Henrique Carlos Fonte Boa CarvalhoCarvalho, Henrique C. F. B

Filtragem Baseada em Comentários para Recomendação de Recursos Educacionais em Plataformas de Conteúdos Diversificados

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2024

Filtragem Baseada em Comentários para Recomendação de Recursos Educacionais em Plataformas de Conteúdos Diversificados

Tese de doutorado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Fabiano Azevedo Dorça
Coorientador: Cristiano Grijó Pitangui

Uberlândia
2024

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

C331
2024

Carvalho, Henrique Carlos Fonte Bôa, 1990-
Filtragem Baseada em Comentários para Recomendação de Recursos Educacionais em Plataformas de Conteúdos Diversificados [recurso eletrônico] / Henrique Carlos Fonte Bôa Carvalho. - 2024.

Orientador: Fabiano Azevedo Dorça.
Coorientador: Cristiano Grijó Pitangui.
Tese (Doutorado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.te.2024.654
Inclui bibliografia.

1. Computação. I. Dorça, Fabiano Azevedo,1979-, (Orient.). II. Pitangui, Cristiano Grijó,1982-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Tese, 35/2024, PPGCO				
Data:	06 de setembro de 2024	Hora de início:	15:10	Hora de encerramento:	18:15
Matrícula do Discente:	12123CCP001				
Nome do Discente:	Henrique Carlos Fonte Boa Carvalho				
Título do Trabalho:	Filtragem Baseada em Comentários para Recomendação de Recursos Educacionais em Plataformas de Conteúdos Diversificados				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	----------				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Cristiano Grijó Pitangui - Dtech/UFSJ (Coorientador) , Márcia Aparecida Fernandes- FACOM/UFU, Rafael Dias Araújo - FACOM/UFU , Alessandro Vivas Andrade - UFVJM - CAMPUS JK, Thiago Rodrigues de Oliveira - UFSJ e Fabiano Azevedo Dorça - FACOM/UFU, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: Cristiano Grijó Pitangui - Ouro Branco/MG, Alessandro Vivas Andrade - Diamantina/MG e Thiago Rodrigues de Oliveira - Ouro Branco/MG . Os outros membros da banca e o aluno participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Fabiano Azevedo Dorça, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação da Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir ao candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Doutor.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Cristiano Grijó Pitangui**, **Usuário Externo**, em 11/09/2024, às 14:17, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Alessandro Vivas Andrade**, **Usuário Externo**, em 11/09/2024, às 16:17, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rafael Dias Araújo**, **Professor(a) do Magistério Superior**, em 11/09/2024, às 17:27, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Fabiano Azevedo Dorça**, **Professor(a) do Magistério Superior**, em 12/09/2024, às 14:42, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Thiago Rodrigues de Oliveira**, **Usuário Externo**, em 16/09/2024, às 09:08, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Márcia Aparecida Fernandes**, **Professor(a) do Magistério Superior**, em 16/09/2024, às 13:14, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5645010** e o código CRC **5C5FE2C8**.

Referência: Processo nº 23117.055654/2024-44 SEI nº 5645010

Este trabalho é uma homenagem às crianças adultas que, desde tenra idade, nutriram o sonho de se tornarem cientistas. É uma homenagem àquela curiosidade sem limites, ao brilho nos olhos ao ver uma experiência bem-sucedida, e à frustração que, ao invés de desencorajar, apenas alimentava mais perguntas. É dedicado para todos que se encantaram com as estrelas, questionaram o movimento das ondas, se intrigaram com o azul do céu, ou se fascinaram com os mistérios por trás da internet, e escolheram embarcar na jornada em busca de respostas. É uma celebração da imaginação, da curiosidade e do desejo de explorar o desconhecido que tantos de nós tivemos em nossa infância.

Esta é uma celebração da jornada do conhecimento, não apenas dos seus destinos. A ciência, afinal, é mais sobre fazer perguntas do que encontrar respostas definitivas. É um convite a olhar o mundo com maravilhamento e curiosidade, e nunca parar de aprender, independentemente da idade ou do caminho profissional escolhido.

Dedicamos também este trabalho às famílias dos sonhadores: pais, mães, irmãos, irmãs, cônjuges e filhos. Sem o vosso suporte inabalável, os sonhos desses aspirantes a cientistas jamais poderiam florescer nem alcançar as estrelas, os sonhos permaneceriam adormecidos, e as estrelas, apenas distantes luzes no céu. Que esta obra seja um reflexo do nosso coletivo desejo de conhecer, um tributo àquela chama eterna de curiosidade que nos define e nos une na maravilhosa jornada da vida.

Agradecimentos

Agradeço sinceramente a todos que contribuíram, direta ou indiretamente, para o sucesso da minha tese.

Em primeiro lugar, um agradecimento especial à minha família. Minha esposa, Aline, e minhas filhas, Sabrina e Catarina, foram a minha grande fonte de amor e inspiração. Agradeço também aos meus pais, Fernando e Sheila, à minha irmã Paloma, ao meu cunhado Fernando e ao meu sobrinho Davi, assim como a todos os meus familiares, sogros, concunhados e sobrinhos, pelo incentivo e encorajamento constantes, que foram essenciais para manter minha motivação e determinação ao longo deste percurso.

Sou profundamente grato aos meus orientadores, Fabiano Dorça e Cristiano Pitangui, cuja orientação valiosa e apoio acadêmico em cada etapa deste projeto foram inestimáveis. Suas críticas construtivas e sugestões enriquecedoras foram cruciais para o desenvolvimento e aprimoramento do meu trabalho e das minhas habilidades de pesquisa.

Agradeço imensamente aos meus amigos do DTI da UFVJM, cujo apoio tornou esta jornada possível. Um agradecimento especial a Clayton, William, Everton, Patrick e Rochelly, cuja amizade e colaboração foram fundamentais.

Estendo também minha gratidão aos meus colegas das pós-graduações em Educação e Ciência da Computação, em especial a Aline, Dayane, Eduardo e Jhonathan.

Por fim, obrigado a todos que tiveram um papel no sucesso deste projeto. Sou profundamente grato pela oportunidade de trabalhar em um tema tão empolgante e significativo, e valorizo cada contribuição que tornou este trabalho possível.

"I am a great believer in luck. The harder I work, the more of it I seem to have."
(COX, 1922)

Resumo

A evolução tecnológica tem promovido uma sociedade cada vez mais conectada, facilitando a interação entre pessoas e o compartilhamento massivo de conteúdos. Esse avanço impacta positivamente várias áreas do conhecimento, incluindo a educação, onde a quantidade de materiais disponíveis cresce exponencialmente. No entanto, essa abundância de recursos educacionais traz desafios, como a dificuldade em identificar e escolher os mais adequados em meio a um vasto acervo de conteúdos. Esses desafios são ainda maiores em repositórios não estritamente educacionais, como Wikipedia, LinkedIn, YouTube, TikTok, Vimeo, entre outros, onde os conteúdos são compartilhados por usuários de diferentes áreas e interesses, incluindo materiais educacionais. O presente trabalho inova ao desenvolver uma abordagem que utiliza os comentários dos usuários juntamente com técnicas de Aprendizado de Máquina (AM) para recomendar Objetos de Aprendizagem (OA) em ambientes com conteúdos diversificados. Para isso, foram utilizados os vocábulos mais frequentes em cada classe, educacional ou não educacional. Duas variações foram desenvolvidas: a variação rígida e a variação flexível. A variação rígida utiliza algoritmos de AM para classificar os vídeos como educacionais ou não educacionais com base nos vocábulos mais frequentes, recomendando vídeos que o algoritmo tem "certeza" de serem educacionais. Já a variação flexível classifica individualmente cada comentário como educacional ou não educacional, analisando a classificação de todos os comentários dos vídeos e recomendando-os com um certo "grau de certeza" de pertencerem à classe educacional. Os resultados obtidos revelaram que os comentários são, de fato, uma excelente característica para a classificação de OAs, especialmente ao utilizar os vocábulos mais frequentes de cada classe. Experimentos indicam que a abordagem permite identificar OAs com uma acurácia impressionante de 95%. Além disso, a variação flexível demonstra maior adaptabilidade para trabalhar com o mundo real, possibilitando uma melhor recomendação de materiais com diferentes quantidades de comentários. Por fim, foi desenvolvido o Learning Object Intelligent Search (LOIS), um Sistema de Recomendação (SR) que auxilia docentes e discentes em Ambientes Virtuais de Aprendizagem (AVA) a encontrar OAs no YouTube.

Palavras-chave: Aprendizado de Máquina, Comentários, Objetos de Aprendizagem, Recursos Educacionais, Sistema de Recomendação, Youtube.

Abstract

The technological evolution has promoted an increasingly connected society, facilitating interaction among people and the massive sharing of content. This advancement positively impacts various areas of knowledge, including education, where the amount of available materials grows exponentially. However, this abundance of educational resources brings challenges, such as the difficulty in identifying and choosing the most suitable ones amidst a vast array of content. These challenges are even greater in non-strictly educational repositories, such as Wikipedia, LinkedIn, YouTube, TikTok, Vimeo, among others, where content is shared by users from different areas and interests, including educational materials. This work innovates by developing an approach that uses user comments along with AM techniques to recommend Learning Objects (LO) in environments with diverse content. For this, the most frequent vocabularies in each class, educational or non-educational, were used. Two variations were developed: the rigid variation and the flexible variation. The rigid variation uses Machine Learning (ML) algorithms to classify videos as educational or non-educational based on the most frequent vocabularies, recommending videos that the algorithm is certain" are educational. The flexible variation classifies each comment individually as educational or non-educational, analyzing the classification of all comments on the videos and recommending them with a certain degree of certainty" of belonging to the educational class. The results obtained revealed that comments are, in fact, an excellent feature for the classification of LOs, especially when using the most frequent vocabularies of each class. Experiments indicate that the approach allows identifying LOs with an impressive accuracy of 95%. Additionally, the flexible variation demonstrates greater adaptability to work with the real world, enabling better recommendation of materials with different quantities of comments. Finally, the LOIS was developed, a Recommendation System (RS) that assists teachers and students in Virtual Learning Environment (VLE) in finding LOs on YouTube.

Keywords: Comments, Educational Resources, Learning Objects, Machine Learning,

Recommender System, Youtube.

Lista de ilustrações

Lista de tabelas

Lista de siglas

AM Aprendizado de Máquina

AVA Ambientes Virtuais de Aprendizagem

BoW Bag of Words

CNN Convolutional Neural Network

DNN Deep Neural Network

LOIS Learning Object Intelligent Search

LO Learning Objects

ML Machine Learning

MT Mineração de Textos

NN Neural Network

OA Objetos de Aprendizagem

RS Recommendation System

RNA Redes Neurais Artificiais

SR Sistema de Recomendação

VLE Virtual Learning Environment

Sumário

Capítulo 1

Introdução

O avanço da tecnologia tem desempenhado um papel fundamental na criação de uma sociedade cada vez mais conectada. Através da *internet* e dos dispositivos eletrônicos, as pessoas têm acesso a uma infinidade de informações ao seu alcance, em qualquer momento e em qualquer lugar. Além disso, a tecnologia proporciona diversas formas de interação social, permitindo que as pessoas se conectem umas com as outras, independentemente da distância geográfica (NASCIMENTO et al., 2017). As redes sociais, por exemplo, possibilitam a troca de mensagens instantâneas, compartilhamento de fotos e vídeos, participação em grupos de discussão, ampliando as oportunidades de interação e fortalecendo os laços sociais.

No entanto, é importante destacar que o avanço tecnológico traz consigo desafios e questões a serem analisadas. Embora o acesso a informações e a interação social sejam vantagens significativas, é necessário ter senso crítico quanto as informações que recebemos assim como verificar a qualidade e a veracidade dessas informações. No universo educacional o aumento de dados nas redes propiciou um enorme avanço, porém, o elevado acervo digital apresenta desafios, quanto a qualidade e a confiabilidade das informações, e a seleção de conteúdos aos que atendam a necessidade. Esse vasto acervo de conteúdo apresenta desvantagem, como apontado por Miranda (2004): "Na área da educação, por exemplo, embora existam muitos materiais sendo criados e disponibilizados, o acesso a eles torna-se um processo cansativo e muitas vezes fracassado".

O fracasso no processo de busca ocorre em grande parte pela enorme quantidade de documentos apresentados ao usuário, o que dificulta e causa confusão no momento de selecionar os mais relevantes (BRAGA; MENEZES, 2014). O excesso de dados e conteúdos disponíveis prejudicam o processo de ensino-aprendizagem, fazendo com que docentes e discentes utilizem a maior parte do tempo em busca de conteúdo, em vez do próprio estudo ou ensino. Neste sentido, a fácil localização e utilização de materiais é de suma importância nesse processo (VIEIRA; NUNES, 2012).

A busca por recursos educacionais em diversos repositórios, como o YouTube, Wikipedia, Linkedin, Reddit, entre outros, não restritos aos específicos de OA, podem apresentar desafios adicionais devido à falta de um planejamento sistemático e padronizado para disponibilizar materiais educacionais de qualidade. Essa falta de estrutura pode dificultar a localização de conteúdo relevante e confiável, exigindo habilidades de pesquisa e avaliação dos seus usuários. Além disso, a diversidade e a variedade de conteúdos disponíveis nesses repositórios podem

tornar a seleção de materiais apropriados um desafio, pois os recursos que são educacionais não são organizados de acordo com um padrão educacional específico. Desta forma, os usuários que buscam conteúdos educacionais precisam adotar uma abordagem cautelosa na busca por recursos que atendam as necessidades, verificando o conteúdo fornecido, a credibilidade dos produtores de conteúdo, avaliando comentários e metadados, e procurando recomendações de fontes confiáveis. Essa busca pode ser exaustiva e demandar tempo e esforço para identificar os materiais adequados, porém, possibilita acessar uma ampla quantidade de materiais educacionais além dos que são fornecidos por sistemas educacionais formais.

Dos diversos tipos de materiais encontrados na atualidade, os vídeos são a forma dominante de mídia entre a nova geração, e sua produção é facilitada pela disponibilidade generalizada de tecnologias, como smartphones e notebooks, permitindo sua criação e compartilhamento de forma mais acessível (PIRHONEN; RASI, 2017; BRAME, 2016).

A combinação de elementos visuais e sonoros nos vídeos torna-os altamente atrativos do ponto de vista cognitivo. Além disso, os repositórios de armazenamento e compartilhamento desempenham um papel essencial na disseminação e popularização dessa mídia. O YouTube, em particular, se destaca como uma plataforma que não só abriga vídeos de conteúdo geral, mas também serve como um espaço onde os usuários disponibilizam materiais educacionais (BARRÉRE et al., 2020).

Considerando a diversidade de aplicativos que também poderiam ser utilizados com propósitos educacionais, a importância dos vídeos no cotidiano da nova geração e a ampla utilização do YouTube como um repositório de vídeos, o estudo direcionou seu foco para essa plataforma em específico. É importante salientar, no entanto, que embora esta pesquisa esteja centrada no Youtube, é possível estender essa análise a outros repositórios, levando em consideração as interações e opiniões dos usuários. Acredita-se que o modelo de aprendizagem de máquina desenvolvido possa ser aplicado com ajustes mínimos, desde que a abordagem seja devidamente adaptada.

O Youtube é uma plataforma amplamente utilizada e com crescente acervo, ele foi criado por Chad Hurley, Steve Chen e Jawed Karim, foi lançado no final de junho de 2005 com o intuito de compartilhar vídeos na *Internet*. Segundo Burgess e Green (2009), possui uma interface simples e intuitiva, onde os usuários podem, sem elevado conhecimento técnico, assistir e publicar vídeos.

A plataforma está voltada para vídeos e possui conteúdos sobre diversos assuntos, podendo encontrar filmes, músicas, *reviews*, entrevistas, entre outros. Segundo Heilbron (2019), a cada minuto são enviadas 400 horas de vídeos, e mais de 2 bilhões de usuários acessam a plataforma mensalmente (YOUTUBE, 2019b). Segundo Youtube (2017), 1,5 bilhões de pessoas conectadas pelo mundo acessam o Youtube mensalmente e 95% da população brasileira conectada acessa a plataforma ao menos uma vez por mês. No Brasil, 9 em cada 10 usuários do Youtube acessam a plataforma com o intuito de aprender algo novo e mais da metade acredita que o Youtube é o lugar onde encontra-se tudo o que se deseja ver e aprender (YOUTUBE, 2019a).

Sob o ponto de vista educacional, o Youtube pode ser considerado um repositório de OA, no sentido que armazena um vasto número de vídeos que podem ser utilizados em caráter educativo. De forma geral, um Objeto de Aprendizagem pode ser entendido como “qualquer entidade, digital ou não digital, que pode ser usada, reutilizada ou referenciada durante a aprendizagem

apoiada por tecnologia" (IEEE, 2002, p. 1). Podem ser animações, mapas, textos, imagens, vídeos e outros (WILEY, 2000). Neste contexto, o Youtube disponibiliza vídeos por meio do seu mecanismo de busca, i.e., a partir da busca por um conteúdo, a plataforma apresenta os vídeos adequados à expressão de busca utilizada.

Apesar de possuir uma interface bastante intuitiva e um extenso acervo de vídeos sobre diversos assuntos, alguns problemas relacionados ao mecanismo de busca do Youtube podem ser identificados. Uma questão importante no mecanismo de busca do Youtube está relacionada aos resultados retornados pela plataforma. Em muitos casos, o número de resultados retornados é muito grande, muitos deles de baixa qualidade (considerando aspectos educacionais) e/ou não relacionados à busca realizada. Nesse sentido, esse número considerável de vídeos incorretos devolvidos pela plataforma pode ser prejudicial para educadores e aprendizes que utilizam o Youtube como recurso para facilitar o ensino e a aprendizagem. Nesse sentido, uma tentativa natural de superar esse impasse é a classificação dos vídeos como educacionais ou não, com o objetivo de melhorar a qualidade do material educativo disponibilizado pela plataforma.

Em resposta aos desafios identificados, a presente investigação introduz uma solução inovadora que representa um avanço significativo em diversas áreas, notavelmente na intersecção entre Inteligência Artificial e Educação. A abordagem proposta utiliza técnicas de Mineração de Textos (MT), Processamento de Texto e AM para avaliar e processar os comentários dos usuários, aplicando-os efetivamente na identificação de materiais educativos pertinentes. Essa estratégia inédita de empregar comentários como um recurso estratégico enriquece a compreensão do conteúdo educativo, ajustando-se ao público-alvo. Combinando Processamento de Texto com técnicas de AM, a abordagem extrai conhecimentos significativos que são essenciais para gerar recomendações que não só atendem, mas também aprimoram o aspecto educacional. Esta inovação é uma contribuição valiosa para a área de Inteligência Artificial aplicada à Educação, oferecendo novas perspectivas para a melhoria de recursos educativos digitais e enriquecendo a experiência de aprendizado em ambientes com vasta diversidade de conteúdos.

Foram desenvolvidas duas variações metodológicas para modelar e recomendar Objetos de Aprendizagem (OAs): as variações "rígida" e "flexível". A variação rígida condensa a representação dos vídeos em uma única entrada, permitindo que os algoritmos de Aprendizado de Máquina (AM) classifiquem os vídeos de forma categórica como educacionais ou não educacionais. Em contraste, a variação flexível representa um vídeo pelo conjunto de seus comentários, classificando cada comentário individualmente. Posteriormente, uma fórmula é aplicada para calcular o "grau de certeza" de o vídeo ser educacional. Esta fórmula considera o número de vezes que o classificador identifica os comentários como educacional ou não, para estabelecer a certeza de sua classificação como um OA.

