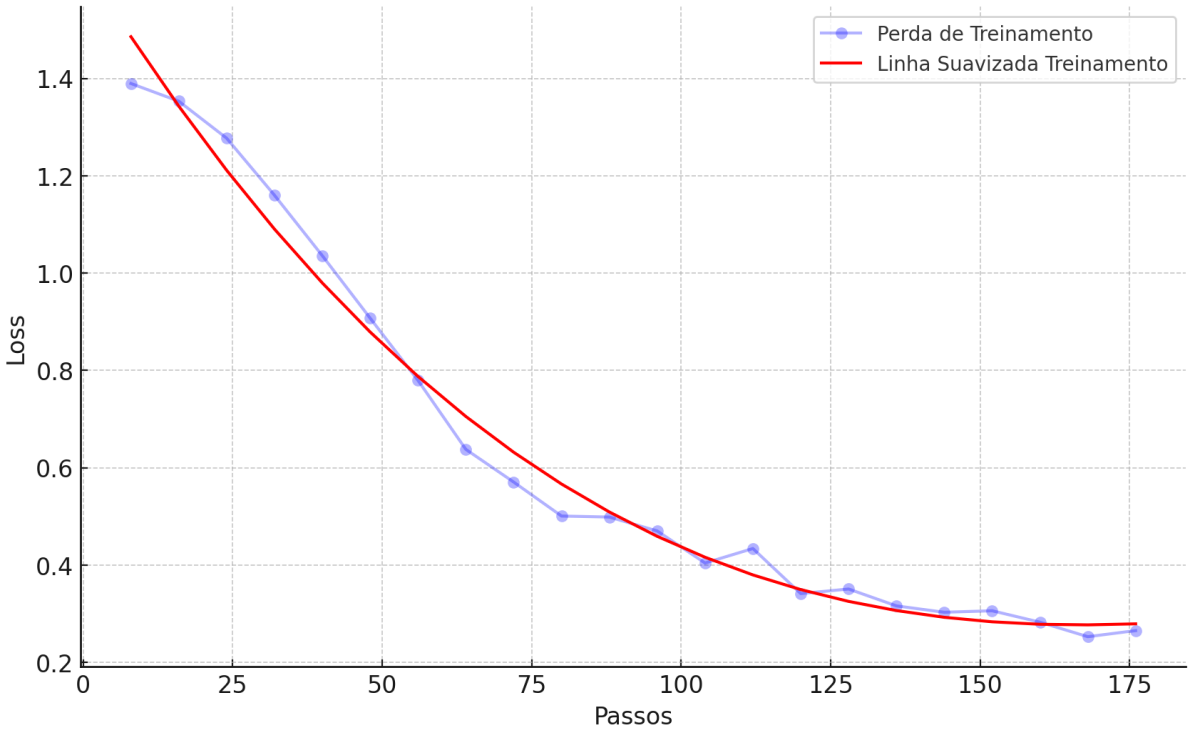

Figura 16 – Curva Loss do treinamento do modelo com dados do prompt 2 com 5 épocas.

Tabela 9 – Resultados gerais dos experimentos: Acurácia e Cohen's Kappa

Experimento	Modelo	Acurácia	Cohen's Kappa
1	XGBoost	0.56	0.35
	RandomForest	0.56	0.34
	MLP	0.55	0.33
2	XGBoost	0.60	0.40
	RandomForest	**0.60**	**0.43**
3	XGBoost (Dados+Aument. BERT)	0.58	0.38
	RandomForest (Dados+Aument. BERT)	0.58	0.40
4	DeBERTa+MLP	0.58	0.38
5	DeBERTa+MLP (Dados+Dados do Prompt 1)	0.59	0.37
	DeBERTa+MLP (Dados+Dados do Prompt 2)	0.59	0.40
	DeBERTa+MLP (Dados+Dados do Prompt 3)	0.59	0.39
	DeBERTa+MLP (Dados+Dados do Prompt 2 com 5 épocas)	**0.62**	**0.42**

rest e MLP foram avaliados sem técnicas de aumentação de dados, com o XGBoost alcançando a maior acurácia de 0.56 e um Cohen's Kappa de 0.35. O Experimento 2 aplicou aprendizado ativo com XGBoost e RandomForest, resultando em uma acurácia de 0.60 para ambos os modelos, mas com o RandomForest apresentando um Cohen's Kappa superior de 0.43. No Experimento 3, a aumentação de dados utilizando BERT gerou resultados modestos, com acurácia de 0.58 para ambos os modelos e melhorias marginais no Cohen's Kappa.