Os resultados obtidos com a abordagem inovadora, juntamente com as metodologias implementadas, demonstraram que é possível alcançar uma acurácia superior a 95% na recomendação de novos OAs. Além da aplicabilidade destas técnicas no setor educacional, os métodos e resultados desta pesquisa também mostraram sua utilidade em outros campos. Um exemplo é o estudo de Carvalho (2024), que emprega uma abordagem semelhante em conjunto com técnicas de Ciência de Redes para identificar notícias falsas (*Fake News*) no Twitter, evidenciando a versatilidade e o potencial expansivo dessas metodologias.

1.1 Motivação

A evolução tecnológica possibilitou maior interação e compartilhamento de dados e informações entre a sociedade, possibilitando que as pessoas possam interagir e disponibilizar conteúdos sobre os mais diversos tipos. Essa evolução possibilitou grande avanço na sociedade, pois permitiu que diversos materiais sejam encontrados, criados e compartilhados. Entretanto, no ambiente educacional, esse crescente acervo torna o processo de seleção de materiais cansativo e muitas vezes fracassado, principalmente quando utiliza-se uma plataforma não restrita a Objetos de Aprendizagem.

Apesar dos benefícios proporcionados pela evolução tecnológica, a ampla variedade de informações disponíveis podem sobrecarregar os educadores e aprendizes e dificultar a identificação e seleção de materiais relevantes para a aprendizagem. Essa tarefa torna-se ainda mais desafiadora quando se utiliza uma plataforma que não oferece restrições ou filtros específicos para OA.

Nesse contexto, é essencial desenvolver soluções que auxiliem os professores e os aprendizes a buscar conteúdos por esse grande acervo de materiais. Ferramentas inteligentes de busca devem ser implementadas, auxiliando na análise da qualidade, da relevância e da adequação dos materiais disponíveis aos propósitos educacionais. Auxiliando, dessa forma, na etapa de ensino-aprendizagem.

O crescente acervo de materiais disponibilizados no YouTube, sua ampla utilização como ferramenta de aprendizado e a crescente necessidade de filtrar e recomendar objetos de aprendizagem destacam a importância de estudos para aprimorar a qualidade do conteúdo oferecido nessa plataforma, visando uma melhor experiência de ensino-aprendizagem.

Com milhões de vídeos educacionais disponíveis, o YouTube se tornou uma fonte popular de informações e conhecimentos em diversos campos. No entanto, a falta de controle de qualidade, a dificuldade de encontrar conteúdos e a fragilidade do seu mecanismo de busca podem comprometer a eficácia dessa plataforma como uma ferramenta educacional confiável.

Diante desse cenário, Carvalho et al. (2020b) iniciaram estudos para investigar a utilização de comentários como possível métrica para a categorização de OA do Youtube. Em seus resultados preliminares, os autores, identificaram que os comentários podem ser significativos para a classificação de OA da plataforma. Entretanto, a pesquisa necessita de ser aprofundada para compreender como os usuários interagem com os materiais educacionais no YouTube. A análise dos padrões dos usuários, a avaliação da eficácia do conteúdo e a recomendação são aspectos cruciais a serem investigados.

Nesse contexto, é necessário desenvolver novas abordagens de filtragem de conteúdo mais sofisticadas, isso permitirá que os educadores e aprendizes encontrem materiais relevantes, confiáveis e alinhados aos objetivos de aprendizagem, melhorando assim a qualidade da experiência educacional no YouTube.

Dessa forma, investir em pesquisas e estudos sobre o YouTube como plataforma educacional desempenha um papel fundamental no avanço do campo da educação. Isso possibilita que educadores, aprendizes tenham acesso a materiais mais qualificados, confiáveis e relevantes. O constante estudo sobre a recomendação de OA não apenas nesse repositório é essencial para promover uma melhor experiência na escolha de materiais educacionais, assim como acompanhar

os avanços tecnológicos e atender às necessidades educacionais da sociedade atual.

É importante ressaltar que as pesquisas relacionadas ao comportamento e à interação dos usuários em relação aos Objetos de Aprendizagem podem ser adaptadas para outros repositórios. Isso significa que a abordagem centrada na opinião dos usuários pode ser ampliada e aplicada em repositórios que apresentam diferentes tipos de materiais educacionais, proporcionando um conhecimento mais aprofundado sobre o comportamento dos usuários que buscam esse tipo de material, assim como, possibilita percepções sobre futuras melhorias para auxiliar o aprendizado. O aprofundamento dos estudos, nessa área, propicia um avanço continuo no campo da educação.

1.2 Objetivos e Desafios da Pesquisa

Observa-se que a recomendação de Objetos de Aprendizagem em ambientes não estritamente educacionais necessitam de investigações, desde a busca até a recomendação desses materiais, além disso, plataformas como o Youtube, apresentam poucos estudos específicos sobre isso. Nesse aspecto, este trabalho visa responder as seguintes Questão de Pesquisa (QP):

- ❏ QP1: É possível recuperar e recomendar materiais educacionais de ambientes não estritamente educacionais mas que são utilizados com víes educacional?
- ❏ QP2: É possível utilizar a opinião dos usuários como métrica para classificar um Objeto de Aprendizagem?
- ❏ QP3: Algoritmos de Aprendizagem de Máquina podem ser utilizados, com elevada acurácia (superior a 90%), para identificar padrões e classificar Objetos de Aprendizagem?
- ❏ QP4: É possível desenvolver uma metodologia ajustável para a recomendação de OA, que possa ser ajustada conforme as necessidades específicas de cada temática?

Para responder as questões de pesquisa, este trabalho tem como objetivo geral a investigação das opiniões dos usuários em vídeos do Youtube com o intuito de identificar diferenças e reconhecer padrões para apresentar aos usuários Objetos de Aprendizagem que os auxiliem a obter novos conhecimentos. A partir das questões de pesquisa, podemos observar os seguintes objetivos específicos:

- ❏ Explorar a opinião dos usuários como meio para compreender a interação dos mesmos em materiais educacionais do Youtube.
- ❏ Analisar como algoritmos de Aprendizagem de Máquina se comportam para classificar Objetos de Aprendizagem através de comentários do Youtube.
- ❏ Desenvolver um modelo de algoritmo de Aprendizagem de Máquina que possa ser utilizado em conjunto com as opiniões de usuários para recomendar Objetos de Aprendizagem do Youtube.
- ❏ Desenvolver um Sistema de Recomendação que auxilie na escolha de materiais educacionais do Youtube.

A partir das Questões de Pesquisa e dos Objetivos estabelecidos, propomos a seguinte hipótese: "A análise de comentários aprimora o processo de recomendação de conteúdos educacionais". Esta hipótese sustenta-se na ideia de que os comentários deixados por usuários em vídeos são ricos em informações pertinentes que, quando eficazmente aproveitadas, têm o potencial de aprimorar a experiência educacional aumentando a pertinência das recomendações de materiais educativos. Além disso, essa hipótese encapsula os principais objetivos deste trabalho.

O escopo deste trabalho abrange um campo extenso e complexo, e enfrentou dificuldades significativas em todas as etapas, desde a concepção inicial até o desenvolvimento do Sistema de Recomendação LOIS. A primeira etapa, a definição do que constitui um vídeo educacional, é crucial e estabelece a base para todo o processo subsequente. A metodologia de coleta dos comentários, essencial para a análise e o desenvolvimento de um sistema de recomendação eficaz que funcione em tempo real, requer atenção especial devido às suas complexidades técnicas e operacionais.

A preparação dos comentários para análise envolve desafios consideráveis, principalmente devido à falta de estrutura e à ampla diversidade dos temas discutidos nos comentários. Esta etapa exige uma análise meticulosa dos comentários e a utilização de técnicas de processamento de textos para torná-los processáveis por algoritmos de aprendizado de máquina. A seleção e a configuração cuidadosa desses algoritmos são fundamentais para otimizar o desempenho e garantir o funcionamento desejado nas recomendações finais do sistema. Além disso, a escolha da linguagem de programação e das ferramentas de desenvolvimento para o projeto e para o sistema LOIS foi crucial, impactando diretamente na eficácia e na eficiência do sistema recomendado.

A abordagem introduzida representa um avanço significativo, especialmente na interseção da Inteligência Artificial com a educação. Utilizando técnicas avançadas de Mineração de Texto para analisar comentários e identificar as palavras-chave mais recorrentes, ela consegue destacar conteúdos educacionais pertinentes de maneira eficiente. O emprego estratégico dos vocábulos mais frequentes facilita uma análise mais detalhada e ajustada do material educativo, alinhando-o melhor às necessidades do público. Essa inovação traz uma valiosa contribuição para a aplicação de Inteligência Artificial na educação, propondo novos caminhos para a personalização do processo de aprendizagem. Adicionalmente, o Sistema de Recomendação, LOIS, se destaca por incorporar estas metodologias inovadoras, apoiando de forma eficaz o processo de ensino-aprendizagem.

1.3 Organização da Tese

Este documento está dividido da seguinte forma: o Capítulo 2 traz a fundamentação teórica da pesquisa, discutindo principalmente o que foi utilizado neste trabalho. O Capítulo 3 apresenta os trabalhos correlatos. O Capítulo 4 apresenta a abordagem proposta neste trabalho, levantando as tecnologias que foram empregadas para atingir os objetivos. O Capítulo 5 traz os experimentos realizados para responder as questões de pesquisa. Por fim, o Capítulo 6 apresenta as considerações finais deste trabalho.

Capítulo 2

Fundamentação Teórica

Esse capítulo aborda os principais conceitos utilizados e os trabalhos relacionados com esta pesquisa.

2.1 Mineração de Dados e Texto

Mineração de Dados pode ser definido como o processo de descoberta de padrões em dados (WITTEN; FRANK; HALL, 2011). Adicionalmente, pode-se dizer que é o processo de análise e exploração de grande quantidade de dados com o intuito de descobrir regras ou padrões significativos (BERRY; LINOFF, 2004).

Enquanto a Mineração de Dados identifica padrões em dados, a Mineração de Textos busca identificar padrões nos textos, i.e., é um processo que possibilita gerar conhecimento e extrair informações relevantes e não triviais de dados textuais. É um campo multidisciplinar que envolve Aprendizado de Máquina, Mineração de Dados, Recuperação da Informação, Processamento de Texto, entre outros (VIJAYARANI; JANANI et al., 2016; JUSOH; ALFAWAREH, 2012). Tal área de estudo pode ser definida como o processo de análise e extração de informações úteis de textos para propósitos específicos (WITTEN; FRANK; HALL, 2011).

A Mineração de Textos trabalha basicamente em três etapas, a saber: pré-processamento de dados; aplicação de técnicas de mineração; e análise do texto (VIJAYARANI et al., 2015; SUKANYA; BIRUNTHA, 2012). Tais etapas são brevemente descritas a seguir.

A etapa de pré-processamento é fundamental no tratamento do texto antes da análise e aplicação das técnicas de mineração. Durante essa etapa, é realizada a padronização do texto, que inclui a conversão de caracteres para minúsculas e a correção de erros ortográficos. Além disso, são removidas palavras irrelevantes, como stopwords, caracteres especiais e numéricos. Essa etapa também envolve a aglomeração de termos similares (HICKMAN et al., 2022).

O objetivo principal do pré-processamento é melhorar a qualidade dos dados que serão analisados, eliminando ruídos e inconsistências. Ao limpar o texto de erros ortográficos, caracteres especiais e outros elementos indesejados, é possível obter uma base de dados mais confiável e pronta para a análise (HACOHEN-KERNER; MILLER; YIGAL, 2020).

A etapa de aplicação de técnicas de mineração consiste em utilizar algoritmos para proces-

sar os textos. Neste sentido, podem ser utilizados algoritmos para Visualização, Sumarização, Extração da Informação, Clusterização, e Categorização (SUKANYA; BIRUNTHA, 2012).

- A Visualização é uma maneira de se melhorar e simplificar a descoberta de informações relevantes. Para isso, organizam-se as informações textuais em uma hierarquia visual que possibilita a interação do usuário com o documento, podendo, este, dimensionar, ampliar e buscar informações. A utilização de técnicas de visualização fornece informações melhores e mais rápidas, possibilitando que os usuários as diferenciem por meio de cores, relacionamentos, distância, entre outros (GAIKWAD; CHAUGULE; PATIL, 2014; SUKANYA; BIRUNTHA, 2012)

- A Sumarização é basicamente a produção de resumos a partir de um documento, realizado principalmente devido a grande quantidade de textos presentes. O objetivo é reduzir a quantidade de textos sem afetar o significado e os pontos principais. A sumarização pode ser realizada a partir de um único ou um grupo de documentos, caso seja através de um conjunto de documentos, esses serão substituídos por um resumo (SUKANYA; BIRUNTHA, 2012).

- A Extração da Informação é o processo de exploração de texto buscando identificar informações relevantes voltadas para a identificação de algum interesse. O processo inclui a extração de relações, entidades, e eventos (HOBBS; RILOFF, 2010). A identificação é feita por meio de um processo denominado correspondência de padrões que é realizado procurando-se sequencias predefinidas no texto (VIJAYARANI et al., 2015; SUKANYA; BIRUNTHA, 2012).

- A Clusterização é uma técnica utilizada para agrupar documentos semelhantes. Seu objetivo é identificar estruturas semelhantes nas informações e organizá-las em subgrupos significativos. É um processo não supervisionado, ou seja, nenhum dado de saída é fornecido, e os objetos são classificados em grupos semelhantes chamados clusteres. O objetivo é agrupar, sem conhecimento prévio, diversos dados não rotulados em clusteres significativos. Todos os rótulos associados são obtidos por meio dos dados fornecidos (SUKANYA; BIRUNTHA, 2012; DANG; AHMAD, 2014).

- A Classificação é uma técnica para categorizar documentos em classes definidas. Diferentemente da Clusterização, a Classificação é supervisionada, i.e., as as classes de cada documento já são conhecidas *a priori*. O objetivo é treinar o classificador em uma base de dados de treinamento e então os exemplos desconhecidos serão categorizados automaticamente por meio do "conhecimento" obtido na base de dados de treino (SUKANYA; BIRUNTHA, 2012; DANG; AHMAD, 2014).

Além das técnicas mencionadas, a Tokenização e a Vetorização são fundamentais para a preparação de dados em análises de texto. A Tokenização, em particular, é uma técnica essencial no processamento de informações, utilizada extensivamente em tarefas que envolvem a extração de conhecimento de textos. Ela consiste em dividir um texto em unidades menores chamadas tokens, que podem ser palavras, frases ou outros elementos significativos (FRIEDMAN, 2023).

Esses tokens são cruciais para a preparação de dados em diversas aplicações. Por exemplo, ao segmentar o texto em tokens, torna-se mais fácil identificar e extrair entidades nomeadas (como nomes de pessoas, locais, etc.) e outras informações específicas. A tokenização também é frequentemente usada para analisar a frequência dos tokens, permitindo percepções sobre os termos mais utilizados e suas implicações no contexto analisado.

A vetorização é necessária para converter texto, uma forma de dado qualitativo, em representações numéricas compreensíveis por computadores. Esse processo de transformação é essencial para que os algoritmos de Aprendizado de Máquina possam processar e aprender a partir de dados textuais. Uma das técnicas mais utilizadas para a vetorização é o *Bag of Words* (Bag of Words (BoW)).

O *Bag of Words* (BoW) é uma abordagem simples e eficaz na mineração de textos para converter linguagem natural em vetores numéricos (VM; R, 2019). No modelo BoW, cada documento é representado por um vetor, onde cada dimensão corresponde a uma palavra única presente no conjunto de dados analisados. Cada componente desse vetor indica a presença ou a frequência da respectiva palavra no documento, contabilizando cada ocorrência de forma exata (ZHAO; MAO, 2017).

A representação vetorial baseada no BoW geralmente se fundamenta nos tokens mais frequentes em todo o corpus (conjunto textual fornecido para análise). Essa técnica permite uma análise quantitativa do texto, embora desconsidere a ordem das palavras e seu contexto sintático ou semântico.

Por fim, a etapa de análise do texto consiste em analisar e identificar as informações relevantes que foram geradas após a etapa de aplicação de técnicas de mineração. Obtêm-se, após esta última etapa, informações e conhecimentos relevantes sobre o texto processado (SUKANYA; BIRUNTHA, 2012).

2.2 Aprendizado de Máquina

Aprendizado de Máquina (AM, ou ML do inglês *Machine Learning*) pode ser definido como o campo de estudo se preocupa em como fornecer ao computador a habilidade de aprender sem ser explicitamente programado (WIEDERHOLD; MCCARTHY, 1992). É o ramo da Inteligência Artificial que utiliza técnicas e algoritmos com o intuito de reconhecer padrões ou de melhorar seu desempenho por meio de sua experiência (MITCHELL, 1997; RUSSELL; NORVIG, 2010). De forma geral, existem três formas de aquisição de conhecimento pelas técnicas de Aprendizado de Máquina, a saber: Aprendizado Supervisionado, Aprendizado não Supervisionado e Aprendizado por Reforço (RUSSELL; NORVIG, 2010).

No Aprendizado Supervisionado, os dados são enviados juntamente com os rótulos, as classes, ou seja, o algoritmo já possui informações prévias sobre como os dados são classificados. Para esse tipo de aprendizado, são fornecidos aos algoritmos os dados de “treinamento” e “teste”. Desta forma, é necessário que os dados sejam divididos nessas duas bases de dados distintas para o classificador “aprender” e depois “validar” seus resultados, ou seja, predizer a qual classe uma nova entrada de dados pertence.

Após a classificação dos dados, é necessário verificar a capacidade do classificador em reconhecer as classes apresentadas. Um dos métodos utilizados e recomendados para validar a predição dos classificadores no Aprendizado Supervisionado é o método de validação cruzada de k-folds (*k-fold cross-validation*). Tal método consiste basicamente em dividir a base de dados em k partes, utilizando-se k-1 partes para a etapa de treinamento e 1 parte para a etapa de teste, repetindo-se este processo k vezes, e modificando-se os conjuntos de treinamento e teste a cada vez. De forma geral utiliza-se k = 10, mas outros valores para k também podem ser adotados (BERRAR, 2019; MITCHELL, 1997).

No Aprendizado não Supervisionado, os algoritmos não possuem informações sobre as classes dos dados, ou seja, eles não possuem informação prévia que os influencie a predizer os novos dados. Neste caso, o próprio algoritmo é, portanto, o responsável por analisar os dados com o intuito de separá-los de acordo com a similaridade e os padrões identificados, agrupando-os em classes ou clusteres distintos.

Por fim, no Aprendizado por Reforço, os algoritmos aprendem por meio de reforços positivos ou negativos. Caso o algoritmo forneça uma resposta "correta", recebe uma recompensa (reforço positivo), e caso forneça uma resposta "incorreta", recebe uma punição (reforço negativo) (RUSSELL; NORVIG, 2010).

As técnicas de Aprendizado de Máquina potencializaram diversos campos, entre eles pode-se citar a Mineração de Dados e Mineração de Textos. De forma geral, Mineração de Dados é um campo multidisciplinar que envolve a Visualização de Dados, Inteligência Artificial, Aprendizado de Máquina, Reconhecimento de Padrões, Banco de Dados, Computação de Alto Desempenho, Aquisição de Conhecimento, Recuperação de Informação e Teoria da Informação (SUMATHI; SIVANANDAM, 2006).

Algoritmos de Aprendizagem de máquina está sendo cada vez mais estudados e analisados no campo da inteligência artificial. Esses algoritmos são capazes de aprender, melhorar e identificar padrões a partir dos dados fornecidos, permitindo que auxiliar e ajudar a tomar decisões complexas e sensíveis se foram explicitamente programados. Com o crescente volume de dados e o avanço do poder computacional, os algoritmos de aprendizagem de máquina tem se apresentado essenciais para auxiliar à solucionar problemas complexos em várias áreas do conhecimento.

Uma maneira de classificar os variados algoritmos de Aprendizado de Máquina é através do método utilizado para representar o conhecimento. Assim, essa representação pode ocorrer através de elementos como Árvores de Decisão, Regras, Agrupamentos (Clusteres), Neurônios Artificiais e Vetores de Suporte.

A categorização dos algoritmos de Aprendizado de Máquina é realizada com base no mecanismo utilizado para representar o funcionamento do algoritmo, assim como, o conhecimento adquirido. Entre as técnicas de representação mais empregadas, encontram-se as Árvores de Decisão, que organizam as decisões e suas possíveis consequências em uma estrutura de árvore; as Regras, que definem condições específicas para a tomada de decisões; os Clusters, que agrupam dados similares para identificar padrões; os Neurônios Artificiais, inspirados na estrutura neural do cérebro humano para processar informações; e as Máquinas de Vetores de Suporte, que criam um hiperplano ótimo para distinguir entre categorias de dados, maximizando a margem entre diferentes classes.

2.2.1 Árvores de Decisão

Os algoritmos baseados em Árvores de Decisão são umas das formas mais simples e mais bem-sucedidas de se classificar dados. Essa abordagem baseia-se na construção de uma estrutura de árvore que representa um conjunto de regras de decisão hierárquicas. As árvores de decisão são especialmente apreciadas devido à sua interpretabilidade, facilidade de compreensão e capacidade de lidar com dados tanto numéricos quanto categóricos (RUSSELL; NORVIG, 2010).

Uma Árvore representa uma função que toma como entrada um conjunto de atributos e retorna uma "decisão". Sua decisão é alcançada executando uma sequência de testes (RUSSELL; NORVIG, 2010). A árvore é composta por um conjunto de nós, onde cada nó representa uma característica ou atributo do conjunto de dados. A partir do nó raiz, ocorrem divisões sucessivas e cada nó interno da árvore corresponde a um teste do valor de um dos atributos de entrada. As divisões ocorrem até que se alcancem os nós folha, que representam as classes ou as decisões finais para uma determinada entrada (KESAVARAJ; SUKUMARAN, 2013; ALLAHYARI et al., 2017).

A construção de uma Árvore de Decisão é um processo iterativo que começa com o nó raiz, representando o atributo com o maior ganho de informação. A árvore é dividida em subárvores de acordo com os valores possíveis desse atributo. A cada nó gerado, o atributo com o maior ganho de informação é utilizado. Os critérios, mais conhecidos, de seleção do atributo com o maior ganho de informação são conhecidos como entropia e índice Gini. Essas medidas avaliam a incerteza ou a impureza dos dados e auxiliam na escolha do melhor atributo para a divisão da árvore (SCIKIT-LEARN DEVELOPERS, 2023).

Entropia é uma medida utilizada na teoria da informação para calcular o ganho de informação em um conjunto de dados. Essa medida quantifica a falta de homogeneidade dos dados em relação à sua classificação. A entropia é máxima, igual a 1, quando o conjunto de dados é heterogêneo, ou seja, quando apresenta uma distribuição equilibrada entre as diferentes classes ou categorias. O cálculo da entropia é dado pela fórmula 1 (MITCHELL, 1997).

$$Entropia(S) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \tag{1}$$

Onde:

- Entropia(S) é a entropia do conjunto de dados S;
- p(i) é a proporção de instâncias do conjunto de dados que pertencem à classe i;
- n é o número total de classes no conjunto de dados.

Após o cálculo da entropia é necessário calcular qual atributo apresenta o maior ganho de informação. O ganho de informação é uma medida que determina o quanto um determinado atributo contribui para a redução da entropia do conjunto de dados. Quanto maior o ganho de informação, mais relevante é o atributo para a construção da árvore de decisão (RUSSELL; NORVIG, 2010; MITCHELL, 1997).

O ganho de informação é calculado a partir da entropia do conjunto de dados original (antes da divisão) e da entropia ponderada dos conjuntos de dados resultantes da divisão. A fórmula para o cálculo do ganho de informação é dada pela formula 2 (RUSSELL; NORVIG, 2010; MITCHELL, 1997).

$$Ganho(A) = Entropia(S) - \sum_{v \in Val(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{2}$$

Onde:

- Ganho(A) é o ganho de informação do atributo A;
- Entropia(S) é a entropia do conjunto de dados original S;
- Val(A) é o conjunto de valores possíveis do atributo A;
- S_v é o subconjunto de dados correspondente ao valor v do atributo A;
- |S| é o número total de instâncias no conjunto de dados S.

O atributo com o maior ganho de informação é escolhido como o próximo nó da árvore de decisão, e o processo de construção continua recursivamente para os subconjuntos resultantes da divisão (RUSSELL; NORVIG, 2010; MITCHELL, 1997).

Outra medida comumente usada para construir árvores de decisão é o índice Gini. O índice Gini mede a impureza dos dados, sendo uma medida de quão frequentemente uma instância escolhida aleatoriamente do conjunto seria incorretamente classificada se fosse rotulada aleatoriamente de acordo com a distribuição das classes no subconjunto. O cálculo do índice Gini é análogo ao cálculo da entropia, mas utilizando o índice Gini no lugar da função de entropia (BREIMAN et al., 1984).

O cálculo do índice Gini para um conjunto de dados S é dado pela fórmula 3:

$$Gini(S) = 1 - \sum_{i=1}^{n} (p(i))^2 \tag{3}$$

Onde:

- Gini(S) é o índice Gini do conjunto de dados S;
- p(i) é a proporção de instâncias do conjunto de dados que pertencem à classe i;
- n é o número total de classes no conjunto de dados.

Para cada atributo A, o índice Gini ponderado é calculado utilizando a fórmula 4:

$$Gini\_Ponderado(A) = \sum_{v \in \mathrm{Val}(A)} \frac{|S_v|}{|S|} \cdot Gini(S_v) \tag{4}$$

Onde:

- Gini_Ponderado(A) é o índice Gini ponderado do atributo A;

- Val(A) é o conjunto de valores possíveis do atributo A;
- S_v é o subconjunto de dados correspondente ao valor v do atributo A;
- |S| é o número total de instâncias no conjunto de dados S.

O atributo com o menor índice Gini ponderado é escolhido como o próximo nó da árvore de decisão, pois representa a divisão que resulta em maior pureza (menor impureza) dos subconjuntos (BREIMAN et al., 1984).

Ambas as medidas, entropia e índice Gini, são utilizadas na construção de árvores de decisão e têm como objetivo encontrar a melhor divisão dos dados, priorizando atributos que reduzam a incerteza ou a impureza das classes nos subconjuntos resultantes. A escolha entre entropia e índice Gini depende do problema em questão e das preferências do usuário (BREIMAN et al., 1984).

O J48 serve como uma representação do algoritmo de Árvore de Decisão, sendo uma implementação específica para o Weka do renomado algoritmo C4.5 de árvores de decisão. Desenvolvido por Ross Quinlan, Quinlan (1993), o C4.5 destaca-se como um dos métodos mais reconhecidos e utilizados amplamente na construção de árvores de decisão.

O Random Forest (Floresta Aleatória) é um modelo, amplamente utilizado, baseado em árvores de decisão. Ele pertence à categoria de algoritmos baseados em ensemble, que consiste em combinar vários modelos individuais para obter um resultado final mais robusto e preciso. O Random Forest constrói diversas árvores de decisão e a classificação de um exemplo é dada pela votação deste conjunto de classificadores (BREIMAN, 2001).

No Random Forest, cada árvore de decisão é treinada em uma amostra aleatória dos dados de treinamento, selecionada através da técnica "*bagging*" (*bootstrap aggregating* ou agregação *bootstrap*). Essa técnica envolve a criação de várias amostras de treinamento, chamadas de subconjuntos *bootstrap*, a partir do conjunto de dados de treinamento original. O processo de criação desses subconjuntos é realizado selecionando aleatoriamente amostras do conjunto de treinamento com reposição, ou seja, cada amostra selecionada é copiada do conjunto de treinamento original e é mantida no conjunto antes da próxima seleção. Como resultado, um subconjunto *bootstrap* pode conter múltiplas cópias de algumas amostras e pode não incluir outras amostras originais (BREIMAN, 2001).

O objetivo do *bagging* é introduzir variação e diversidade nas amostras de treinamento por meio do processo de amostragem. Como resultado, cada árvore é gerada a partir de uma amostra dos dados originais, o que permite que as árvores sejam diferentes entre si. Essa técnica possibilita compensar as fraquezas individuais de cada árvore e melhorar a capacidade do modelo de generalizar para novos dados (BREIMAN, 2001; BREIMAN, 1996).

Durante a etapa de previsão, cada árvore do *Random Forest* faz uma previsão individual para uma determinada amostra dos dados de teste. No caso da classificação, a previsão final é determinada pela votação majoritária das previsões de todas as árvores. No caso da regressão, a previsão final é obtida pela média das previsões das árvores (BREIMAN, 2001).

A principal vantagem do *Random Forest* é a capacidade de lidar com uma ampla gama de dados, incluindo variáveis categóricas e numéricas. Além disso, ele permite avaliar a importância relativa das características utilizadas na construção das árvores.

É importante destacar que, embora o *Random Forest* utilize técnicas semelhantes à árvore de decisão, as árvores construídas com partes aleatórias do conjunto de dados podem ser completamente diferentes entre si, resultando em resultados mais robustos. Esse modelo é especialmente útil para evitar o *overfitting*, quando o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados.

Os algoritmos baseados em árvores de decisão são amplamente utilizados devido à interpretabilidade das árvores de decisão resultantes e sua capacidade de lidar com diferentes tipos de dados, como atributos categóricos e numéricos. Sendo possível serem visualizadas de forma gráfica, tornando mais fácil compreender como as decisões estão sendo tomadas. Além disso, é percorrer a árvore, extraindo um conjunto de regras que descrevem as condições para cada decisão tomada. Essas regras podem ser úteis para interpretar o modelo ou para utilizar em sistemas especialistas.

2.2.2 Regras

Os algoritmos que operam com base em regras constituem uma estratégia amplamente adotada no domínio do Aprendizado de Máquina, representando uma alternativa notável às árvores de decisão. Essa abordagem tem ganhado destaque pela sua proficiência em resolver questões de classificação e regressão. É possível derivar regras diretamente de árvores de decisão, o que proporciona um meio eficaz e interpretável de expressar o conhecimento obtido (WITTEN; FRANK; HALL, 2011).

As regras são compostas por um antecedente (ou condição) e um consequente (ou ação). As regras iniciam com um conjunto de verificações (testes), estabelecendo critérios que devem ser cumpridos para a aplicação da regra. O consequente, então, designa a ação a ser tomada, seja ela a atribuição de uma classe específica ou uma distribuição entre várias classes. Em essência, o mecanismo das regras se baseia no princípio condicional: SE a condição X for satisfeita, ENTÃO atribui-se a sequência de características a uma determinada classe (WITTEN; FRANK; HALL, 2011).

Normalmente, os antecedentes das regras são construídos utilizando operadores lógicos “E” (AND), onde todos os testes devem ser bem-sucedidos para que a regra seja ativada. No entanto, em algumas formulações de regras, os antecedentes podem ser expressões lógicas mais gerais, permitindo uma maior flexibilidade. É comum pensar nas regras individuais como estando efetivamente ligadas por um operador lógico “OU” (OR): se qualquer uma das regras se aplicar, a classe (ou distribuição de probabilidade) especificada em sua conclusão será aplicada à instância (WITTEN; FRANK; HALL, 2011).

Uma das principais vantagens dos algoritmos baseados em regras é a sua interpretação intuitiva. Cada regra representa um conhecimento independente. Novas regras podem ser adicionadas a um conjunto existente sem perturbar as regras já presentes, diferentemente das árvores de decisão, que geralmente requerem uma remodelagem completa. No entanto, a ordem de execução das regras é crítica para o resultado final (WITTEN; FRANK; HALL, 2011).

Embora as regras sejam simples e intuitivas, é importante considerar como elas são executadas em conjunto. Se as regras forem interpretadas em ordem, como uma lista de decisões

sequenciais, algumas regras podem estar corretas individualmente, mas podem levar a conclusões incorretas quando consideradas isoladamente e fora de contexto. Por outro lado, se a ordem de interpretação for considerada irrelevante, podem surgir problemas quando diferentes regras levam a conclusões diferentes para a mesma instância (WITTEN; FRANK; HALL, 2011).

O algoritmo JRIP, também conhecido como RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*), é um dos algoritmos baseados em regras amplamente utilizados. Ele foi proposto por (COHEN, 1995). Ele é um algoritmo de aprendizagem de máquina que gera regras de classificação com base em exemplos de uma classe específica de cada vez. Ele segue uma abordagem incremental, tratando uma classe de cada vez e gerando regras para essa classe antes de passar para as demais classes. Esse processo é repetido até que todas as classes tenham sido tratadas (COHEN, 1995).

Essa abordagem incremental do JRip permite uma maior especialização das regras para cada classe, já que o algoritmo foca em aprender as características distintivas de cada uma delas separadamente. Ao lidar com uma classe de cada vez, o algoritmo tem a oportunidade de explorar a relação entre os atributos e a classe em questão de forma mais precisa (COHEN, 1995).

Durante a geração das regras, o JRip utiliza o método de poda (*pruning*) incremental para reduzir o erro de classificação. Isso significa que, à medida que novas regras são adicionadas, o algoritmo realiza uma análise de cada regra existente e, se necessário, pode podá-la para melhorar sua generalização e evitar o overfitting (COHEN, 1995).

O JRip é especialmente útil quando se lida com conjuntos de dados complexos, nos quais diferentes classes possuem características distintas e regras específicas. A abordagem incremental do algoritmo permite que ele se adapte a essas complexidades e aprenda de forma mais precisa as regras de classificação para cada classe individualmente.

Em resumo, o algoritmo JRIP é um algoritmo poderoso, capaz de gerar regras de classificação especializadas para cada classe em conjuntos de dados complexos. Sua abordagem incremental e o uso do método de poda incremental contribuem para sua eficácia na obtenção de resultados precisos e generalizáveis.

O PART (Partial Decision Trees) é um algoritmo de aprendizagem de máquina utilizado para construir árvores de decisão parciais. A cada iteração ele constrói uma árvore parcial e transforma o caminho até o melhor nó em uma regra de classificação (FRANK; WITTEN, 1998). Ao contrário das árvores de decisão tradicionais, que visam criar uma árvore completa, o PART permite a construção de árvores parciais, ou seja, a árvore gerada não expande todos os nós possíveis.

O algoritmo PART é útil, principalmente, para lidar com grandes e complexos conjuntos de dados, nos quais construir uma árvore completa apresentaria um elevado custo computacional ou que poderia levar a uma árvore muito complexa e propensa a overfitting. A construção das árvores de decisão parciais no PART é feita seguindo uma abordagem de divisão recursiva, ela usa um critério de divisão baseado no ganho de informação relativo, em vez do ganho de informação absoluto (FRANK; WITTEN, 1998).

Ao construir uma árvore de decisão parcial, o PART realiza a divisão do conjunto de dados com base no atributo que fornece o maior ganho de informação relativo. Isso significa que o

PART prioriza a divisão que reduz a incerteza dos dados em relação à classificação, levando em consideração a incerteza já existente no nó pai (FRANK; WITTEN, 1998).

Uma vez que a divisão é feita, o PART continua a construção da árvore apenas para os ramos que fornecem um ganho de informação relativo acima de um determinado limite pré-definido. Isso evita que a árvore cresça demais e permite que ela se torne parcial, focando apenas nas divisões mais relevantes e informativas (FRANK; WITTEN, 1998).

Uma característica interessante do PART é que ele permite a presença de regras incompletas nos nós folha da árvore. Isso significa que nem todas as instâncias de treinamento são necessariamente classificadas em um nó folha, permitindo a expressão de incerteza ou falta de informações em certos casos.

Em resumo, o PART é um algoritmo de aprendizagem de máquina que constrói árvores de decisão parciais, focando nas divisões mais relevantes com base no ganho de informação relativo. É uma abordagem útil para lidar com conjuntos de dados grandes e complexos, permitindo a construção de modelos mais simples e eficientes.

2.2.3 Clusteres

Clustering (agrupamento) é uma técnica amplamente utilizada no campo da aprendizagem de máquina que visa agrupar dados não rotulados em conjuntos homogêneos chamados de clusters. Essa abordagem permite a descoberta de padrões intrínsecos nos dados, identificação de grupos similares e organização dos objetos de acordo com suas características comuns. Os algoritmos de clustering são do tipo não supervisionado e pertencem a uma classe de métodos indutivos que agrupam os dados de acordo com suas características similares (HICKMAN et al., 2022).

No processo de clustering, o conjunto de dados é dividido em um conjunto pré-definido de clusters e, de forma iterativa, busca-se a melhor forma de realizar essa divisão com base nas características dos dados. O objetivo é que os objetos dentro de cada cluster sejam o mais similares possível entre si, enquanto os objetos de clusters diferentes sejam o mais dissimilares possível. Essa tarefa de agrupamento pode ser desafiadora, especialmente quando os dados possuem alta dimensionalidade ou quando a estrutura dos clusters não é facilmente distinguível. A Figura 1 apresenta como os dados fornecidos são separados em 3 clusteres diferentes.

O algoritmo GenClust++ é uma técnica de clustering que combina os princípios do K-Means e um algoritmo genético, utilizando um novo arranjo de operadores genéticos para realizar a clusterização. Esta abordagem híbrida possui a capacidade de identificar, como soluções iniciais, um conjunto de cromossomos de alta qualidade, permitindo uma melhor exploração do espaço de busca (ISLAM et al., 2018).

A principal inspiração do GenClust++ vem do K-Means, que é um dos algoritmos de clustering mais conhecidos. O K-Means busca minimizar a soma dos quadrados das distâncias entre os pontos de dados e os centroides dos clusters. No entanto, o K-Means pode ser sensível à inicialização dos centroides, o que pode levar a resultados subótimos (ISLAM et al., 2018).

Para superar esse desafio, o GenClust++ utiliza a meta-heurística dos algoritmos genéticos. Os algoritmos genéticos são técnicas de busca e otimização inspiradas na evolução biológica.

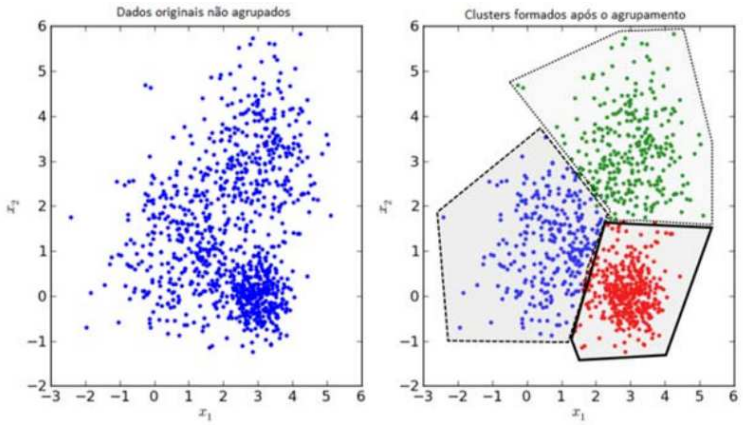


Figura 1 – Separação dos dados em clusteres.

Fonte: (RAMOS et al., 2016)

Eles operam em uma população de soluções candidatas (indivíduos), que evoluem ao longo de várias gerações (ISLAM et al., 2018).