O Experimento 4 avaliou o uso do modelo DeBERTa combinado com MLP, resultando em uma acurácia de 0.58 e um Cohen's Kappa de 0.38, comparável aos resultados do Expe-

rimento 3. No Experimento 5, diferentes prompts para aumentação de dados com o modelo DeBERTa foram testados, todos alcançando uma acurácia de 0.59. A configuração com dados do Prompt 2 e 5 épocas de treinamento obteve a melhor performance, com acurácia de 0.62 e Cohen's Kappa de 0.42. Estes resultados indicam que a combinação de aumentação de dados e um maior número de épocas pode potencializar a eficácia do modelo. Contudo, é necessário avaliar a relação custo-benefício dessas abordagens, considerando os recursos computacionais necessários.

5.6 RESUMO DO CAPÍTULO

Neste capítulo, apresentamos os resultados dos experimentos realizados para classificar as categorias de presença cognitiva em fóruns de discussão online. Utilizamos diversas abordagens de aprendizado de máquina, incluindo aprendizado ativo com modelos como XGBoost e RandomForest, técnicas de aumentação de dados com BERT e GPT-4, e modelos baseados na arquitetura Transformer, como o DeBERTa. Avaliamos o desempenho dos modelos utilizando métricas de acurácia e o coeficiente de Cohen's Kappa para medir a concordância entre as classificações previstas e as reais. Os resultados indicaram que o RandomForest com aprendizado ativo alcançou um Cohen's Kappa de 0,43, enquanto o DeBERTa combinado com aumentação de dados e mais épocas de treinamento atingiu um Cohen's Kappa de 0,42. Embora as técnicas aplicadas tenham proporcionado melhorias, destacamos que o incremento no desempenho deve ser avaliado em relação ao custo computacional envolvido, considerando os ganhos marginais obtidos.

No próximo capítulo, abordaremos as considerações finais deste trabalho, discutindo as limitações encontradas e apontando direções para pesquisas futuras. Serão apresentadas reflexões sobre os resultados obtidos e sugestões para avanços na classificação da presença cognitiva.

6 CONCLUSÃO

A transformação digital no ensino superior destacou a importância do ensino à distância como um elemento crucial para a continuidade da educação em situações adversas. Neste contexto, os fóruns de discussão online, integrados aos Ambientes Virtuais de Aprendizagem (AVA), emergiram como espaços essenciais para a interação e construção de conhecimento entre estudantes e professores (MILMAN, 2015; LIM et al., 2017). A eficácia desses fóruns depende significativamente da capacidade de monitorar e analisar as interações neles contidas, visando otimizar o suporte educativo e promover um ambiente de aprendizado colaborativo e engajador (XIA; FIELDER; SIRAGUSA, 2013).

O modelo de Comunidade de Investigação (CoI), aplicado em ambientes online, é amplamente sustentado pela literatura acadêmica, destacando as presenças social, cognitiva e de ensino como componentes fundamentais para uma experiência educacional eficiente no ensino superior (GARRISON; ANDERSON; ARCHER, 1999). A análise das interações nesse modelo oferece uma visão ampla sobre o processo de aprendizagem, engajamento dos alunos e a qualidade da educação oferecida. Contudo, as metodologias convencionais de análise, principalmente a Análise de Conteúdo Quantitativa (QCA), enfrentam desafios de escalabilidade e aplicação em tempo real devido à natureza laboriosa e prolongada da codificação manual dos dados (DONNELLY; GARDNER, 2011; KOVANOVIC et al., 2014a).

A necessidade de métodos de análise mais ágeis e adaptativos justifica a investigação de abordagens automatizadas que possam processar grandes volumes de dados de forma eficiente. A automação não apenas acelera o processo de análise, mas também proporciona uma resposta educativa mais rápida e informada, potencialmente transformando a dinâmica do aprendizado em cursos online (GAŠEVIĆ et al., 2016). Além disso, a implementação de técnicas de aprendizagem de máquina e análise de redes em fóruns de discussão pode descobrir padrões complexos de interação e conhecimento que são difíceis de identificar manualmente (FERREIRA-MELLO et al., 2019).

No entanto, a eficácia dessas técnicas depende da disponibilidade de dados anotados de alta qualidade, que são escassos e onerosos para produzir, especialmente em um domínio tão especializado como o do ensino superior (FERREIRA et al., 2020). Portanto, este estudo também se dedica a explorar abordagens inovadoras para a geração de dados, como o uso de dados sintéticos, além de técnicas de aprendizagem de máquina, para superar essas limitações

(BARBOSA et al., 2020; ROLIM et al., 2021). Dessa forma, a motivação desta pesquisa é dupla: melhorar a compreensão e a gestão das interações em fóruns online no ensino superior através da aplicação de métodos automatizados de análise de texto e, simultaneamente, enfrentar os desafios associados à preparação de dados para aprendizagem de máquina em ambientes educacionais.