No GenClust++, cada indivíduo representa uma possível configuração de clusters, codificada como um cromossomo. Os genes do cromossomo representam os centroides dos clusters. O algoritmo evolui esses indivíduos por meio de seleção, crossover e mutação, buscando otimizar a qualidade dos clusters encontrados (ISLAM et al., 2018).

A combinação do K-Means com algoritmos genéticos no GenClust++ permite uma abordagem mais robusta para o problema de clustering. Os operadores genéticos adaptados para o contexto de clustering permitem a exploração eficiente do espaço de soluções e a identificação de configurações de clusters de alta qualidade.

2.2.4 Neurônios Artificiais

Os algoritmos baseados em neurônios artificiais são modelos inspirados no comportamento cérebro humano e no funcionamento dos neurônios. Os neurônios são células especializadas do sistema nervoso que desempenham um papel fundamental na transmissão e processamento de informações no cérebro. Neurônios são, na verdade, elementos de processamento muito simples (COPPIN, 2015). Cada neurônio é composto por um soma, que é o corpo do neurônio, um axônio e vários dendritos (COPPIN, 2015; KOVÁCS, 2002). A figura 2 apresenta um neurônio biológico e o fluxo das informações.

Os neurônios são, basicamente, um dispositivo computacional elementar, do sistema biológico, com várias entradas e uma saída, ou seja, os neurônios são capazes de receber vários estímulos e fornecer uma saída (KOVÁCS, 2002). Os dentritos são responsáveis pelo recebimento e condução das informações vindas de outros neurônios ou do meio externo onde podem estar em contato. Um potencial de ativação é produzido e indicará se os impulsos nervosos serão ou não conduzidos para outros neurônios por meio do axônio. O corpo celular é encarregado de coletar e processar as informações enviadas pelos dentritos. A terminação dos axônios é ramificada e recebe a denominação de terminações sinápticas, que se conectam, mesmo que sem

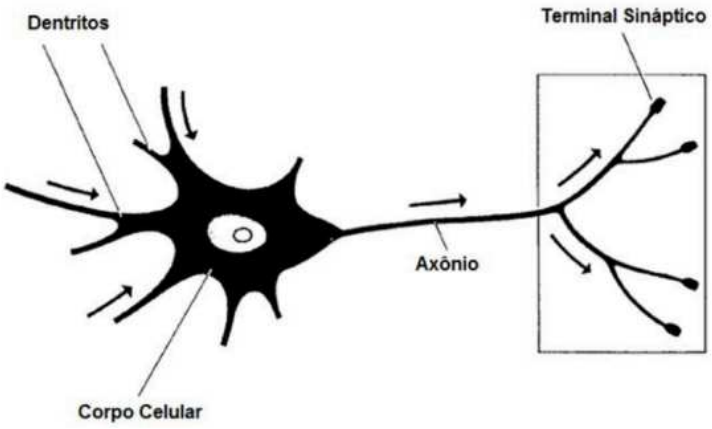


Figura 2 – Neurônio biológico

Fonte: Adaptado de (BIANCHINI, 2001).

contato físico com os dentritos de outros neurônios. Esse contato que se encarrega de transferir os impulsos nervosos de um neurônio para outro é chamado de sinapse.

A sinapses é onde ocorre a transmissão de sinais elétricos e químicos. As sinapses são unidades estruturais e funcionais elementares que medeiam as interações entre os neurônios. O tipo mais comum de sinapse é a sinapse quimica, que opera da seguinte forma: um processo pré-sinaptico libera uma substância transmissora que se difunde através da junção sinaptica entre neurônios e então age sobre um processo pós-sinaptico. Assim, uma sinapse converte um sinal elétrico pré-sináptico em um sinal químico e então de volta em um sinal elétrico pós-sináptico (SHEPHERD; KOCH, 1990).

O surgimento das Redes Neurais Artificiais (RNA) teve início com a proposta de Warren McCulloch e Walter Pitts em 1943, por meio do modelo matemático do neurônio biológico, conhecido como neurônio MCP (McCulloch-Pitts) (MCCULLOCH; PITTS, 1943). O modelo MCP descreve um conjunto de n entradas, multiplicadas por pesos sinápticos correspondentes, cujos resultados são somados e comparados a um limiar para determinar a ativação do neurônio. Esse modelo estabeleceu as bases para o desenvolvimento das RNAs.

Posteriormente, Frank Rosenblatt, em 1958 Rosenblatt (1958), realizou avanços significativos ao aprimorar o modelo MCP e introduzir o Perceptron. O Perceptron tornou-se um componente fundamental nas redes neurais modernas. Ele consiste em uma camada de neurônios em que cada neurônio é um MCP aprimorado, com a adição de um processo de treinamento baseado no algoritmo de aprendizado supervisionado. Essa abordagem permitiu que o Perceptron aprendesse a realizar tarefas de classificação e reconhecimento de padrões.

De maneira computacional as redes neurais podem são definidas por Haykin (2001) como: “um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso”. Ela é semelhante ao cérebro em dois pontos:

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.

2. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

A figura 3 apresenta um modelo de neurônio artificial segundo (HAYKIN, 2001).

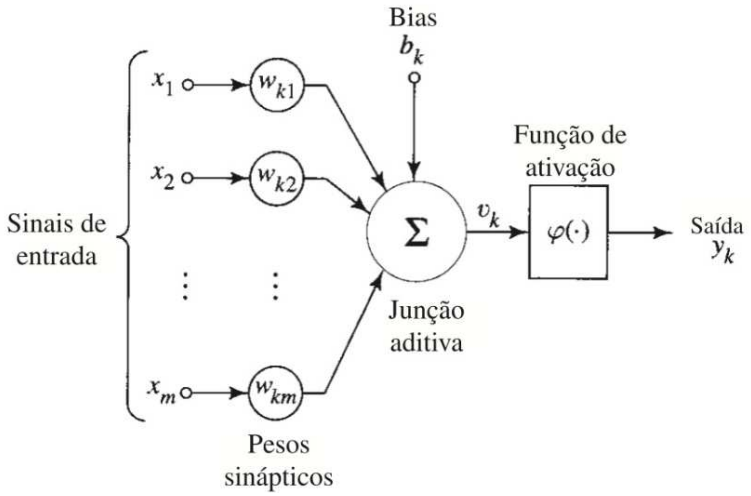


Figura 3 – Neurônio artificial

Fonte: (HAYKIN, 2001).

A função de ativação desempenha um papel importante ao restringir a amplitude da saída. Essa função é comumente denominada função restritiva, pois limita o intervalo de amplitude do sinal de saída a um valor finito. Tipicamente, o intervalo normalizado da amplitude da saída de um neurônio é representado pelo intervalo unitário fechado [0, 1] ou, alternativamente, (-1, 1]. Essa restrição assegura que a saída do neurônio esteja dentro desse intervalo, independentemente dos valores dos sinais de entrada e dos pesos sinápticos. A escolha do intervalo de saída depende das necessidades específicas da aplicação da rede neural, sendo [0, 1] comum em problemas de classificação binária e (-1, 1] frequentemente utilizado em problemas de regressão ou quando é necessário representar valores contínuos simetricamente em torno de zero (HAYKIN, 2001).

O bias b_k (viés) em um neurônio artificial tem o efeito de adicionar um termo constante à entrada líquida da função de ativação. Esse termo pode aumentar ou diminuir a entrada líquida, dependendo se o bias é positivo ou negativo, respectivamente. Em outras palavras, o bias afeta o ponto de partida da função de ativação, deslocando-a para cima ou para baixo ao longo do eixo vertical. Isso permite ao neurônio realizar um ajuste adicional na saída, além da influência dos pesos sinápticos, permitindo maior flexibilidade na modelagem de relações não-lineares entre os sinais de entrada e a saída do neurônio (HAYKIN, 2001).

Podemos definir um neurônio u_k através das fórmulas 5 e 6

$$u_k = \sum_{j=1}^{m} w_{kj} \cdot x_j \tag{5}$$

e

$$y_k = \varphi(u_k + b_k) \tag{6}$$

Onde:

- X_m são sinais de entrada da rede;
- w_{km} são pesos ou pesos sinápticos dos sinais de entrada;
- b_k é o termo bias;
- $\sum$ é a função de somatório;
- $\varphi(.)$ é função de ativação;
- y_k é a saída, em direção a outros neurônios.

O bias b_k tem o efeito de aplicar uma transformação afim à saída u_k do combinador linear no modelo do neurônio artificial na figura 3 e é dada pela fórmula 7

$$v_k = u_k + b_k \tag{7}$$

Para facilitar a compreensão, a figura 4 ilustra um neurônio artificial de forma simplificada. Nesse modelo, o neurônio recebe três entradas (X, Y e Z). Para cada entrada, são aplicados pesos sinápticos correspondentes (W_x, W_y e W_Z). Em seguida, o valor do bias é somado à combinação linear das entradas ponderadas. Finalmente, é aplicada a função de ativação ($f(a)$) ao resultado para obter a saída do neurônio. Essa saída pode ser transmitida a outros neurônios na rede. Essa representação visual simplificada nos ajuda a visualizar o fluxo de informações e as operações realizadas dentro de um neurônio artificial.

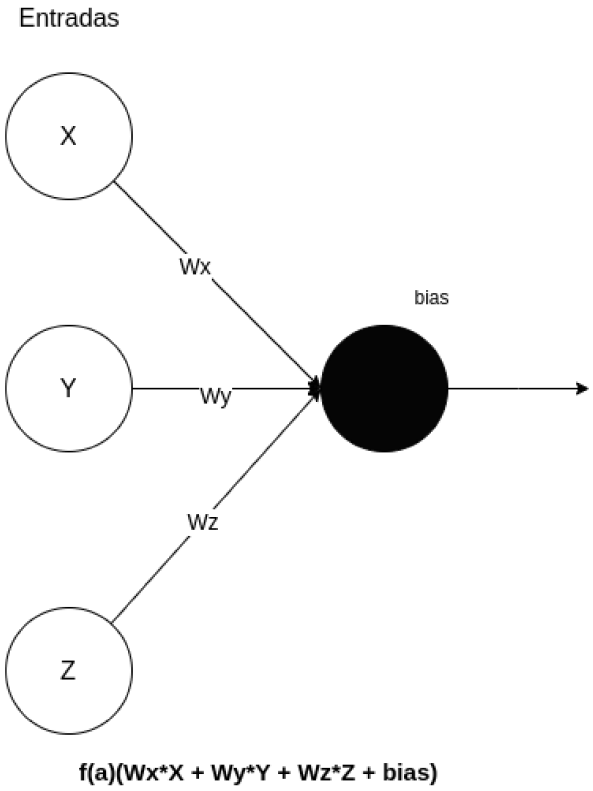


Figura 4 – Neurônio artificial simplificado

Fonte: Adaptado de (ASSEMBLYAI, 2022)

A aprendizagem em redes neurais ocorre por meio do ajuste ordenado dos pesos sinápticos, com o objetivo de realizar uma tarefa específica (HAYKIN, 2001). O verdadeiro potencial das redes neurais está na sua capacidade de se conectar a diversos outros neurônios, o que possibilita alcançar resultados notáveis no reconhecimento de padrões e em diversas outras aplicações. Através dessas conexões, as redes neurais podem capturar relações complexas nos dados de entrada, aprender com exemplos e generalizar o conhecimento adquirido para realizar predições ou tomar decisões em situações não vistas anteriormente. Essa característica de conexão em larga escala é uma das principais vantagens das redes neurais em relação a outros modelos de aprendizado de máquina.

Computacionalmente, podemos simular diversos tipos de Redes Neurais Artificiais, sendo que cada uma apresenta características significativas para resolver determinados problemas. Entre esses tipos de redes, temos as Redes Neurais Simples, Redes Neurais Profunda Densa, Redes Neurais Convolucionais, Redes Neurais Generativas e Redes Neurais de Memória de Curto Prazo, entre outras. Neste projeto, utilizamos as Redes Neurais Simples, a Rede Neural Profunda Densa e a Rede Neural Convolucional.

Uma Rede Neural Simples (ou Neural Network (NN)) representa a forma mais básica de uma rede neural artificial (ROSENBLATT, 1958). Esta estrutura é composta por três camadas principais: a camada de entrada, uma camada oculta (de processamento) e a camada de saída (GOODFELLOW; BENGIO; COURVILLE, 2016).

A camada de entrada é responsável por receber os dados de entrada, os quais são tipicamente representados como um vetor de características. Esta camada atua como uma interface inicial para os dados que serão processados pela rede (GOODFELLOW; BENGIO; COURVILLE, 2016).

Segue-se a camada oculta, que desempenha a função de processamento intermediário dos dados. Nesta camada, ocorrem transformações lineares seguidas pela aplicação de funções de ativação não-lineares. Por último, a camada de saída gera a resposta final da rede. Dependendo do problema em questão, esta resposta pode assumir a forma de uma classificação, uma probabilidade (em tarefas de classificação probabilística) ou um valor contínuo (em tarefas de regressão) (GOODFELLOW; BENGIO; COURVILLE, 2016).

Na Figura 5, ilustramos a estrutura básica de uma rede neural simples, onde a camada de saída é representada por apenas um neurônio, adequado para problemas de classificação binária. Contudo, é essencial destacar que, conforme a complexidade do problema e o número de categorias ou valores a prever aumenta, a camada de saída pode ser expandida para incluir múltiplos neurônios, cada um representando uma classe distinta ou um aspecto diferente da previsão desejada.

A simplicidade das Redes Neurais Simples advém da presença de apenas duas camadas de processamento, uma camada oculta e uma camada de saída. Essa configuração é adequada para lidar com problemas menos complexos, que não exigem uma representação de alta dimensionalidade. Embora a estrutura básica seja composta por apenas três camadas, é importante destacar que o número de neurônios em cada camada pode ser ajustado para melhor representar o problema e obter resultados mais precisos (GOODFELLOW; BENGIO; COURVILLE, 2016).

Na camada oculta, o número de neurônios pode ser maior que o tamanho do vetor de ca-

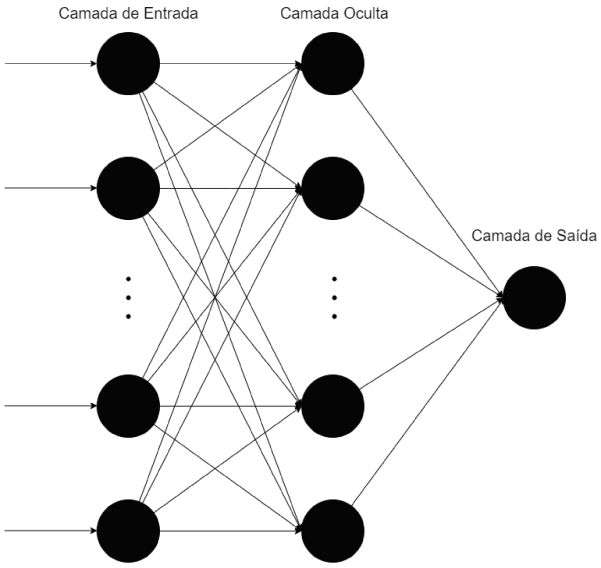


Figura 5 – Rede Neural Simples

Fonte: Elaborado pelo autor

racterísticas da camada de entrada. Na camada de saída, o número de neurônios depende da natureza do problema e do tipo de resposta desejada. Por exemplo, em um problema de classificação binária, pode ser suficiente ter um único neurônio de saída que represente a probabilidade de pertencer a uma das duas classes. No entanto, em problemas de classificação com várias classes, pode ser necessário ter um neurônio de saída para cada classe.

Portanto, embora as Redes Neurais Simples sejam caracterizadas pela presença de três camadas, a flexibilidade no ajuste do número de neurônios em cada camada permite adaptá-las às necessidades do problema em questão, maximizando seu desempenho e capacidade de representação.

Uma Rede Neural Profunda (ou Deep Neural Network (DNN)) é uma arquitetura que se diferencia das Redes Neurais Simples pela presença de múltiplas camadas ocultas para o processamento dos dados, entre a camada de entrada e a camada de saída (GOODFELLOW; BENGIO; COURVILLE, 2016). Essas camadas ocultas adicionais permitem que a rede neural aprenda representações hierárquicas e complexas dos dados de entrada (LECUN; BENGIO; HINTON, 2015). Existem vários tipos de Redes Neurais Profundas, incluindo a Rede Neural Profunda Densa e a Rede Neural Convolucional, entre outras (LECUN; BENGIO; HINTON, 2015). A principal distinção entre esses tipos reside na forma como os neurônios nas camadas ocultas estão conectados entre si (GOODFELLOW; BENGIO; COURVILLE, 2016).

Neste estudo, o conceito de Rede Neural Profunda é aplicado de forma específica às Redes Neurais Profundas Densas, que se distinguem pelo seu padrão de conectividade total: cada neurônio presente em uma camada oculta está diretamente ligado a todos os neurônios da camada subsequente (GOODFELLOW; BENGIO; COURVILLE, 2016). O termo “Densa” reflete essa característica de conexões completas entre as camadas (GOODFELLOW; BENGIO; COURVILLE, 2016).

A Figura 6 ilustra um modelo básico de uma rede neural profunda densa. É importante notar

que, dependendo da complexidade do problema a ser resolvido, podem-se adicionar camadas adicionais à rede, aumentando sua capacidade de realizar cálculos mais sofisticados. O número de neurônios na camada de saída é determinado pelo problema em questão, levando em conta o número de classes ou resultados desejados.

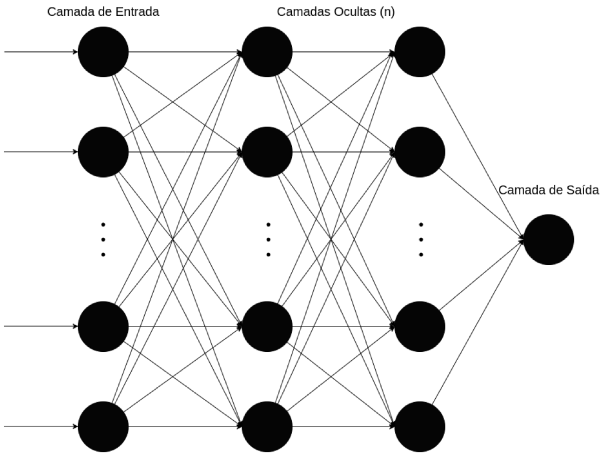


Figura 6 – Rede Neural Profunda

Fonte: Elaborado pelo autor

Cada camada oculta em uma Rede Neural Profunda Densa consiste em vários neurônios que estão totalmente conectados aos neurônios das camadas adjacentes (GOODFELLOW; BENGIO; COURVILLE, 2016). Esta configuração totalmente conectada ou *fully-connected* significa que cada neurônio em uma camada recebe sinais de todos os neurônios da camada anterior e, por sua vez, envia sinais a todos os neurônios na camada subsequente (GOODFELLOW; BENGIO; COURVILLE, 2016).

Esta densa rede de conexões facilita que cada neurônio processe informações derivadas de várias fontes da camada precedente e transmita essas informações adiante, promovendo a propagação de sinais ao longo das camadas (LECUN; BENGIO; HINTON, 2015). Tal mecanismo possibilita à rede neural capturar e aprender representações complexas e hierárquicas dos dados de entrada, identificando relações e características significativas inerentes aos dados (LECUN; BENGIO; HINTON, 2015).

O treinamento de uma Rede Neural Profunda geralmente é realizado usando o algoritmo de backpropagation, que é uma técnica de treinamento supervisionado (RUMELHART; HINTON; WILLIAMS, 1986). O backpropagation ajusta os pesos e os vieses (bias) das conexões entre os neurônios, visando minimizar o erro de predição e melhorar o desempenho da rede (RUMELHART; HINTON; WILLIAMS, 1986).