6.1 DISCUSSÃO DOS RESULTADOS

Os experimentos conduzidos nesta pesquisa envolveram a utilização de diferentes técnicas de aprendizado de máquina para a classificação das categorias de presença cognitiva em fóruns de discussão online. Foram utilizados modelos como Random Forest, XGBoost e MLP, além de abordagens de aumento de dados com BERT e GPT-4 para lidar com o desbalanceamento das categorias da presença cognitiva.

Os resultados mostraram que:

- **Random Forest e XGBoost**: Sem técnicas de aumento de dados, esses modelos alcançaram uma acurácia de 0.60, com o Random Forest apresentando um coeficiente de Cohen's Kappa de 0.43. Aumentação de dados com BERT não trouxe melhorias significativas.
- **DeBERTa combinado com MLP**: Este modelo alcançou uma acurácia de 0.58 e um Cohen's Kappa de 0.38. Quando combinados com diferentes prompts de aumento de dados, os resultados variaram, com a melhor configuração atingindo uma acurácia de 0.62 e um Cohen's Kappa de 0.42.
- **Random Forest com Features Linguísticas**: Esta abordagem, utilizando apenas 18% da base de dados, superou o trabalho de referência na área (FARROW; MOORE; GAŠEVIĆ, 2019), demonstrando uma acurácia de 0.60 e um Cohen's Kappa de 0.43, destacando-se como a melhor abordagem dentre as testadas.

Os resultados indicam que a combinação de técnicas avançadas de processamento de linguagem natural e métodos de aumento de dados pode melhorar a precisão e a capacidade de generalização dos modelos. Contudo, a relação custo-benefício dessas abordagens deve ser cuidadosamente avaliada, especialmente em termos de recursos computacionais necessários. Adicionalmente, observou-se que o uso de aprendizado ativo permitiu otimizar o processo de

treinamento, selecionando as instâncias mais informativas para anotação adicional, o que é crucial em contextos com dados limitados.

Apesar dos bons resultados alcançados com DeBERTa e MLP, é importante destacar que o custo computacional associado a esses modelos é substancialmente maior, incluindo o tempo de processamento e a necessidade de *hardware* especializado. Em contraste, a abordagem com Random Forest e características linguísticas mostrou-se não apenas mais eficiente em termos de recursos, mas também mais eficaz em termos de desempenho, utilizando uma fração menor da base de dados.

Uma possível direção para pesquisas futuras é explorar mais a fundo a combinação de características linguísticas e contextuais. A utilização de ferramentas como LIWC e Coh-Metrix, em conjunto com modelos de linguagem avançados, pode fornecer uma representação mais abrangente e detalhada das interações nos fóruns. Além disso, o investimento em técnicas de aprendizado ativo mostrou-se promissor para otimizar o processo de treinamento. Ao selecionar de forma inteligente as instâncias mais informativas para anotação adicional, é possível melhorar a precisão dos modelos mesmo com um volume limitado de dados anotados.

Portanto, a integração de aprendizado ativo com a combinação de características linguísticas e contextuais pode representar um caminho eficiente para alcançar melhorias significativas na classificação das categorias de presença cognitiva. Isso não apenas potencializa a capacidade dos modelos de capturar aspectos sutis das interações textuais, mas também oferece uma solução mais custo-efetiva em termos de recursos computacionais.

RQ1: Como as técnicas de aprendizagem de máquina, combinadas com diferentes abordagens de mineração de texto e representação textual, podem ser utilizadas para classificar as categorias da presença cognitiva e qual é o impacto dessas abordagens na acurácia dos modelos?

Os resultados demonstraram que técnicas de aprendizagem de máquina, como Random Forest, XGBoost e MLP, combinadas com modelos baseados em Transformers, como DeBERTa, podem ser utilizadas com eficácia para classificar automaticamente as categorias da presença cognitiva. Em particular, técnicas recentes de aprendizagem profunda, especialmente aquelas baseadas em modelos Transformers, revelaram-se extremamente eficazes na captura de características contextuais e semânticas das interações textuais. Essa capacidade de representar profundamente o contexto do texto contribuiu significativamente para melhorar a acurácia dos

modelos de classificação.

A integração do aprendizado ativo com técnicas como Random Forest e XGBoost teve um impacto considerável, especialmente quando comparada com os trabalhos de referência (ver o Capítulo 3), permitindo uma análise mais precisa e detalhada das interações dos alunos. O aprendizado ativo demonstrou ser particularmente valioso ao reduzir significativamente a necessidade de dados anotados. Em nossos experimentos, utilizamos apenas 18% dos dados originais e conseguimos superar os resultados obtidos com 100% da base de dados. Este ganho ressalta o potencial do aprendizado ativo para otimizar o uso de recursos e melhorar a eficiência do processo de anotação, resultando em classificações mais acuradas e robustas. Essa abordagem aproxima a aplicação desses modelos em ambientes reais, auxiliando os educadores a proporcionar uma experiência educacional mais satisfatória para os estudantes.