O processo de backpropagation envolve analisar a saída produzida pelos neurônios na camada de saída e compará-la com o resultado correto desejado. A partir dessa comparação, o erro é retroalimentado pelas camadas ocultas, permitindo o ajuste gradual dos pesos e vieses das conexões, de forma a reduzir o erro em cada iteração (RUMELHART; HINTON; WILLIAMS, 1986). Esse processo iterativo permite que a rede neural se aproxime cada vez mais do resultado correto, aprimorando sua capacidade de fazer previsões precisas (RUMELHART; HINTON; WILLIAMS, 1986). A figura 7 ilustra o backpropagation.

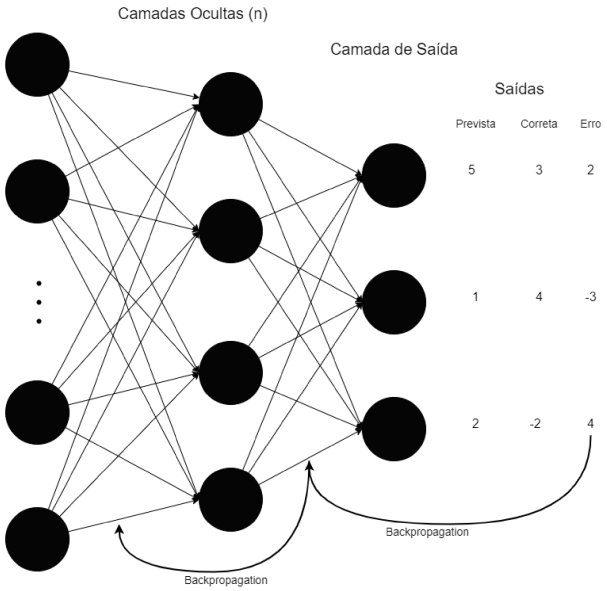


Figura 7 – Rede Neural Profunda - Backpropagation

Fonte: Elaborado pelo autor

Uma das principais vantagens das Redes Neurais Profundas é sua capacidade de aprender e extrair automaticamente características complexas dos dados, o que as torna poderosas em áreas como reconhecimento de padrões, processamento de linguagem natural, visão computacional, entre outras.

No entanto, é importante mencionar que o treinamento e a modelagem das Redes Neurais Profundas podem ser desafiadores em comparação com redes mais simples. Requer um grande conjunto de dados de treinamento para fornecer exemplos suficientes para a rede neural aprender as características relevantes. Além disso, o treinamento de redes neurais profundas pode exigir um tempo computacional significativo devido à complexidade dos cálculos envolvidos.

A escolha da arquitetura da rede, incluindo o número de camadas ocultas e a quantidade de neurônios em cada camada, também é um desafio. Uma arquitetura inadequada pode levar ao sobreajuste ou subajuste dos dados, afetando o desempenho da rede.

Para lidar com esses desafios, técnicas como regularização, dropout e normalização podem ser aplicadas (SRIVASTAVA et al., 2014). A regularização ajuda a controlar a complexidade da rede, evitando o sobreajuste (GOODFELLOW; BENGIO; COURVILLE, 2016). O dropout é uma técnica que desativa aleatoriamente uma porcentagem dos neurônios durante o treinamento, o que ajuda a evitar o sobreajuste e a aumentar a capacidade de generalização da rede (SRIVASTAVA et al., 2014). A normalização dos dados de entrada também é importante para garantir que a rede neural esteja trabalhando com valores adequados, o que pode melhorar o treinamento e a eficiência da rede (IOFFE; SZEGEDY, 2015).

Além disso, é essencial realizar experimentação e ajustes na arquitetura da rede para encontrar a configuração mais adequada para o problema específico. Isso pode envolver a tentativa de diferentes combinações de camadas ocultas, tamanhos de neurônios e funções de ativação (GOODFELLOW; BENGIO; COURVILLE, 2016).

Por fim, a Rede Neural Convolucional (ou Convolutional Neural Network (CNN)) é um tipo

especializado de rede neural que se destaca no processamento de dados com estrutura espacial, como imagens e vídeos (LECUN et al., 1998). Embora as Redes Neurais Convolucionais tenham sido inicialmente desenvolvidas para processar imagens, elas também podem ser adaptadas para outras tarefas com estrutura espacial, como processamento de sinais, reconhecimento de fala e análise de sequências temporais (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Essa adaptabilidade amplia o escopo de aplicação das CNNs e demonstra sua eficácia em várias áreas (GOODFELLOW; BENGIO; COURVILLE, 2016).

A principal característica das Redes Neurais Convolucionais é a camada de convolução. Essa camada aplica filtros de convolução a regiões locais dos dados de entrada, permitindo a extração de características relevantes em diferentes escalas e orientações (LECUN et al., 1998). Esses filtros são capazes de detectar padrões locais e invariantes, como bordas, texturas e formas, contribuindo para uma análise detalhada das informações (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

As CNNs também podem incluir camadas de pooling, como a camada de max pooling, que reduzem a dimensionalidade espacial dos dados, preservando as características mais relevantes e diminuindo a sensibilidade a pequenas variações espaciais (GOODFELLOW; BENGIO; COURVILLE, 2016). Isso torna a representação mais robusta e eficiente.

As Redes Neurais Convolucionais são amplamente utilizadas em tarefas de visão computacional, como classificação de imagens, detecção de objetos, segmentação semântica e reconhecimento de faces. Seu desempenho impressionante nessas áreas superou abordagens tradicionais e destacou a capacidade das CNNs em lidar com a complexidade das informações visuais (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Além disso, as CNNs têm a capacidade de aprendizado de recursos hierárquicos, capturando informações contextuais e identificando padrões complexos à medida que as camadas mais profundas são alcançadas (LECUN et al., 1998). Isso permite que as CNNs construam uma representação rica e significativa dos dados (GOODFELLOW; BENGIO; COURVILLE, 2016).

Em resumo, as Redes Neurais Convolucionais são um tipo especializado de rede neural que se destacam no processamento de dados com estrutura espacial, como imagens e vídeos. Sua capacidade de analisar localmente e capturar características relevantes, combinada com a aprendizagem hierárquica de recursos, tornam-nas poderosas ferramentas para tarefas de visão computacional. Com avanços contínuos e aplicação de técnicas aprimoradas, as CNNs estão impulsionando a inovação em áreas como reconhecimento de padrões, detecção de objetos e análise de dados espaciais em geral.

2.2.5 Vetores de Suporte

Os algoritmos de Máquinas de Vetores de Suporte (SVM, da sigla em inglês *Support Vector Machine)* têm como objetivo encontrar um hiperplano de separação ótimo entre os vetores de suporte, permitindo classificar novos exemplos com base em sua posição em relação a esse hiperplano. O classificador SVM foi desenvolvido com base nas ideias da teoria de aprendizagem estatística proposta por Vapnik (VAPNIK; VAPNIK, 1998). Essa teoria estabelece métricas

que devem ser seguidas para obter um classificador com boa generalização, ou seja, capaz de identificar corretamente a classe de novos dados do mesmo domínio (LORENA; CARVALHO, 2007).

As SVMs têm recebido grande atenção da comunidade de Aprendizado de Máquina devido à sua eficácia em lidar com problemas de classificação e regressão. Em muitos casos, elas demonstram resultados superiores em comparação a outros algoritmos de aprendizagem de máquina, como Redes Neurais, devido à sua habilidade de encontrar um hiperplano de separação ótimo. Esse hiperplano busca maximizar a margem, que é a distância entre o hiperplano e os vetores de suporte, e minimizar o erro nos dados de treinamento. Essa capacidade de maximizar a margem e minimizar o erro contribui para a excelente capacidade de generalização das SVMs, tornando-as aplicáveis em diversas áreas do conhecimento (CHAVES, 2006).

Em termos conceituais, uma SVM encontra um hiperplano que separa os dados de duas classes em um conjunto de dados. O objetivo é minimizar os erros marginais, ou seja, reduzir os erros nos dados de teste e treinamento, respectivamente, resultando em um hiperplano ótimo. Os vetores de suporte, que são amostras do conjunto de dados, determinam esse hiperplano. A ideia fundamental por trás das SVMs é maximizar a margem, que é a distância entre os dados de treinamento, conforme explicado porCortes e Vapnik (1995), Lorena e Carvalho (2007), e (SMOLA; SCHÖLKOPF, 1998).

A Figura 8 ilustra a distribuição do hiperplano de separação ótimo em um conjunto de dados de duas classes. O SVM busca encontrar esse hiperplano de forma a maximizar a margem entre as duas classes e utilizar os pontos de suporte para determinar sua posição (NASCIMENTO et al., 2009).

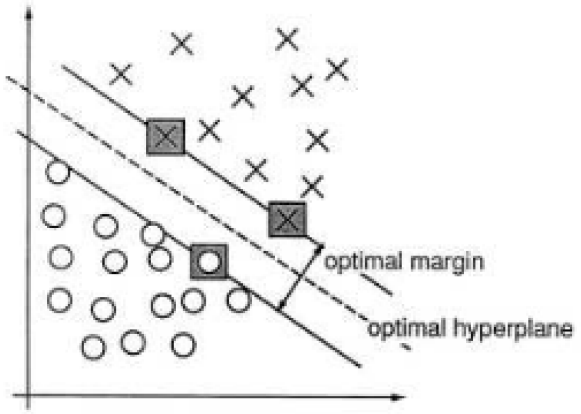


Figura 8 – Distribuição do hiperplano em uma SVM.

Fonte: Adaptado de Cortes e Vapnik (CORTES; VAPNIK, 1995).

A decisão de classificação em uma SVM é dada pela função de decisão, representada pela Equação 9, onde $f(x)$ é a função de decisão, α_i são os multiplicadores de Lagrange, y_i é o rótulo da classe, e $K(x, x_i)$ é o kernel utilizado. O vetor de pesos $W(\alpha)$ é calculado como a soma dos produtos dos multiplicadores de Lagrange e dos rótulos das classes, ponderados pelo kernel aplicado aos vetores de características (CORTES; VAPNIK, 1995).

$$f(x) = \sum_{i} \alpha_i y_i K(x, x_i) \tag{8}$$

$$W(\alpha) = \sum_{i}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{9}$$

$$Onde = \sum_{i=1}^{N} y_i \alpha_i; \forall_{i=1}^{N} : 0 \leq \alpha_i \leq C \tag{10}$$

A função de decisão $f(x)$ permite realizar a classificação de novos exemplos com base nos valores obtidos. A escolha do kernel é uma parte fundamental do modelo SVM, pois ele define o mapeamento dos dados de entrada para um espaço de maior dimensionalidade, onde a separação linear pode ser mais facilmente alcançada (HORTA et al., 2011). Existem diferentes tipos de funções de kernel, tais como:

- Linear: $K(x, x') = (x^T \cdot x')$, onde x e x' são vetores de características.

- Polinomial: $K(x, x') = (\gamma \cdot x^T \cdot x' + 1)^d$, onde γ é um fator de escala e d é o grau do polinômio.

- Radial Basis Function (RBF): $K(x, x') = \exp(-\gamma \cdot \|x - x'\|^2)$, onde γ é um fator de escala e $\|x - x'\|$ é a distância euclidiana entre os vetores x e x'.

- Sigmoidal: $K(x, x') = \tanh(\gamma \cdot x^T \cdot x' + r)$, onde γ é um fator de escala e r é uma constante.

A escolha do kernel depende da natureza dos dados e da complexidade do problema de classificação. Cada tipo de kernel tem suas próprias características e propriedades de mapeamento, e é importante selecionar aquele que melhor se ajusta aos dados e à tarefa em questão.

Além disso, o parâmetro C é especificado pelo usuário e controla a relação entre a complexidade do modelo e o número de amostras de treinamento incorretamente classificadas. Ele pode ser visto como um parâmetro de penalização, onde um valor maior de C permite uma classificação mais precisa nos dados de treinamento, mas pode levar a um modelo mais complexo e suscetível a overfitting (HORTA, 2008; SEMOLINI, 2002).

Em resumo, as Máquinas de Vetores de Suporte são algoritmos poderosos e versáteis para problemas de classificação e regressão. Elas buscam encontrar um hiperplano de separação ótimo entre os dados de treinamento, maximizando a margem e minimizando o erro. A escolha do kernel e do parâmetro de penalização são importantes para obter um modelo SVM eficaz.

O Capítulo 2 apresentou a fundamentação teórica que norteia esta pesquisa. Nele, foram colocados aspectos fundamentais para o desenvolvimento do trabalho. A seguir, serão abordados trabalhos correlatos a esta pesquisa.

Capítulo 3

Trabalhos Correlatos

Neste capítulo, são apresentados trabalhos sobre Recomendação de OA utilizando a *Wikipédia* e o *Youtube*, e trabalhos que utilizam técnicas de Aprendizado de Máquina com foco no *Youtube*. Por fim, descrevemos a evolução do nosso trabalho no estado da arte. Ademais, aponta-se que os trabalhos relacionados foram coletados por meio das plataformas de busca Google, Google Scholar e Portal de Periódicos Capes.

A Tabela 1 apresenta uma comparação entre os trabalhos relacionados e a presente pesquisa. Neste sentido, a coluna "Autores" apresenta os nomes dos autores e ano de publicação do trabalho. A coluna "Plataforma" apresenta em qual plataforma o estudo foi realizado (*Wikipédia* e/ou *Youtube*). A coluna "OA" aponta se o trabalho está relacionado ao tema de Objetos de Aprendizagem. A coluna "Class.?", refere-se ao uso de alguma técnica para classificação da base de dados utilizada. A coluna "Característica?" refere-se à característica do material escolhido que foi utilizado durante a pesquisa. Por fim, a coluna "Com. Edu.?" aponta se o trabalho utilizou os comentários da plataforma com foco educacional.

De forma geral, observa-se que apenas a presente proposta e o trabalho desenvolvido por Carvalho et al. (2020b) utilizam os comentários e realizam a análise de comentários educacionais (comentários obtidos de vídeos educacionais), mas que apenas a abordagem proposta realiza a classificação de vídeos do *Youtube* por meio de técnicas de Aprendizado de Máquina em comentários. Observa-se, adicionalmente, que a abordagem proposta, apresenta diferentes experimentos utilizando os comentários como forma de classificação dos OA, onde são realizados experimentos de classificação determinística (educacional ou não educacional) e probabilística (probabilidade do vídeo ser educacional).

Menolli, Malucelli e Reinehr (2011) objetivam gerar OA, através da *Wikipedia*, utilizando tecnologias semânticas e o padrão *Learning Object Metadata* (LOM) com a Web 2.0. Em sua proposta, os conteúdos inseridos na plataforma são acessados, e realizada a mineração de textos para extração e classificação dos conteúdos de acordo com o padrão LOM. Utilizar esse padrão, possibilita encontrar os atributos e metadados da página, gerando um *XML-schema* com os metadados trabalhados. Concluem sobre a necessidade dessa abordagem, por facilitar a utilização dos conteúdos, uma vez que as ferramentas wikis não consideram como o conteúdo será utilizado.

Abu-El-Haija et al. (2016) abordam a classificação de vídeos do *Youtube* visando desenvolver

Tabela 1 – Comparação entre os trabalhos relacionados e a abordagem proposta

Autores	Plataforma	OA?	Class.?	Característica?	Com. Edu.?
(MENOLLI; MALUCELLI; REINEHR, 2011)	Wikipedia	Sim	Sim	Textos	Não
(ABU-EL-HAIJA et al., 2016)	Youtube	Não	Sim	Frames / imagens	Não
(JÚNIOR; DORÇA, 2018)	Wikipedia	Sim	Sim	Textos	Não
(PINHEIRO et al., 2018)	Youtube	Sim	Sim	? (Não informado)	Não
(THELWALL, 2018)	Youtube	Não	Sim	Comentários	Não
(AFONSO; DUQUE, 2019)	Youtube	Não	Não	Comentários	Não
(CARVALHO et al., 2020a)	Youtube	Sim	Não	Categoria do vídeo	Não
(CARVALHO et al., 2020c)	Youtube	Sim	Não	Categoria do vídeo	Não
(AMANDA; NEGARA, 2020)	Youtube	Não	Sim	Títulos e descrição	Não
(TRINDADE et al., 2020)	Youtube	Sim	Não	Títulos, descrição e tags	Não
(CARVALHO et al., 2020b)	Youtube	Sim	Não	Comentários	Sim
(ZHENG et al., 2021)	Youtube	Não	Não	Comentários	Não
Abordagem proposta	Youtube	Sim	Sim	Comentários	Sim

um sistema de classificação múltipla de vídeos. A base de dados utilizada conta com aproximadamente 8 milhões de vídeos, englobando o total de 1,9 bilhão de quadros, e 500 mil horas de vídeos categorizados. A pesquisa foi realizada em duas etapas, a saber: 1) os rótulos dos vídeos foram obtidos por meio do *Knowledge Graph entities*; 2) os vídeos foram procesados *frame* a *frame* e categorizados por uma Rede Neural Convolucional pré-treinada no *ImageNet*. O *ImageNet* é um banco de dados visual com diversos objetos/entidades já classificados. Através do processamento de mais de 50 anos de vídeos, gerando 2 bilhões de *frames*, e mais de 8 milhões de vídeos que podem ser rapidamente modelados em uma única máquina, a o trabalho aponta no sentido de auxiliar o desenvolvimento de pesquisas sobre compreensão de vídeos. Apesar das diversas classes de categorização, não foi encontrada uma categoria específica para vídeos educacionais. O trabalho cita a categoria "*Jobs & Education*" na qual estão universidades, salas de aulas, palestras, etc. Assim sendo, um vídeo que possua imagens de um campus universitário, por exemplo, será enquadrado nessa categoria, apesar de não necessariamente ser um vídeo educacional.

Júnior e Dorça (2018) apresentam uma abordagem para criação e recomendação de OA por meio da plataforma *Wikipédia*. A abordagem é definida por três etapas: 1) enriquecimento da ontologia por meio dos metadados das seções wiki; 2) recomendação dos OA - utilizando técnicas de Problema de Cobertura de Conjuntos combinados com Algoritmo Genético; 3) uso de operações CRUD (*Create, Read, Update, Delete*). O trabalho conclui que a abordagem adotada resolve o problema da recomendação de OA, retornando soluções de elevada qualidade.

Pinheiro et al. (2018) apresentam o *Easy Youtube*, um Sistema de Recomendação de OA baseado no *Youtube*. O sistema tem seu funcionamento em seis etapas, como segue: 1) enriquecimento de consultas - estabelecimento de temas pré-definidos, cadastrados por especialistas; 2) extração de vídeos - busca de vídeos, que pode ser realizada por meio de pesquisa ou de temas

pré-definidos; 3) pré-processamento - tratamento dos textos (em português), com remoção de pontuação, espaços, etc.; 4) classificação - utilização de algoritmo para classificar os vídeos considerados como educacionais e de qualidade; 5) engenho de recomendação - o sistema recebe os vídeos considerados "bons" e classifica-os; 6) coletor de *feedback* - o usuário avalia a recomendação fornecida pelo sistema por meio de notas de uma a sete estrelas. O trabalho indica suas principais contribuições nos pontos: 1) o Sistema de Recomendação desenvolvido pode ser utilizado como solução para vários domínios de aplicação; 2) o sistema serviu como prova de conceito para melhorar as recomendações, por meio de características do *Youtube*, como a avaliações dos usuários, e linguagem nativa do vídeo. O trabalho apresentado, porém, não detalha questões importantes da pesquisa. Por exemplo, para a classificação de vídeos considerados de qualidade, afirma-se que foi utilizado um conjunto de treinamento de 100 vídeos, os quais foram avaliados por especialistas e alunos do tema "Orientação a Objetos/Herança". Entretanto, não se explica de que maneira esta análise foi realizada, e quais características dos vídeos foram consideradas. Outro ponto que causa confusão é a declaração de que, devido ao prazo exíguo para a realização da pesquisa, o foco do trabalho foi em "algumas características para o experimento". Tais características não foram descritas.