Portanto, o impacto do uso de diferentes técnicas de mineração de texto e representação textual foi significativo. Modelos baseados em Transformers, como o DeBERTa, demonstraram ser especialmente eficazes na captura de nuances semânticas e contextuais, melhorando a acurácia e a robustez dos modelos de classificação. Separadamente, as abordagens tradicionais de aprendizagem de máquina, como Random Forest e XGBoost, também apresentaram resultados promissores, especialmente quando combinadas com técnicas de aprendizado ativo. Ambas as abordagens mostraram potencial em diferentes contextos, oferecendo aos educadores ferramentas eficazes para a análise automatizada da presença cognitiva e para obter uma compreensão mais detalhada do desenvolvimento cognitivo dos estudantes em ambientes de ensino à distância.

RQ2: Quais são as melhores práticas para lidar com a escassez de dados anotados e o desbalanceamento das categorias da presença cognitiva?

Para enfrentar a escassez de dados anotados e o desbalanceamento das categorias, o estudo mostrou que métodos de aumento de dados, utilizando modelos como BERT e GPT-4, são altamente eficazes. A aplicação de aprendizado ativo revelou-se uma estratégia eficiente para otimizar o uso de dados anotados, permitindo que os modelos selecionem as instâncias mais informativas para anotação adicional. Isso resultou em melhorias na precisão e na capacidade de generalização do modelo baseado em redes neurais.

No entanto, não houve ganhos significativos ao aplicar essas técnicas a modelos tradicionais, como Random Forest e XGBoost. Esse resultado indica que a adição de dados gerados

por modelos de linguagem deve ser analisada mais profundamente. É crucial avaliar se as características naturais dos dados originais são preservadas para que a modelagem das categorias da presença cognitiva seja realmente aprimorada. A integridade das características semânticas e contextuais deve ser mantida para garantir que a classificação das categorias de presença cognitiva possa capturar informações importantes e, assim, melhorar efetivamente os modelos.

A implementação de técnicas de aprendizado de máquina para a análise automatizada da presença cognitiva em fóruns de discussão é não só viável, mas também potencialmente enriquecedora para a prática educacional em ambientes virtuais. Essas técnicas permitem monitorar continuamente e de forma detalhada as interações dos alunos, fornecendo aos educadores ferramentas poderosas para apoiar o desenvolvimento cognitivo dos estudantes e melhorar a qualidade do ensino à distância.

6.2 LIMITAÇÕES

Esta pesquisa apresenta várias limitações que devem ser consideradas ao interpretar os resultados e ao planejar estudos futuros. Primeiramente, a utilização de apenas uma base de dados para conduzir os experimentos limita a capacidade de generalização dos modelos desenvolvidos. A eficácia das técnicas de aprendizagem de máquina e dos modelos baseados em Transformers, como DeBERTa, pode variar quando aplicadas a diferentes conjuntos de dados ou contextos educacionais. Portanto, é essencial validar os achados em bases de dados adicionais para confirmar a robustez e a aplicabilidade das abordagens propostas.

Outra limitação significativa é a ausência de integração do modelo gerado em um ambiente real de ensino à distância. Embora os resultados experimentais sejam promissores, a aplicação prática dos modelos em ambientes educacionais reais pode revelar desafios não previstos, como a necessidade de ajustes contínuos nos modelos para lidar com dados dinâmicos e variáveis contextuais específicas. A implementação e avaliação do modelo em um sistema de gestão de aprendizado real seriam passos cruciais para testar sua eficácia e usabilidade em cenários práticos.

Além disso, o treinamento dos modelos foi realizado em um período de tempo limitado, o que pode ter impactado a performance dos modelos mais robustos, como os baseados em Transformers. Treinamentos mais prolongados e extensivos podem levar a melhorias consideráveis nos resultados, permitindo que os modelos capturem melhor as complexidades das interações textuais. Investigações futuras devem considerar a alocação de mais tempo e recur-

sos computacionais para explorar plenamente o potencial desses modelos.

Essas limitações indicam áreas importantes para pesquisas futuras, destacando a necessidade de validação adicional, integração prática e otimização contínua dos modelos para melhorar a análise automatizada da presença cognitiva em ambientes de ensino à distância.

6.3 TRABALHOS FUTUROS

A pesquisa realizada abre várias direções promissoras para trabalhos futuros, que podem ampliar e aprofundar os achados deste estudo.

6.3.1 Validação em Diferentes Bases de Dados

Uma das principais limitações desta pesquisa foi a utilização de uma única base de dados. Para validar a generalização dos modelos desenvolvidos, é essencial aplicar e testar essas abordagens em diferentes conjuntos de dados provenientes de variados contextos educacionais. A diversidade nas bases de dados ajudará a avaliar a robustez e a adaptabilidade dos modelos, garantindo que eles sejam aplicáveis a uma ampla gama de situações e populações estudantis.

6.3.2 Integração em Ambientes Reais

Outro passo crucial é a integração dos modelos desenvolvidos em sistemas de gestão de aprendizagem reais. Testar os modelos em ambientes de ensino à distância permitirá observar como eles se comportam com dados dinâmicos e variáveis contextuais específicas. Além disso, essa integração pode fornecer feedback prático sobre a usabilidade e eficácia dos modelos, ajudando a identificar possíveis melhorias e ajustes necessários para uma aplicação bem-sucedida no dia a dia educacional.

6.3.3 Treinamento de Modelos Mais Robustos

Os modelos baseados em Transformers, como DeBERTa, mostraram um potencial significativo, mas ainda há espaço para otimização. Investir em treinamentos mais prolongados e com maior poder computacional pode levar a melhorias substanciais nos resultados. Futuras pesquisas devem explorar o treinamento desses modelos com mais dados e por períodos mais

longos, investigando como ajustes finos e técnicas de otimização podem aprimorar ainda mais a acurácia e a robustez das classificações.

6.3.4 Combinação de Features Linguísticas e Contextuais

A combinação de características linguísticas e contextuais mostrou-se promissora. Trabalhos futuros devem investigar mais profundamente como integrar ferramentas como LIWC e Coh-Metrix com modelos de linguagem avançados para obter uma representação mais rica e detalhada das interações nos fóruns. Isso pode incluir o desenvolvimento de novas técnicas para a extração e combinação de features que capturem melhor as características das interações textuais.

6.3.5 Desenvolvimento de Interfaces Interativas para Educadores

Para maximizar o impacto prático das análises automatizadas, o desenvolvimento de interfaces interativas que apresentem os resultados de maneira intuitiva e acionável para os educadores é essencial. Essas interfaces podem fornecer insights em tempo real sobre o desenvolvimento cognitivo dos estudantes, ajudando os educadores a tomar decisões informadas e a fornecer feedback personalizado. Pesquisas futuras devem focar no design e na implementação dessas ferramentas, garantindo que elas sejam eficazes e fáceis de usar.

Essas direções futuras oferecem um caminho claro para expandir e enriquecer a pesquisa sobre a análise automatizada da presença cognitiva, contribuindo para a melhoria contínua da educação à distância.

6.4 PUBLICAÇÕES

Durante o desenvolvimento desta pesquisa, foram realizadas várias publicações em conferências e revistas acadêmicas, destacando os avanços e resultados obtidos. As principais publicações incluem:

- "Analysing Social Presence in Online Discussions Through Network and Text Analytics", publicado na conferência ICALT do ano de 2019 (**Publicado**);

- "An Analysis of the use of Good Feedback Practices in Online Learning Courses", publicado na conferência ICALT do ano de 2019 (**Publicado**);
- "Identifying Students' Weaknesses and Strengths Based on Online Discussion using Topic Modeling", publicado na conferência ICALT do ano de 2019 (**Publicado**);
- "A Chatbot to Support Basic Students Questions", publicado na conferência LACLO do ano de 2021;
- "Uma Análise entre Boas Práticas de Feedback em Ambientes Virtuais de Aprendizagem", publicado na conferência SBIE do ano de 2020 (**Publicado**);
- "Análise de Discussões em Fóruns Educacionais Usando Mineração de Texto e Análise de Grafos", publicado como capítulo de livro no JAIE (CBIE) do ano de 2020 (**Publicado**);
- "Towards automatic content analysis of social presence in transcripts of online discussions", publicado na conferência LAK do ano de 2020 (**Publicado**);
- "Let's shine together!: a comparative study between learning analytics and educational data mining", publicado na conferência LAK do ano de 2020 (**Publicado**);
- "Reducing the size of training datasets in the classification of online discussions", publicado na conferência ICALT do ano de 2021 (**Publicado**);
- "Automatic Content Analysis of Online Discussions for Cognitive Presence: A Study of the Generalizability Across Educational Contexts", publicado na revista IEEE Transactions on Learning Technologies do ano de 2021 (**Publicado**).
- "Three-Layer Denoiser: Denoising Parallel Corpora for NMT Systems", publicado na conferência ENIAC do ano de 2023 (**Publicado**).
- "A Comparative Analysis Between Good Feedback Descriptors on Online Courses", publicado na conferência SBIE do ano de 2023 (**Publicado**).
- "Towards Explainable Automatic Punctuation Restoration for Portuguese Using Transformers", publicado na revista Expert Systems With Applications no ano de 2024 (**Publicado**).