Thelwall (2018) por sua vez, analisa os comentários de vídeos do *Youtube* relacionados a estilos de dança. A base de dados utilizada contém 36.702 vídeos. O trabalho objetiva identificar, por meio dos comentários postados nos vídeos da plataforma, os tipos de dança, relações quanto aos gêneros (masculino e feminino), sentimentos expressados, e discussões referentes aos estilos de dança. Utiliza-se, para tanto, o método denominado *Comment Term Frequency Comparison* (CTFC) na tentativa de identificação de subtópicos/subtemas das discussões sobre determinado tópico nos comentários do *Youtube*, questões de gênero, sentimentos, e relacionamento entre tópicos. O método utilizado define com sucesso diversas atitudes predominantes em homens e mulheres. Os 10 termos homem-associados foram: *shit, fuck, shuffle, man, fucking, crip, dude, bro, shuffling, hardstyle*. Por outro lado, os 10 termos mulher-associados foram: *she, amazing, her, beautiful, cute, omg, belly, ballet, really, workout*. A análise de sentimentos forneceu ideias plausíveis dos motivos pelos quais as danças eram apreciadas. Os 10 termos positivos mais utilizados foram: *please, nice, wow, beautiful, loved, job (e.g. nice/great/good job), pretty, hope, perfect, keep (going/up the good work/it up)*. Por sua vez, os 10 termos negativos mais utilizados foram: *shit, fuck, killed, stupid, wtf, hate, idiot, dislike, die, dead*. Os autores afirmam que os resultados poderiam servir de partida para análises mais aprofundadas sobre o tema e que a pesquisa destacaria as diferenças de gêneros, sentimentos, e subtópicos entre as danças. Consideram que o método utilizado pode ser útil para discutir, em larga escala, fenômenos específicos do *Youtube*, bem como pode ser útil em outros contextos para fornecer análise exploratória inicial de certo problema que não havia sido pesquisado anteriormente.

Afonso e Duque (2019) realizaram análise de sentimentos em comentários de vídeos do *Youtube* utilizando técnicas de Aprendizado de Máquina supervisionada. O trabalho coletou 918 comentários de um vídeo que apresenta análises e críticas ao filme "Batman versus Superman: a origem da justiça". Os comentários coletados foram classificados como positivos, negativos ou neutros. Foram realizados três experimentos, a saber: 1) três classes de polaridade: positiva, negativa e neutra; 2) duas classes: negativa e não negativa; 3) utilizaram-se apenas os

comentários que apresentavam a referência "filme Batman vs Superman", e consideraram-se as classes negativa e não negativa. Utilizou-se o classificador SMO *(Sequential Minimal Optimization algorithm for training a support vector classifier)* e metodologia de *8-fold cross-validation*. A acurácia máxima obtida foi de 81%. Os autores apontam que talvez seja possível aumentar a acurácia por meio da coleta de comentários de outros vídeos.

Carvalho et al. (2020a) apresentam o sistema Educavídeos, um sistema de Recomendação de vídeos do *Youtube* que realiza a busca de vídeos por meio de suas categorias. Observaram que que o uso da categoria "Education" pelo Educavídeos apresenta melhores resultados para identificação de vídeos educacionais do que a plataforma *Youtube* em modo padrão de busca.

Carvalho et al. (2020c) realizam uma análise de diversos vídeos do *Youtube* e as categorias atribuídas aos mesmos. Os autores identificaram que a grande maioria dos vídeos se apresentam categorizados erroneamente, apontando, desta maneira, que as categorias fornecidas pela plataforma não devem ser consideradas como determinantes para busca de vídeos. Ademais, os autores apontam falhas no mecanismo de busca do *Youtube*, a saber: falta de informação quanto as categorizações dos vídeos, dificuldade em se adicionar termos para refinar as buscas, e impossibilidade de realizar buscas por vídeos utilizando suas categorias.

Amanda e Negara (2020) aplicaram técnicas de Aprendizado de Máquina para classificar vídeos do *Youtube* entre "*Kesenian*" e "*Sains*", que, em tradução do Indonésio, significa "Arte" e "Ciência", respectivamente. Os autores utilizaram o motor de busca da plataforma com as palavras *Kesenian* e *Sains* e obtiveram seus dados experimentais na forma de links, títulos e descrições de vídeos. Foram utilizados 3 classificadores, a saber: Random Forest, SVM, e Naïve Bayes. O classificador com melhor avaliação foi o Naïve Bayes que obteve acurácia de 88%, enquanto os classificadores Random Forest e SVM obtiveram acurácia de 82%.

Trindade et al. (2020) apresentam a proposta de um sistema de Recomendação de OA baseado em vídeos do *Youtube*. O sistema realiza buscas na plataforma e também fornece um conjunto de vídeos que atenda aos conteúdos demandados pelo usuário. O modelo de recomendação proposto é baseado no problema min-max para o Problema de Cobertura de Conjuntos, que objetiva minimizar o custo dos OA e o número máximo de repetição de conceitos apresentados. Os metadados dos vídeos como título, *likes*, *dislikes* e *views* são utilizados no processo de recomendação. Os autores afirmam que o sistema proposto pode ser expandido para utilização em outras plataformas e que a abordagem utilizada é valida, pois possibilita reduzir a carga mental e evitar desmotivação dos alunos causada pela repetição de conteúdos.

Carvalho et al. (2020b) analisaram 200 vídeos, 100 educacionais e 100 não educacionais, e identificaram diferenças relevantes entre os termos e vocábulos mais frequentes empregados nos comentários dos vídeos de ambas as categorias. Neste sentido, notaram que termos como "melhor professor" e "ótima aula" estão presentes apenas na lista dos termos mais frequentes nos comentários dos vídeos educacionais. De modo similar, apontaram que que os radicais "obrig", "aul" e "profes" aparecem com alta frequência em comentários de vídeos educacionais. O estudo sugeriu que os comentários dos usuários do *Youtube* apresentam potencial para serem utilizados para categorizar os vídeos da plataforma.

Zheng et al. (2021) analisaram comentários postados em vídeos diários do primeiro-ministro canadense durante a pandemia de COVID-19. Foram analisados 46.732 comentários, em inglês,

obtidos de 57 vídeos postados entre 13 de Março a 22 de Maio de 2020. O objetivo do estudo era analisar os comentários fornecidos por usuários sobre os resumos diários de COVID-19 do primeiro-ministro canadense com o intuito de avaliar a mudança das opiniões e preocupações do público. Os autores afirmam que o estudo estabelece um feedback em tempo real entre a sociedade e a autoridade de saúde, possibilitando identificar e trabalhar de acordo com as preocupações da sociedade, podendo assim, aumentar a confiança entre o público e o governo.

Durante as pesquisas sobre o tema abordado, não foram identificados trabalhos que identificassem e auxiliassem o processo de escolha de Objetos de Aprendizagem por meio da opinião dos usuários, tampouco vídeos educacionais da plataforma *YouTube*. Inicialmente, o trabalho de Abu-El-Haija et al. (2016) foi identificado quanto à classificação/categorização dos vídeos do *YouTube*. Esta pesquisa realiza a classificação de vídeos da plataforma, porém sem foco educacional, e a categorização é realizada por meio dos *frames* dos vídeos. Essa ideia, apesar de poder ser utilizada, pode demandar elevada capacidade de processamento, caso fosse desenvolvido um sistema em tempo real para a classificação dos vídeos da plataforma. Nesse sentido, continuou-se com o objetivo de identificar outros meios de classificação dos vídeos educacionais do *YouTube*. Posteriormente, identificou-se o trabalho de Pinheiro et al. (2018); porém, a pesquisa não detalha partes importantes sobre sua metodologia, e não foi possível encontrar informações detalhadas sobre o que foi de fato desenvolvido. Em seguida, identificou-se o trabalho de Thelwall (2018), que analisa os comentários do *YouTube* com o intuito de classificar tipos de danças. Este trabalho despertou a ideia da utilização dos comentários para a classificação dos vídeos educacionais do *YouTube*.

Na pesquisa, foi desenvolvida uma abordagem inovadora que utiliza comentários para classificar Objetos de Aprendizagem (OAs). Além disso, duas variações foram criadas: a "rígida" e a "flexível". Na variação "rígida", todos os comentários de um determinado vídeo são agregados como um único "documento", e o classificador atribui uma classificação determinística, determinando se o vídeo é Educacional ou Não Educacional. Por outro lado, na variação "flexível", o classificador analisa individualmente cada comentário, classificando-os como educacional ou não educacional. Em seguida, é calculado e apresentado o "grau de certeza" de que o vídeo, como um todo, é Educacional, ou seja, um OA.

Capítulo 4

Abordagem Desenvolvida

Este capítulo descreve a abordagem empregada para combinar tecnologias de aprendizado de máquina e mineração de texto, objetivando fornecer recomendações personalizadas de objetos de aprendizagem. A abordagem proposta tem como finalidade facilitar a busca de materiais educacionais por aprendizes e educadores em repositórios de conteúdo não exclusivamente educacionais. Diferenciando-se de estudos, este trabalho emprega técnicas de processamento de texto sobre os comentários de usuários para recomendar materiais pertinentes. Embora o Youtube tenha sido escolhido como cenário experimental para esta pesquisa, a abordagem é adaptável a diferentes plataformas com ajustes mínimos. O Sistema de Recomendação desenvolvido, denominado LOIS, oferece suporte tanto a docentes e aprendizes por meio de sua interface web quanto a AVA, por meio de respostas no formato JSON.

As subseções a seguir detalham cada etapa da abordagem adotada neste projeto. A primeira subseção fornece um quadro geral das fases do projeto e introduz brevemente o conteúdo das subseções subsequentes. A segunda subseção oferece uma visão geral do processo, passando por todas as etapas e fornecendo uma explicação preliminar do que será explorado mais detalhadamente na sequência.

A terceira subseção aborda a análise dos vídeos e a coleta dos comentários, estabelecendo a base para o tratamento dos dados. A quarta subseção detalha o pré-processamento desses dados, uma fase crítica para a preparação eficaz dos mesmos para análises subsequentes. A quinta subseção explica como os comentários são modelados para uso em algoritmos de aprendizado de máquina, os quais têm o objetivo de classificar e recomendar Objetos de Aprendizagem.

Finalmente, a sexta subseção discute como é realizada a filtragem baseada em comentários e apresenta o LOIS, o sistema que integra todas as etapas do projeto. É importante destacar que as etapas de pré-processamento, análise dos dados, modelagem e preparação dos dados, bem como a filtragem baseada em comentários, fazem parte do campo da Mineração de Texto. A filtragem de comentários, especificamente, está interligada com técnicas de Aprendizado de Máquina, visando à classificação efetiva de textos.

4.1 Metodologia

A metodologia adotada para o desenvolvimento é ilustrada na Figura 9 e descrita a seguir.

A metodologia implementada neste projeto foi estrategicamente desenvolvida para superar os desafios associados à recomendação de Objetos de Aprendizagem (OAs) em ambientes digitais não exclusivamente educacionais. A primeira fase envolve a definição e análise dos materiais educacionais. Isso inclui a clara delimitação dos materiais que serão analisados e utilizados na composição da base de dados. Esta fase crucial compreende a identificação precisa do que configura um vídeo educacional, garantindo que a seleção dos materiais seja relevante e esteja alinhada aos objetivos do estudo.

Durante esta etapa, a coleta de comentários dos usuários é essencial, demandando uma compreensão detalhada e a automação desse processo. Além disso, é imperativo avaliar meticulosamente as ferramentas disponíveis para apoiar o desenvolvimento contínuo do projeto. Para esse fim, empregamos duas ferramentas principais: o aplicativo *Mozdeh Big Data Text Analysis* e a *API* do YouTube. Cada ferramenta apresenta particularidades e limitações: o Mozdeh enfrenta desafios com análises em tempo real e restrições linguísticas, sendo primordialmente ajustado para o inglês; a *API* do YouTube, embora robusta, está sujeita a limites de cotas e restringe o número de resultados por consulta.

A segunda etapa, de pré-processamento de texto, e a terceira etapa, de análise e exploração de dados, requerem atenção especial por serem complexas e impactarem diretamente no desenvolvimento subsequente. No pré-processamento, aplicam-se técnicas para adicionar mais significado e relevância aos comentários, reduzindo ruídos nos dados e destacando características textuais essenciais para a classificação dos vídeos.

Após o pré-processamento, segue-se a análise e exploração de dados, onde se realiza uma investigação sobre os comentários e sua correlação com cada classe. Esta exploração é vital para identificar termos e expressões indicativos de conteúdo educacional e não educacional, sendo crítica para compreender o engajamento dos usuários com os vídeos e determinar quais características dos textos podem ser utilizadas para aprimorar a classificação e a recomendação de OAs. O sucesso nesta etapa pode aprimorar significativamente a relevância das recomendações do sistema.

A etapa subsequente envolve planejar como utilizar os comentários para classificar os OAs. O desenvolvimento da abordagem metodológica e suas variações é uma das fases mais complexas do projeto. Nesta fase, define-se como os textos serão utilizados e modelados para funcionarem como métricas para a classificação, enfrentando o desafio de representar textualmente dados que o computador interpreta como números.

É essencial destacar que a modelagem eficaz dos textos possibilita a extração de conhecimento e a representação adequada de cada classe, permitindo o uso por algoritmos de aprendizado de máquina. A adequada representação dos textos é crucial para capturar e aproveitar as características relevantes.

Finalmente, chegamos à etapa de filtragem baseada em comentários. Neste estágio, definimos e aplicamos algoritmos de aprendizado de máquina para validar e verificar a eficácia das abordagens propostas em atingir os objetivos do projeto. A seleção desses algoritmos envolve

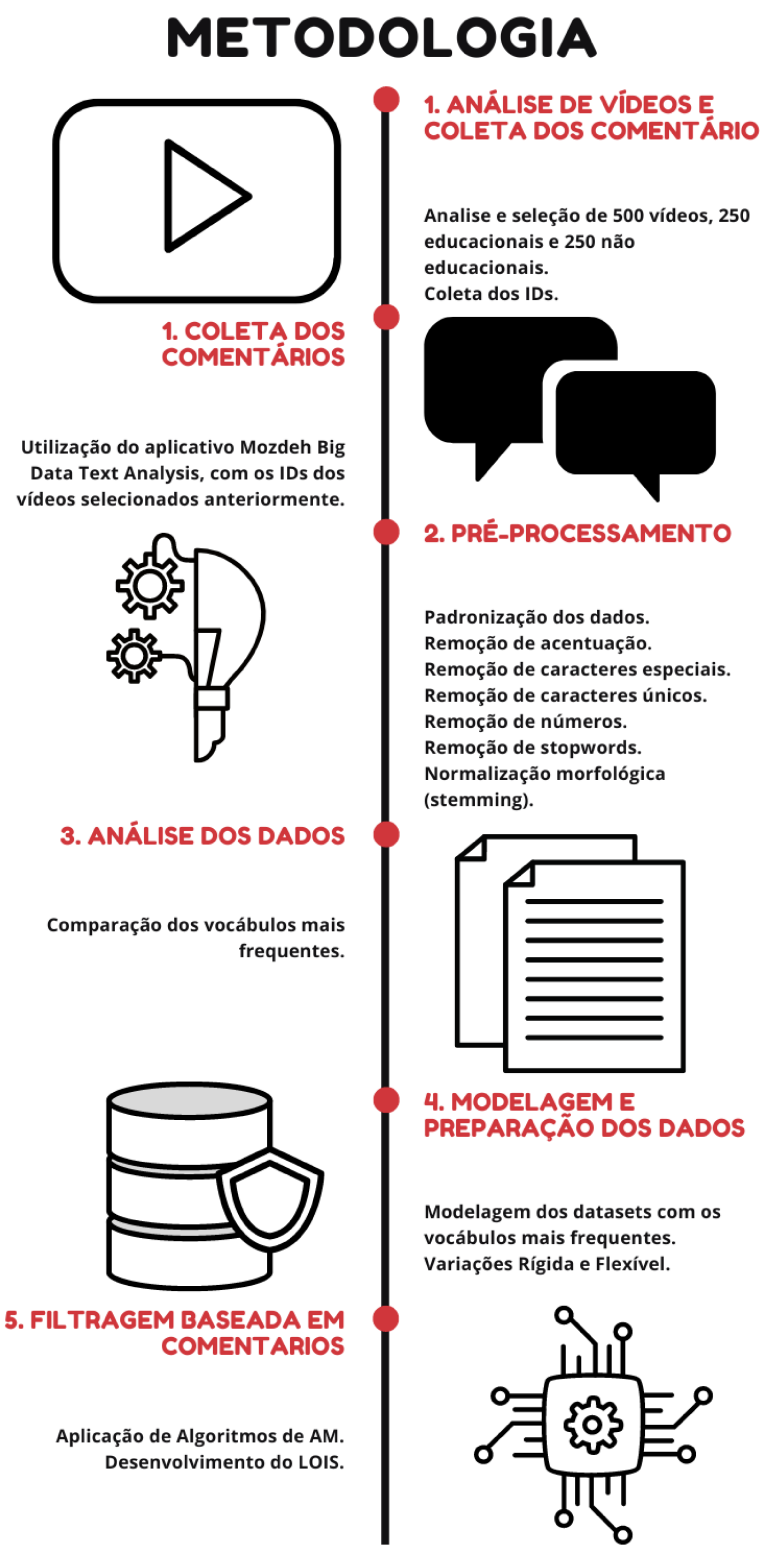


Figura 9 – Metodologia utilizada para a pesquisa.

uma fase de experimentação abrangente, abordando tanto algoritmos supervisionados quanto não supervisionados, dependendo das necessidades específicas de classificação e análise.

A escolha adequada dos algoritmos é crucial e depende intensamente dos resultados obtidos nos experimentos. Durante esse processo, é fundamental ajustar corretamente a modelagem das redes neurais, visto que isso tem um impacto direto na precisão e eficácia dos resultados. As redes neurais necessitam de uma configuração minuciosa e podem variar em complexidade desde alterar o número de neurônios até adicionar camadas adicionais.

Além disso, o desenvolvimento do Sistema de Recomendação LOIS consolida todos os passos anteriores em uma aplicação prática e funcional. É vital que este sistema possua uma interface intuitiva e amigável, permitindo que usuários de todos os níveis técnicos possam utilizá-lo com facilidade. O sistema também deve ser capaz de integrar-se com outros sistemas, facilitando que ambientes virtuais de aprendizagem recomendem os Objetos de Aprendizagem (OAs) mais adequados aos seus estudantes.

Em resumo, a abordagem empregada no projeto é uma sequência lógica e integrada de etapas que se estendem desde a coleta e análise de dados até o desenvolvimento e implementação de um sistema de recomendação. Cada etapa é meticulosamente projetada para construir sobre a anterior, utilizando técnicas avançadas de análise de dados e aprendizado de máquina para converter comentários de usuários em recomendações precisas e valiosas.

4.2 Visão Geral

Esta abordagem inovadora fundamenta-se amplamente no feedback — comentários e impressões — de uma diversidade de usuários, incluindo estudantes, educadores e indivíduos com interesses variados. A inovação desse trabalho consiste em utilizar essas opiniões coletadas diretamente dos vídeos para serem utilizados por algoritmos de aprendizado de máquina, permitindo assim uma classificação e recomendação mais precisas e contextualizadas de conteúdos educacionais. Ao analisar os comentários, o sistema é capaz de identificar padrões linguísticos e temáticos que distinguem vídeos educacionais de outros tipos, utilizando essas percepções para o processo de seleção e recomendação de Objetos de Aprendizagem. Este enfoque não só melhora a relevância e a personalização das recomendações oferecidas aos usuários, mas também potencializa a utilidade prática do sistema em ambientes educacionais diversificados.