Além desses, houveram outros trabalhos submetidos, porém que não foram publicados, desses listamos:

- "Sobre o que Falamos? Uma Análise dos Últimos 19 Anos do Simpósio Brasileiro de Informática na Educação", submetido para a conferência SBIE do ano de 2020 (**Não publicado**);
- "Quem Somos? Análise de Autores em 19 anos do Simpósio Brasileiro de Informática na Educação", submetido para a conferência SBIE do ano de 2020 (**Não publicado**);
- "Análise de Ligações Implícitas em Fóruns Educacionais Usando Redes Epistêmicas", submetido para a conferência SBIE do ano de 2020 (**Não publicado**);
- "Assessing Topic Modeling Techniques on Multiple Corpora: What Should We Use?", submetido para a conferência BRACIS do ano de 2021 (**Não publicado**);

REFERÊNCIAS

ANDERSON, L. W.; SOSNIAK, L. A. *Bloom's taxonomy*. [S.l.]: Univ. Chicago Press Chicago, IL, 1994.

ANDERSON, T.; ROURKE, L.; GARRISON, D. R.; ARCHER, W. Assessing Teaching Presence in a Computer Conferencing Context. *Journal of Asynchronous Learning Networks*, v. 5, p. 1–17, 2001.

BA, S.; HU, X.; STEIN, D.; LIU, Q. Assessing cognitive presence in online inquiry-based discussion through text classification and epistemic network analysis. *British Journal of Educational Technology*, Wiley Online Library, v. 54, n. 1, p. 247–266, 2023.

BARBOSA, A.; FERREIRA, M.; MELLO, R. F.; LINS, R.; DUEIRE, R.; GAšEVIĆ, D. The impact of automatic text translation on classification of online discussions for social and cognitive presences. In: ACM. *Proc. 11th Int. Conf. Learning Analytics & Knowledge (LAK'21)*. [S.l.], 2021. p. 77–87.

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.

BIGGS, J. B.; COLLIS, K. F. *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. [S.l.]: Academic Press, 2014.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006.

BOZKURT, A.; SHARMA, R. C. Challenging the status quo and exploring the new boundaries in the age of algorithms: Reimagining the role of generative ai in distance education and online learning. *Asian Journal of Distance Education*, v. 18, n. 1, 2023.

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.

BREIMAN, L. et al. Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. p. 785–794, 2016.

CHEN, T.; HE, T.; BENESTY, M.; KHOTILOVICH, V.; TANG, Y.; CHO, H.; CHEN, K.; MITCHELL, R.; CANO, I.; ZHOU, T. et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, v. 1, n. 4, p. 1–4, 2015.

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.

CORICH, S.; HUNT, K.; HUNT, L. Computerised content analysis for measuring critical thinking within discussion forums. *Journal of e-learning and knowledge society*, Italian e-Learning Association, v. 2, n. 1, 2006.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, Kluwer Academic Publishers, v. 20, n. 3, p. 273–297, 1995.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DONNELLY, R.; GARDNER, J. Content analysis of computer conferencing transcripts. *Interactive learning environments*, Taylor & Francis, v. 19, n. 4, p. 303–315, 2011. Disponível em: <https://doi.org/10.1080/10494820903075722>.

ERICKSON, N.; MUELLER, J.; SHIRKOV, A.; ZHANG, H.; LARROY, P.; LI, M.; SMOLA, A. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

FARROW, E.; MOORE, J.; GAŠEVIĆ, D. Analysing discussion forum data: a replication study avoiding data contamination. In: *LAK' 2019*. [S.l.: s.n.], 2019. p. 170–179.

FARROW, E.; MOORE, J.; GAŠEVIĆ, D. Dialogue attributes that inform depth and quality of participation in course discussion forums. In: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. [S.l.: s.n.], 2020. p. 129–134.

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.

FERREIRA-MELLO, R.; ANDRÉ, M.; PINHEIRO, A.; COSTA, E.; ROMERO, C. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, p. e1332, 2019. Disponível em: <https://doi.org/10.1002/widm.1332>.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, Elsevier, v. 55, n. 1, p. 119–139, 1997.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

GARRISON, D. R. Cognitive presence for effective asynchronous online learning: The role of reflective inquiry, self-direction and metacognition. *Elements of quality online education: Practice and direction*, v. 4, n. 1, p. 47–58, 2003.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education*, 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. Disponível em: <https://doi.org/10.1080/08923640109527071>.

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.; DAWSON, S.; ROGERS, T.; GASEVIC, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, Elsevier, v. 28, p. 68–84, 2016.

GAŠEVIĆ, D.; JOKSIMOVIĆ, S.; EAGAN, B. R.; SHAFFER, D. W. Sens: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior*, Elsevier, 2018.

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*, v. 3, n. 1, p. 63–78, 2017.

GENG, S.; NIU, B.; FENG, Y.; HUANG, M. Understanding the focal points and sentiment of learners in mooc reviews: A machine learning and sc-liwc-based approach. *British Journal of Educational Technology*, Wiley Online Library, v. 51, n. 5, p. 1785–1803, 2020.

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.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. [S.l.]: Springer Series in Statistics, 2009.

HE, P.; GAO, J.; CHEN, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

HE, P.; LIU, X.; GAO, J.; CHEN, W. Deberta: Decoding-enhanced bert with disentangled attention. *International Conference on Learning Representations*, 2021.

HE, X.; PAN, J.; JIN, O.; XU, T.; LIU, B.; XU, T.; SHI, Y.; ATALLAH, A.; HERBRICH, R.; BOWERS, S.; CANDELA, J. Q. Practical lessons from predicting clicks on ads at facebook. In: ACM. *Proceedings of the eighth international workshop on data mining for online advertising*. [S.l.], 2014. p. 1–9.

HOSLER, K. A.; AREND, B. D. The importance of course design, feedback, and facilitation: Student perceptions of the relationship between teaching presence and cognitive presence. *Educational Media International*, Taylor & Francis, v. 49, n. 3, p. 217–229, 2012.

HU, Y.; MELLO, R. F.; GAŠEVIĆ, D. Automatic analysis of cognitive presence in online discussions: An approach using deep learning and explainable artificial intelligence. *Computers and Education: Artificial Intelligence*, Elsevier, v. 2, p. 100037, 2021.

KARUMBAIAH, S.; LAN, A.; NAGPAL, S.; BAKER, R. S.; BOTELHO, A.; HEFFERNAN, N. Using past data to warm start active machine learning: Does context matter? In: *LAK21: 11th International Learning Analytics and Knowledge Conference*. [S.l.: s.n.], 2021. p. 151–160.

KHOSHGOFTAAR, T. M.; ALLEN, E. B. Controlling overfitting in classification-tree models of software quality. *Empirical Software Engineering*, Springer, v. 6, p. 59–79, 2001.

KILAG, O. K.; OBANER, E.; VIDAL, E.; CASTAÑARES, J.; DUMDUM, J. N.; HERMOSA, T. J. Optimizing education: Building blended learning curricula with lms. *Excellencia: International Multi-disciplinary Journal of Education (2994-9521)*, v. 1, n. 4, p. 238–250, 2023.

KONOPKA, C. L.; ADAIME, M. B.; MOSELE, P. H. Active teaching and learning methodologies: some considerations. *Creative Education*, Scientific Research Publishing, v. 6, n. 14, p. 1536–1545, 2015.

KOVANOVIC, V.; JOKSIMOVIC, S.; GASEVIC, D.; HATALA, M. Automated cognitive presence detection in online discussion transcripts. In: *LAK Workshops*. [S.l.: s.n.], 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: ACM. *Proceedings of the sixth international conference on learning analytics & knowledge*. [S.l.], 2016. p. 15–24.

KOVANOVIć, V.; JOKSIMOVIć, S.; WATERS, Z.; GAsEVIć, 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*. New York, NY, USA: Association for Computing Machinery, 2016. (LAK '16), p. 15–24. ISBN 9781450341905. Disponível em: <https://doi.org/10.1145/2883851.2883950>.

KOVANOVIć, V.; JOKSIMOVIć, S.; GAšEVIć, D.; HATALA, M. Automated cognitive presence detection in online discussion transcripts. In: *Proceedings of the Workshops at the LAK 2014 Conference co-located with 4th International Conference on Learning Analytics and Knowledge (LAK'14)*. Indianapolis, IN: [s.n.], 2014. Disponível em: <http://ceur-ws.org/Vol-1137/>.

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 (LAK'16)*. New York, NY, USA: ACM, 2016. (LAK '16), p. 15–24.

KUMAR, V.; CHOUDHARY, A.; CHO, E. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020.

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

LEE, J.; SOLEIMANI, F.; IV, J. H.; SOYLU, M. Y.; FINKELBERG, R.; CHATTERJEE, S. Predicting cognitive presence in at-scale online learning: Mooc and for-credit online course environments. *Online Learning*, ERIC, v. 26, n. 1, p. 58–79, 2022.

LIM, V.; WEE, L.; TEO, J.; NG, S. Massive open and online courses and open education resources in singapore. *arXiv preprint arXiv:1708.08743*, 2017.

LIU., Z.; KANG., L.; DOMANSKA., M.; LIU., S.; SUN., J.; FANG., C. Social network characteristics of learners in a course forum and their relationship to learning outcomes. In: INSTICC. *Proceedings of the 10th International Conference on Computer Supported Education - Volume 2: CSEDU,*. [S.l.]: SciTePress, 2018. p. 15–21. ISBN 978-989-758-291-2.

MANNING, C. D.; MANNING, C. D.; SCHÜTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT press, 1999.

MCCARTHY, P. M.; LEWIS, G. A.; DUFTY, D. F.; MCNAMARA, D. S. Analyzing writing styles with coh-metrix. In: *FLAIRS Conference*. [S.l.: s.n.], 2006. p. 764–769.