Adicionalmente, o Sistema de Recomendação LOIS se destaca por ser pioneiro na utilização da opinião dos usuários para oferecer recomendações de novos Objetos de Aprendizagem diretamente do YouTube. Esta característica coloca o LOIS como o primeiro a implementar tal classificação em vídeos da plataforma, oferecendo uma abordagem inovadora que aprimora significativamente a experiência educacional ao recomendar recursos educativos relevantes.

O principal objetivo deste estudo é explorar e investigar os comentários de usuários com o intuito de prover recomendações precisas de Objetos de Aprendizagem (OAs) em plataformas que não estão restritas apenas a materiais educacionais. Para alcançar essa meta, é crucial distinguir efetivamente entre conteúdos educacionais e não educacionais. Além disso, torna-se necessário analisar os comentários através de técnicas de Mineração de Texto, a fim de transformá-los em

métricas úteis para os algoritmos de Aprendizado de Máquina (AM).

A metodologia implementada se articula em cinco etapas fundamentais: (1) Análise de vídeos e coleta de comentários; (2) Pré-processamento dos comentários; (3) Análise dos dados; (4) Metodologias para modelagem e preparação dos dados; e (5) Filtragem baseada em comentários. As etapas 2, 3, 4 e 5 estão intrinsecamente associadas à Mineração de Textos, envolvendo técnicas como visualização de dados e extração de conhecimento, bem como a aplicação de algoritmos de aprendizado de máquina. Especificamente, a etapa 5 está diretamente conectadas ao uso de algoritmos de aprendizado de máquina para aperfeiçoar a modelagem e realizar a filtragem dos vídeos através dos comentários.

Identificar vídeos não educacionais mostra-se relativamente mais simples, categorizando conteúdos como clipes musicais, apresentações de comédia stand-up e peças teatrais como não educacionais. Contudo, o desafio reside em estabelecer e discernir os critérios que definem um vídeo como educacional, exigindo a formulação de uma definição abrangente e precisa para o conteúdo educacional, que constitui parte essencial do projeto.

Uma definição clara do que constitui um vídeo educacional foi utilizada e apresentada por Gomes (2008), o passo seguinte é a coleta de comentários. Esta seleção de vídeos foi meticulosamente conduzida com base em critérios definidos sobre o que é considerado um vídeo educacional. Inicialmente, analisamos um conjunto de 200 vídeos, sendo 100 educacionais e 100 não educacionais, seguido por uma expansão que incluiu mais 300 vídeos ao estudo, totalizando 500 vídeos, onde 250 pertenciam a cada classe. Estes foram cuidadosamente selecionados e categorizados manualmente, após uma avaliação visualização do completa do conteúdo.

Após a categorização manual dos vídeos em cada categoria, tornou-se crucial identificar métodos eficazes para coletar dados que representassem adequadamente cada grupo. Recorremos, então, ao aplicativo *Mozdeh Big Data Text Analysis*[1], especificamente para a coleta de comentários. Este software gera um arquivo com várias informações, incluindo a hora da publicação do comentário, o ID do vídeo, o nome do usuário que comentou, entre outros. No entanto, focamos exclusivamente nos comentários, apesar do aplicativo oferecer suporte a múltiplos idiomas, com especial atenção para o inglês, que não era o principal foco de nossa pesquisa.

Embora o Mozdeh ofereça uma rica variedade de dados, sua incapacidade de coletar informações em tempo real limitou sua utilidade contínua para nosso estudo. Para superar essa limitação e aproveitar a coleta de dados em tempo real, migramos para a utilização da *API* do YouTube. A *API* permite acesso instantâneo aos dados do YouTube e pode ser operada por meio de *scripts* em Python, a linguagem escolhida para o desenvolvimento do projeto. Essa mudança não apenas resolveu a questão da coleta em tempo real mas também alinhou melhor as práticas de desenvolvimento do sistema com as necessidades do projeto.

A sequência do estudo engloba as etapas cruciais para seu desenvolvimento, inseridas no âmbito da Mineração de Texto, que contempla diversas fases destinadas ao trabalho com a linguagem natural escrita. Entre estas, destacam-se o pré-processamento e a análise de comentários, fundamentais para o desenvolvimento da metodologia proposta.

Após a obtenção dos dados, procedemos ao pré-processamento dos comentários, uma etapa crucial para aprimorar a qualidade das informações coletadas e facilitar a diferenciação entre as

[1] Desenvolvido por Thelwall (2018) e disponível em: http://mozdeh.wlv.ac.uk/installation.html

categorias de vídeos. Este processo envolve a padronização dos comentários, remoção de informações irrelevantes e identificação de palavras-chave significativas para cada categoria. Segue-se para análise dos dados, a qual oferece uma compreensão aprofundada das interações dos usuários com as duas categorias de vídeos, proporcionando informações importantes sobre o comportamento dos usuários e evidenciando diferenças notáveis entre as classes, permitindo a modelagem apropriada dos dados para a subsequente aplicação de técnicas de Aprendizado de Máquina, visando alcançar os objetivos propostos.

Após o processamento eficaz dos comentários, torna-se essencial estabelecer metodologias que modelem os dados e os preparem para os algoritmos de aprendizado de máquina. Essas metodologias são cruciais, dado que os algoritmos, isoladamente, não têm capacidade para interpretar textos. Portanto, um tratamento adequado dos dados é fundamental para atingir os objetivos desejados.

Inicialmente, desenvolveu-se uma metodologia "rígida", na qual um vídeo e seus comentários são unificados para serem representados como um único vetor de entrada para a classificação como educacional ou não educacional. Dessa forma, os algoritmos de aprendizado de máquina recomendam vídeos nos quais têm "confiança" em sua natureza educacional. Em uma fase posterior, para superar limitações técnicas e obter uma classificação mais adaptável, foi introduzida a metodologia "flexível". Nesta variação, cada vídeo é representado pela quantidade de seus comentários, que são classificados individualmente como educacionais ou não. Posteriormente, os vídeos são recomendados com base em um "grau de confiança" de serem educacionais, considerando a quantidade de comentários classificados como educacionais e não educacionais.

O resultado desse trabalho extensivo é a criação do LOIS (Learning Object Intelligent Search), um sistema pioneiro de recomendação de vídeos educacionais do Youtube através da opinião dos usuários. Com base nos dados coletados e nas análises realizadas, o LOIS é capaz de filtrar e sugerir conteúdo educacional, utilizando técnicas avançadas de processamento de texto e algoritmos de aprendizado de máquina, ativados a cada nova pesquisa realizada no sistema. Isso não só demonstra a viabilidade prática da pesquisa, mas também seu potencial impacto ao facilitar o acesso a materiais educacionais relevantes em plataformas de conteúdo diversificado, contribuindo significativamente para a educação mediada por tecnologia.

A inovação deste trabalho reside na aplicação prática de técnicas de Processamento de Textos em um contexto educacional, diferenciando-se pelo seu enfoque na classificação e recomendação de Objetos de Aprendizagem a partir de plataformas não exclusivamente educacionais. Enquanto o Processamento de Textos não constitui uma novidade por si só, a maneira como essas técnicas são integradas com algoritmos de Aprendizado de Máquina para criar um sistema capaz de discernir conteúdos educacionais efetivamente em meio a um vasto acervo de informações representa uma abordagem pioneira. Este método não apenas melhora a acessibilidade e a seleção de recursos educativos para educadores e alunos, mas também estabelece um novo paradigma no uso da Inteligência Artificial para apoiar e enriquecer o processo educacional em ambientes digitais com conteúdos diversificados.

4.3 Análise dos Vídeos e Coleta dos Comentários

A análise dos vídeos é uma das partes mais cruciais do trabalho. Essa etapa é o que define o sucesso ou o fracasso da abordagem, pois através de conceitos bem definidos que teremos a possibilidade de identificar corretamente o que é um vídeo educacional. E, através da análise correta dos vídeos que é possível construir uma amostra correta e significativa que forneça boa base para alcançar os objetivos propostos. Para isso, é necessário obter uma boa definição do que é um vídeo educacional.

Para a identificação de vídeos educacionais foi abordada a definição de Gomes (2008) acerca de vídeos educativos: "produto específico, produzido com intenção didático-pedagógica e que considera seu contexto de recepção como especialmente a escola e a sala de aula, sendo, portanto, intrinsecamente diferente dos vídeos de documentários, entrevistas, reportagens, etc.". Acredita-se que esta definição apresenta-se abrangente o suficiente para a análise dos vídeos, sendo utilizada para a categorização de um vídeo como educacional ou não.

Após uma identificação de uma definição do que é um vídeo educacional é necessário identificar os exemplos. Nesse momento é necessário coletar diversos exemplos para possibilitar a generalização de características específicas de cada classe por um algoritmo de Aprendizado de Máquina. A identificação, análise, dos vídeos ocorreu de forma manual, ou seja, foram visualizados os vídeos e categorizados na classe educacional ou não, seguindo a definição apresentada anteriormente.

A análise e coleta dos vídeos ocorreram em duas etapas, primeiro foram coletados 200 vídeos, 100 da classe educacional e 100 da classe não educacional, posteriormente, a coleta foi expandida e passou para 500 vídeos, sendo 250 de cada classe. O objetivo foi manter a proporção em cada classe, entretanto, no mundo real, essa proporção não é seguida rigorosamente, através da coleta, percebeu-se que vídeos educacionais tendem a ter menos comentários que vídeos não educacionais.

De forma geral, a seleção dos vídeos foi realizada sem se considerar assuntos ou critérios específicos, portanto, foram selecionados vídeos sobre os mais diferentes assuntos e temas, com objetivo de se construir uma base de dados bastante diversa.

Em relação à seleção de vídeos educacionais, ressalta-se que parte dos vídeos elencados encontram-se no *Youtube Edu*, um canal do *Youtube/Google* em parceria com a fundação Lemann com o intuito de fornecer conteúdos educacionais gratuitos e de qualidade, em português. O conteúdo dos vídeos do canal é voltado para o Ensino Fundamental e Ensino Médio, e aborda temas das disciplinas de Língua Portuguesa, Química, Física, Biologia, Matemática, História, Geografia, Língua Inglesa, e Língua Espanhola. Alguns exemplos de vídeos retirados deste canal para comporem da base de dados são: Sy_LUnePfRE - Acentuação Gráfica Malha Funk da Acentuação [Prof Noslen]; e sQewkYR4_sg - Geografia - Aula 01 - Orientação e Cartografia. Destaca-se, ainda, que a seleção dos vídeos ocorreu independente do ambiente em que eles foram gravados, por exemplo, o vídeo _bKzJP0Q778 - Cursos Unicamp: Cálculo I - Aula 8 - Regras de Cálculo de Limite - parte 1, foi selecionado para compor a base de dados, é gravado por uma pessoa em uma sala de aula onde o professor utiliza um quadro negro.

Embora alguns vídeos sejam encontrados no canal *Youtube Edu*, é importante esclarecer que

o canal atua mais como um "repositório", divulgando vídeos que originalmente pertencem a outros canais, e não é o criador dos conteúdos.

A seleção de vídeos não educacionais é composta de vídeos de diversos tipos, como musicas, *reviews*, notícias, jogos, comédia, entre outros. Alguns exemplos de vídeos selecionados para comporem a base de dados são: j7v8dMisr78 - Primeiras doses da vacina Coronavac estão chegando ao Brasil; hcuABOkWqSA - XJ6: tudo que você precisa saber - Review; e XPqy0ozwN94 - Filhote de Onça Pintada é encontrado ferido em um lixão de Santa Luzia.

Por fim, é importante destacar que durante a seleção dos vídeos, observou-se que alguns destes não continham ou desabilitaram os comentários. Desta forma, estes vídeos não foram adicionados à base de dados, pois, deve-se buscar outras alternativas para proceder a classificação dos mesmos.

Após a análise e seleção dos vídeos, foi necessário coletar os comentários dos vídeos escolhidos. Esta etapa envolveu a coleta dos comentários dos vídeos selecionados previamente, realizada por meio do *Mozdeh Big Data Text Analysis*, ferramenta também citada em trabalhos anteriores como os de Thelwall (2018) e Carvalho et al. (2020b).

O aplicativo oferece um vasto conjunto de informações sobre cada comentário. Dentre os dados disponibilizados estão: Search (Busca), Title (Título), VideoID (ID do Vídeo), CommentID (ID do Comentário), CommentPublished (Data de Publicação do Comentário), CommentUpdated (Data de Atualização do Comentário), CommentTextDisplay (Texto do Comentário), CommentAuthorName (Nome do Autor do Comentário), CommentAuthorURI (URI do Autor do Comentário), CommentCanReply (Possibilidade de Resposta ao Comentário), CommentTotalReplyCount (Contagem Total de Respostas ao Comentário), CommentisPublic (Publicidade do Comentário), CommentLikeCount (Contagem de Likes do Comentário), CommentVewerRating (Avaliação do Comentário pelo Visualizador), IsReply (Se é Resposta), CommentPosterInfo (Informações do Postador do Comentário). Apesar da riqueza de metadados fornecidos, apenas os textos dos comentários, disponíveis em CommentTextDisplay, foram utilizados.

O *Mozdeh Big Data Text Analysis* foi inicialmente empregado apenas para a coleta preliminar de comentários. Apesar de oferecer uma variedade rica de metadados e funcionalidades, seu foco principal é o processamento de textos em língua inglesa, e sua incapacidade de operar em tempo real limita sua utilidade para os propósitos deste trabalho. Para superar essas limitações e desenvolver um Sistema de Recomendação que funcione em tempo real, optou-se pela utilização da *API* do YouTube. Esta escolha permitiu a coleta contínua e integrada de dados através de requisições automatizadas feitas por *scripts* em Python, adequando-se melhor às necessidades de um sistema dinâmico e interativo.

Apesar de a *API* do YouTube funcionar bem em tempo real, ela está sujeita a limites de cotas que restringem o número de resultados por consulta. As principais dificuldades relacionadas a essas limitações dizem respeito à quantidade de requisições permitidas. Uma busca por vídeos pode retornar até 50 resultados, e se forem necessários mais vídeos, outra requisição deve ser feita para a próxima "página". Da mesma forma, a coleta de comentários retorna até 100 comentários por requisição, sendo necessário realizar requisições adicionais para obter mais comentários, repetidamente. Além disso, cada requisição consome uma parte da cota disponível, que é limitada pela *API*. Portanto, é essencial utilizar essas requisições de maneira estratégica

para recomendar materiais com o menor custo possível.

Finalizada as coletas, o conjunto de dados contendo 200 vídeos, 100 educacionais e 100 não educacionais, apresentava 158.559 comentários e, o conjunto de dados com 500 vídeos, sendo 250 educacionais e 250 não educacionais possuía 738.653 comentários.

4.4 Pré-processamento dos comentários

A fase de pré-processamento dos comentários é uma etapa crítica no tratamento de dados textuais, especialmente quando estes provêm de fontes informais de comunicação como comentários de vídeos online. Esses textos, coletados das interações dos usuários, são ricos em informações, mas também em irregularidades e ruídos que podem interferir na análise subsequente. Portanto, a limpeza e a formatação adequadas são fundamentais para garantir a compreensão das percepções obtidas com os dados.

O processo de pré-processamento envolve várias técnicas detalhadas a seguir, cada uma com o objetivo de refinar os dados para uma análise mais eficaz:

- Padronização dos Dados: Esta etapa envolve a seleção criteriosa dos dados que serão efetivamente utilizados nas análises subsequentes. Uma prática comum é a transformação do texto para um formato padrão, geralmente convertendo todas as letras para minúsculas. Isso é crucial para evitar duplicidades ou variações de uma mesma palavra, como "Casa"e "casa", que seriam interpretadas como entidades distintas pelo computador.

- Remoção de Acentuação: Dada a variabilidade linguística, especialmente em idiomas como o português, a acentuação pode criar variações desnecessárias das palavras. A remoção de acentos padroniza o texto, simplificando a análise.

- Remoção de Caracteres Especiais: Caracteres como exclamações, arrobas e hashtags são frequentes em textos online, mas geralmente não agregam valor semântico. Sua remoção ajuda a uniformizar as expressões, tratando variações de uma mesma frase como equivalentes. No primeiro momento foram removidos os caracteres especiais, como: !, #, @, dentre outros. Isso é necessário, por exemplo, para que comentários tais como, “muito bom!” e “muito bom”; “melhor professor!” e “melhor professor”. Foram removidos os caracteres: [”, ’"’, ’#’, ’$’, ’%’, ’&’, ’’’, ’(’, ’)’, ’*’, ’+’, ’,’, ’-’, ’.’, ’/’, ’:’, ’;’, ’<’, ’=’, ’>’, ’?’, ’@’, ’[’, ’\’, ’]’, ’^’, ’_’, ’`’, ’{’, ’|’, ’}’, ’~’]

- Eliminação de Caracteres Isolados: Letras isoladas, que frequentemente não contribuem para o sentido do texto, são removidas. Isso inclui letras que, fora de um contexto específico, não possuem significado relevante para a análise. Eliminação de caracteres isolados, como “e”, “a”, dentre outros.

- Remoção de Espaços Extras: Espaços redundantes entre palavras ou após pontuações podem ser removidos sem prejuízo ao significado do texto, garantindo uma padronização na formatação.

- Exclusão de Números: Em muitos casos, os números presentes nos textos não são relevantes para a análise desejada e, portanto, são eliminados para focar no conteúdo textual.

- Eliminação de Stopwords: Stopwords são palavras comuns que aparecem frequentemente nos textos, mas que, por sua natureza genérica (como preposições, artigos, pronomes), não contribuem significativamente para a interpretação do conteúdo. A remoção dessas palavras reduz o volume de dados a ser analisado e permite um foco maior em termos que realmente carregam significado. Estas podem ser artigos, preposições, advérbios, pronomes, e outras palavras auxiliares (MORAIS; AMBRÓSIO, 2007). De forma geral, de 20 a 30% das palavras de um texto são *stopwords* (KANNAN et al., 2014). Alguns exemplos de *stopwords* são "tem", "isto", "aos", dentre outras. Foram removidas as stopwords: a, à, ao, aos, aquela, aquelas, aquele, aqueles, aquilo, as, às, até, com, como, da, das, de, dela, delas, dele, deles, depois, do, dos, e, é, ela, elas, ele, eles, em, entre, era, eram, éramos, essa, essas, esse, esses, esta, está, estamos, estão, estar, estas, estava, estavam, estávamos, este, esteja, estejam, estejamos, estes, esteve, estive, estivemos, estiver, estivera, estiveram, estivéramos, estiverem, estivermos, estivesse, estivessem, estivéssemos, estou, eu, foi, fomos, for, fora, foram, fôramos, forem, formos, fosse, fossem, fôssemos, fui, há, haja, hajam, hajamos, hão, havemos, haver, hei, houve, houvemos, houver, houvera, houverá, houveram, houvéramos, houverão, houverei, houverem, houveremos, houveria, houveriam, houveríamos, houvermos, houvesse, houvessem, houvéssemos, isso, isto, já, lhe, lhes, mais, mas, me, mesmo, meu, meus, minha, minhas, muito, na, não, nas, nem, no, nos, nós, nossa, nossas, nosso, nossos, num, numa, o, os, ou, para, pela, pelas, pelo, pelos, por, qual, quando, que, quem, são, se, seja, sejam, sejamos, sem, ser, será, serão, serei, seremos, seria, seriam, seríamos, seu, seus, só, somos, sou, sua, suas, também, te, tem, têm, temos, tenha, tenham, tenhamos, tenho, terá, terão, terei, teremos, teria, teriam, teríamos, teu, teus, teve, tinha, tinham, tínhamos, tive, tivemos, tiver, tivera, tiveram, tivéramos, tiverem, tivermos, tivesse, tivessem, tivéssemos, tu, tua, tuas, um, uma, você, vocês, vos.