MCHUGH, M. L. Interrater reliability: the kappa statistic. *Biochemia medica*, Medicinska naklada, v. 22, n. 3, p. 276–282, 2012.

MCKLIN, T. E. *Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network*. Tese (Doutorado), Atlanta, GA, USA, 2004. AAI3190967. Disponível em: <https://scholarworks.gsu.edu/msit_diss/1/>.

MCNAMARA, D. S.; GRAESSER, A. C.; MCCARTHY, P. M.; CAI, Z. *Automated evaluation of text and discourse with Coh-Metrix*. [S.l.]: Cambridge University Press, 2014.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 3111–3119.

MILMAN, N. B. Distance education. Elsevier, 2015.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, Inc., 1997.

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.

NETO, V. et al. Automatic content analysis of online discussions for cognitive presence: A study of the generalizability across educational contexts. *IEEE Transactions on Learning Technologies*, 2021.

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.

NIELSEN, M. A. *Tree boosting with XGBoost - Why does XGBoost win 'every' machine learning competition?* 2016. Disponível em: <http://mlexplained.com/2016/04/05/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition/>.

OLIVARES, D.; MELLO, R. F. L. de; ADESOPE, O.; ROLIM, V.; GAŠEVIC, D.; HUNDHAUSEN, C. Using social network analysis to measure the effect of learning analytics in computing education. In: IEEE. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*. 2019. v. 2161, p. 145–149. Disponível em: <https://doi.org/10.1109/ICALT.2019.00044>.

PENNEBAKER, J. W.; BOYD, R. L.; JORDAN, K.; BLACKBURN, K. *The development and psychometric properties of LIWC2015*. [S.l.], 2015.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: . [S.l.: s.n.], 2014. p. 1532–1543.

PRATAMA, M. P.; SAMPELOLO, R.; LURA, H. Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. *Klasikal: Journal of education, language teaching and science*, v. 5, n. 2, p. 350–357, 2023.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, n. 1, p. 81–106, 1986.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. et al. Improving language understanding by generative pre-training. OpenAI, 2018.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, v. 21, n. 140, p. 1–67, 2020.

ROLIM, V.; MELLO, R. F.; NASCIMENTO, A.; LINS, R. D.; GAšEVIć, D. Reducing the size of training datasets in the classification of online discussions. In: *2021 International Conference on Advanced Learning Technologies (ICALT)*. [S.l.: s.n.], 2021. p. 179–183.

ROLIM, V.; MELLO, R. F. L. de; FERREIRA, M.; CAVALCANTI, A. P.; LIMA, R. Identifying students' weaknesses and strengths based on online discussion using topic modeling. In: IEEE. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.], 2019. v. 2161, p. 63–65.

ROURKE, L.; ANDERSON, T.; GARRISON, D. R.; ARCHER, W. Assessing social presence in asynchronous text-based computer conferencing. Journal of Distance Education, 1999.

ROURKE, L.; ANDERSON, T.; GARRISON, D. R.; ARCHER, W. Methodological issues in the content analysis of computer conference transcripts. *International journal of artificial intelligence in education (IJAIED)*, v. 12, p. 8–22, 2001.

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

SCOTT, J. *Social network analysis*. [S.l.]: Sage, 2017.

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

SETTLES, B. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences, 2009.

SHARMA, M.; BILGIC, M. Evidence-based uncertainty sampling for active learning. *Data Mining and Knowledge Discovery*, Springer, v. 31, n. 1, p. 164–202, 2017.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. [S.l.]: MIT Press, 2018.

SWAN, K. Developing social presence in online course discussions. *Learning and teaching with technology: Principles and practices*, p. 147–164, 2003.

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.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

WHITESIDE, A. L.; DIKKERS, A. G.; SWAN, K. *Social presence in online learning: Multiple perspectives on practice and research*. [S.l.]: Taylor & Francis, 2023.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M. et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. [S.l.: s.n.], 2020. p. 38–45.

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

YUSOF, N.; RAHMAN, A. A. et al. Students' interactions in online asynchronous discussion forum: A social network analysis. In: IEEE. *2009 International Conference on Education Technology and Computer*. [S.l.], 2009. p. 25–29. ISSN 2155-1812.

ZOU, W.; HU, X.; PAN, Z.; LI, C.; CAI, Y.; LIU, M. Exploring the relationship between social presence and learners' prestige in mooc discussion forums using automated content analysis and social network analysis. *Computers in Human Behavior*, Elsevier, v. 115, p. 106582, 2021.

ZOU, W.; PAN, Z.; LI, C.; LIU, M. Does social presence play a role in learners' positions in mooc learner network? a machine learning approach to analyze social presence in discussion forums. In: SPRINGER. *Proc. 3rd Int. Conf. Quantitative Ethnography (ICQE'21)*. [S.l.], 2021. p. 248–264.