- Normalização Morfológica (Stemming): A aplicação de técnicas de stemming reduz as palavras à sua raiz ou forma base, diminuindo a variabilidade morfológica. Por exemplo, palavras como "correr", "correu" e "correndo" seriam simplificadas para o mesmo radical "corr", assim como, "professor" e "professora", seriam reduzidas a "profes". Essa normalização facilita a identificação de temas ou tópicos comuns nos textos, independentemente das variações verbais.

O pré-processamento dos comentários é uma etapa indispensável na análise de dados textuais, pois prepara o dataset para a aplicação de métodos de análise mais complexos, como classificação. Ao limpar e formatar os dados, reduz-se o ruído e destaca-se a informação relevante, aumentando significativamente a qualidade das percepções observadas. Este processo não apenas otimiza a capacidade de identificar padrões relevantes nos dados, mas também diminui a carga computacional necessária para processar informações desnecessárias ou redundantes.

4.5 Análise dos dados

Após refinar o conjunto de dados, eliminando inconsistências e ruídos, avançamos para a análise detalhada dos dados. Esta fase é fundamental, pois é aqui que discernimos os padrões nos comentários que distinguem vídeos educacionais dos não educacionais e examinamos como os usuários interagem com esses conteúdos. Para isso, empregamos *scripts* em Python, utilizando bibliotecas como NLTK para processamento de linguagem natural, Pandas para manipulação de dados, Numpy para cálculos matemáticos e Matplotlib para visualização.

Nessa etapa, utilizamos técnicas de mineração de textos, visualização de dados e extração de informações. O objetivo dessa fase é extrair conhecimento dos dados, percebendo as diferenças entre as duas classes e avaliando a possibilidade de usar os comentários para classificar os vídeos. Essa análise, em conjunto com a visualização, fornece percepções sobre se o projeto pode continuar conforme o planejado ou se ajustes são necessários. A visualização ajuda a discernir melhor as diferenças entre as classes. Destacam-se, nesta etapa, técnicas importantes como tokenização e análise de frequência.

Inicialmente, nossa análise se concentrou em uma amostra preliminar de 200 vídeos, igualmente divididos entre educacionais e não educacionais. Investigamos as diferenças qualitativas e quantitativas nos comentários dessas duas categorias, explorando como cada grupo se expressa e se engaja com o conteúdo. Encorajados pelas percepções iniciais, expandimos nossa análise para incluir um total de 500 vídeos, mantendo a divisão equitativa entre as categorias.

A investigação subsequente aprofundou-se na análise dos comentários, revelando diferenças significativas nos padrões de engajamento e nos termos usados pelas duas categorias. Termos educacionais, por exemplo, mostraram-se prevalentes em vídeos de sua respectiva classe, enquanto expressões típicas de descontração dominavam na outra. Esta análise detalhada validou a hipótese de que é possível utilizar os comentários como um meio eficaz para diferenciar tipos de conteúdo, além de proporcionar percepções sobre a interação dos usuários com os vídeos.

Este estágio do projeto se mostrou extremamente revelador, desvendando não apenas as peculiaridades do vocabulário usado nos comentários das duas categorias, mas também estabelecendo um ponto de partida robusto para o desenvolvimento do projeto. As descobertas desta fase foram cruciais para planejar a modelagem dos dados que seriam usados em algoritmos de aprendizado de máquina.

Os resultados desta análise forneceram uma base sólida para as etapas subsequentes de modelagem e aplicação de algoritmos de classificação. A percepção de que as palavras frequentes em uma categoria variam significativamente na outra foi instrumental para desenvolver a abordagem de modelar os dados baseando-se nos termos mais frequentes em ambas as classes.

Adicionalmente, o uso eficaz das informações extraídas dos comentários confirmou que eles constituem uma fonte valiosa para compreender melhor as preferências e comportamentos dos usuários, além de servirem como um critério eficaz de classificação de OAs. Essas percepções propiciaram significativamente o desenvolvimento das metodologias aplicadas na classificação e na recomendação de conteúdos educativos.

4.6 Metodologias para Modelagem e Preparação dos Dados

Na etapa anterior, observou-se diferenças significativas nas frequências das palavras mais recorrentes entre as classes, decorrentes da aplicação de técnicas de tokenização e análise de frequência. Essas palavras frequentes foram utilizadas para modelar os comentários, empregando esses vocábulos como características distintivas. No entanto, para que possam ser utilizadas de maneira adequada, é crucial modelar o dataset (características e textos) de forma a representar adequadamente todo o texto nas características especificadas.

Após analisar os dados, constatamos a necessidade de desenvolver uma metodologia robusta para a modelagem desses dados, de forma que possam ser eficazmente utilizados por algoritmos de aprendizado de máquina na classificação de Objetos de Aprendizagem (OA). Esta etapa é fundamental em projetos que envolvem dados textuais, pois requer a transformação dos textos — intrinsecamente qualitativos e repleto de nuances — em um formato quantitativo que possa ser processado eficientemente por computadores.

A vetorização é uma técnica comum na modelagem de dados textuais, que transforma textos em vetores numéricos. Uma das técnicas mais utilizadas é a Bag of Words (BoW), que identifica as palavras mais frequentes em um corpus (conjunto de todos os textos) e as usa como características para representar os documentos. Nesta abordagem, cada documento é representado pela frequência de cada palavra característica no texto, ou seja, quantas vezes cada palavra aparece no texto.

Para este projeto, adotamos uma abordagem mais sofisticada do que o BoW tradicional. Decidimos compilar um vocabulário equilibrado de características, incorporando as palavras mais frequentes encontradas tanto em vídeos educacionais quanto em vídeos não educacionais. Este vocabulário equilibrado visa capturar uma representação mais abrangente e precisa dos conteúdos analisados.

O uso de um vocabulário balanceado oferece vantagens significativas para o modelo de classificação. Esta abordagem assegura uma representação equitativa dos dados, mitigando o risco de viés que poderia advir de uma possível predominância de uma das classes, seja por um volume maior de dados ou por um engajamento diferenciado dos usuários. A integração das palavras mais utilizadas de cada categoria capacita o modelo a identificar eficazmente os padrões linguísticos que caracterizam cada tipo de conteúdo.

Essa distribuição equilibrada de características facilita o reconhecimento pelo algoritmo de aprendizado de máquina das diferenças entre os tipos de vídeos, permitindo-lhe também usar essas informações para prever de forma mais acurada a classificação de vídeos novos. Durante a análise dos dados, a variação significativa na frequência e relevância dos termos entre as classes foi decisiva para adotar um vocabulário balanceado como características no modelo. Isso é especialmente importante considerando que os vídeos não educacionais tendem a apresentar maior engajamento.

Para demonstrar como ocorre a geração de características para a montagem do dataset, vamos considerar o uso de 10 vocábulos. O dataset de 10 vocábulos é formado pelos cinco

termos mais frequentes nos vídeos educacionais e pelos cinco mais frequentes nos vídeos não educacionais. Os vocábulos “aul”, “profess”, “obrig”, “vide” e “aprend” são os predominantes nos vídeos educacionais e constituem as primeiras cinco características. Para os vídeos não educacionais, os termos mais recorrentes são “pra”, “ta”, “vc”, “faz” e, inicialmente, “vide”. No entanto, para evitar repetições e enriquecer o dataset, substituímos “vide” pelo próximo termo mais frequente, “music”. Assim, o dataset será estruturado com os vocábulos “aul”, “profess”, “obrig”, “vide”, “aprend”, “pra”, “ta”, “vc”, “faz” e “music”. A modelagem para outros datasets segue a mesma metodologia estabelecida, variando apenas no número de vocábulos utilizados.

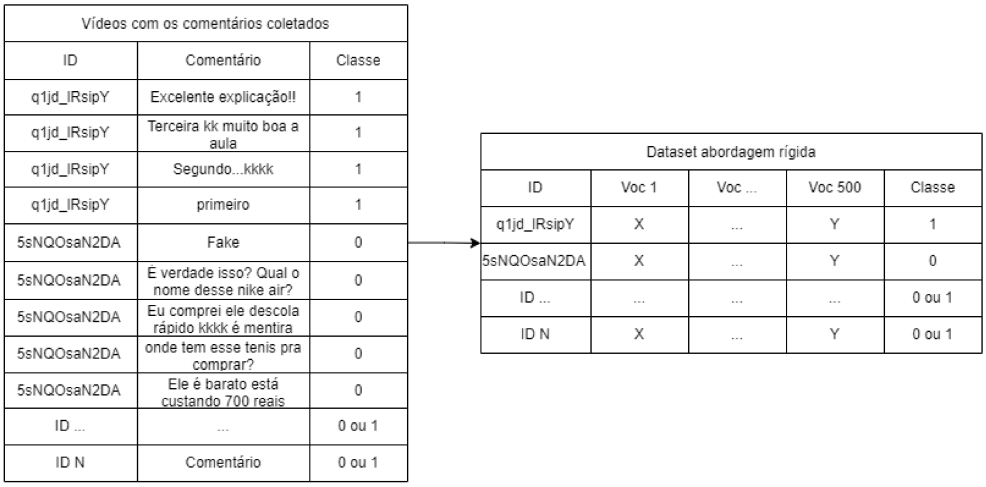
Após definir as características (vocábulos), transformamos os dados textuais para representar a frequência com que cada vocábulo aparece nos comentários, ou seja, é realizada a contagem de repetições que determinado vocábulo aparece nos comentários. A representação do dataset irá variar se modelarmos para que cada entidade do dataset represente um vídeo ou para que represente cada comentário individualmente, uma distinção crucial para a apresentação dos dados ao algoritmo e para a classificação realizada pelo algoritmo.

A modelagem marca uma mudança significativa na representação dos dados e define como serão as entradas e saídas nos algoritmos de aprendizado de máquina. Inicialmente, desenvolvemos uma variação “rígida”, estruturando o dataset para que cada vídeo seja representado por uma entrada, permitindo que os algoritmos de aprendizado de máquina classifiquem os vídeos como educacionais ou não. À medida que a pesquisa avançou e restrições foram descobertas, introduzimos uma variação mais “flexível”, que estrutura os vídeos como um conjunto de comentários individuais. Essa abordagem permite recomendar vídeos com determinado “grau de certeza” de serem educacionais, através da classificação individualizada de cada comentário de um vídeo, ampliando a aplicabilidade das classificações e superando as dificuldades identificadas.

Em suma, a análise dos dados mostrou que os vocábulos mais frequentes de cada classe têm impactos distintos. Com essa informação, decidimos modelar o dataset usando esses termos como características para criar um vocabulário equilibrado, fundamental para a classificação eficaz dos OAs. As Figuras 10 e 11 mostram como os comentários são vetorizados para serem representados pela variação rígida e pela variação flexível, respectivamente. Para exemplificação, utilizaremos dois vídeos presentes no dataset: o vídeo com Id “q1jd_lRsipY” intitulado “Porcentagem”, que é um vídeo educacional, e o vídeo com Id “5sNQOsaN2DA” intitulado “Tênis da Nike que muda a COR! / Nike shoes change color.”, que é não educacional e contém 4 e 5 comentários, respectivamente.

Como ilustrado nas figuras, a metodologia “rígida” consolida todos os comentários de um vídeo em uma única linha do dataset modelado, onde todos os comentários são representados coletivamente, e cada elemento do vetor que representa um vídeo é a frequência com que essa característica aparece em todos os comentários do vídeo. Essa linha única é utilizada pelos algoritmos de aprendizado de máquina (AM) para classificar o vídeo como educacional ou não educacional.

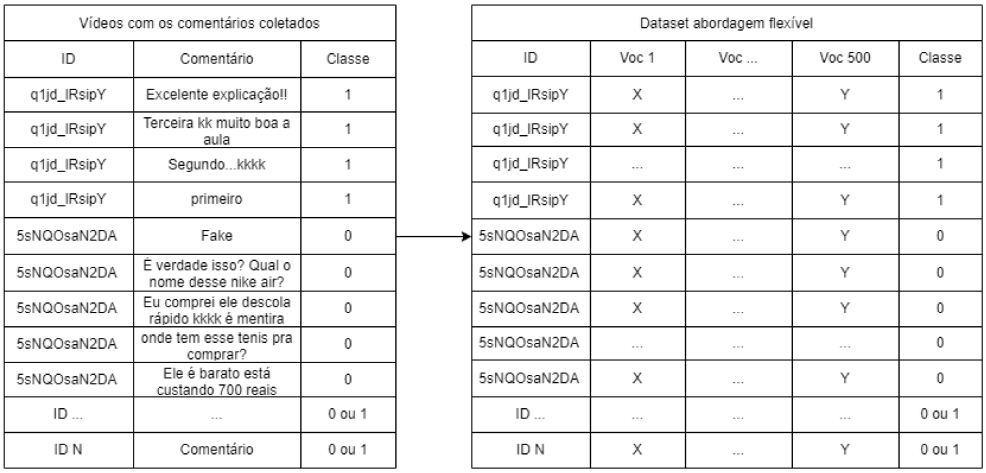
Por outro lado, na metodologia “flexível”, cada comentário é vetorizado e transformado individualmente em uma linha dentro do dataset, e cada posição do vetor representa a frequência de determinada característica apenas naquele comentário. Nesse modelo, os algoritmos de AM classificarão cada comentário separadamente, em vez do vídeo como um todo. Assim, indepen-



Vídeos com os comentários coletados		
ID	Comentário	Classe
q1jd_IRsipY	Excelente explicação!!	1
q1jd_IRsipY	Terceira kk muito boa a aula	1
q1jd_IRsipY	Segundo...kkkk	1
q1jd_IRsipY	primeiro	1
5sNQOsaN2DA	Fake	0
5sNQOsaN2DA	É verdade isso? Qual o nome desse nike air?	0
5sNQOsaN2DA	Eu comprei ele descola rápido kkkk é mentira	0
5sNQOsaN2DA	onde tem esse tenis pra comprar?	0
5sNQOsaN2DA	Ele é barato está custando 700 reais	0
ID ...	...	0 ou 1
ID N	Comentário	0 ou 1

Dataset abordagem rígida				
ID	Voc 1	Voc ...	Voc 500	Classe
q1jd_IRsipY	X	...	Y	1
5sNQOsaN2DA	X	...	Y	0
ID ...	...	...	...	0 ou 1
ID N	X	...	Y	0 ou 1

Figura 10 – Modelagem dos comentários processados para a metodologia “Rígida”.



Vídeos com os comentários coletados		
ID	Comentário	Classe
q1jd_IRsipY	Excelente explicação!!	1
q1jd_IRsipY	Terceira kk muito boa a aula	1
q1jd_IRsipY	Segundo...kkkk	1
q1jd_IRsipY	primeiro	1
5sNQOsaN2DA	Fake	0
5sNQOsaN2DA	É verdade isso? Qual o nome desse nike air?	0
5sNQOsaN2DA	Eu comprei ele descola rápido kkkk é mentira	0
5sNQOsaN2DA	onde tem esse tenis pra comprar?	0
5sNQOsaN2DA	Ele é barato está custando 700 reais	0
ID ...	...	0 ou 1
ID N	Comentário	0 ou 1

Dataset abordagem flexível				
ID	Voc 1	Voc ...	Voc 500	Classe
q1jd_IRsipY	X	...	Y	1
q1jd_IRsipY	X	...	Y	1
q1jd_IRsipY	...	...	...	1
q1jd_IRsipY	X	...	Y	1
5sNQOsaN2DA	X	...	Y	0
5sNQOsaN2DA	X	...	Y	0
5sNQOsaN2DA	X	...	Y	0
5sNQOsaN2DA	...	...	...	0
5sNQOsaN2DA	X	...	Y	0
ID ...	...	...	...	0 ou 1
ID N	X	...	Y	0 ou 1

Figura 11 – Modelagem dos comentários processados para a metodologia “Flexível”.

dentemente de quantos comentários um vídeo possuir, na metodologia “rígida”, o vídeo será representado como apenas uma entrada, uma linha. Já na metodologia “flexível”, um vídeo será representado pela quantidade de comentários que possuir.

Além disso, ao representar os vídeos em uma única linha, os valores de cada célula dos vetores, que representam a quantidade de vezes que um vocábulo - característica - aparece, tendem a ser muito superiores na metodologia “rígida” em comparação com a “flexível”. Por exemplo, o vídeo com ID “uWa2WLOveaQ” e título “Geografia pro Enem - O que é Espaço Geográfico? Lugar, Paisagem, Região e Território” teve a palavra “aul” representada com uma frequência de 79 na metodologia “rígida”, necessitando de 229 comentários para alcançar essa utilizaçãp. Entretanto, na metodologia “flexível”, é improvável que um único comentário contenha a palavra “aul” 79 vezes.

É importante ressaltar que a quantidade de comentários necessários para que um vocábulo atinja uma frequência alta, como 79, varia conforme o engajamento dos usuários em cada vídeo. No entanto, devido ao fato da metodologia “rígida” possuir uma representação única (um vetor) para todos os comentários de um vídeo, os valores de cada linha do dataset tendem a ter

frequências muito maiores em comparação com cada linha do dataset na metodologia "flexível".

As figuras também mostram o uso de até "Voc 500", indicando que o dataset inclui 500 características ou vocábulos. Esses vocábulos são utilizados pelos algoritmos para diferenciar e classificar os vídeos ou comentários, enquanto os valores X e Y representam a frequência com que cada vocábulo aparece em um determinado vídeo ou comentário.

As próximas subseções fornecerão detalhes adicionais sobre as metodologias "rígida" e "flexível".

4.6.1 Metodologia Rígida

A "metodologia rígida" foi desenvolvida inicialmente com o intuito de diferenciar e obter uma classificação clara sobre os vídeos serem educacionais ou não educacionais. A partir da abordagem previamente descrita, que envolve a seleção equilibrada de vocábulos frequentes para uso na classificação, foi necessário modelar o dataset para atingir os objetivos propostos.

Ademais, nos experimentos utilizando a metodologia rígida, foram gerados datasets com diferentes números de características. Inicialmente, foram preparados oito conjuntos contendo 10, 20, 40, 60, 80, 100, 200 e 6 vocábulos, este último foi gerado através do dataset que continha 200 vocábulos, seguidos por um adicional contendo 500 termos, refinando a abordagem conforme a análise progredia. Os conjuntos de dados foram organizados incluindo o identificador (ID) do vídeo, as características correspondentes aos vocábulos mais frequentes, e a classificação do vídeo, que pode ser educacional ou não.

Após a seleção dos vocábulos a serem usados como características, é essencial converter os textos para um formato numérico. Neste processo de modelagem, cada vídeo é representado por uma única linha em uma tabela. A primeira coluna da tabela identifica o ID do vídeo. As colunas seguintes representam a frequência em que cada característica (vocábulo) aparece nos comentários. A última coluna especifica a classificação do vídeo, indicando se é educacional ou não educacional.

Para demonstrar o processo desde a coleta dos comentários até a classificação dos vídeos, utilizaremos dois vídeos específicos do conjunto de dados, e consideraremos que cada um conta com apenas três comentários. O primeiro vídeo de caráter educacional, possui o ID "_QS0oy42bWg" e é intitulado "Aula Biologia - Origem da Vida - Visão Geral e Histórico para o Enem e Vestibulares - STOODI". O segundo, classificado como não educacional, tem o ID "eNA906t98LI" e é denominado "Qual notebook eu uso?". A Tabela 2 apresenta a amostra dos comentários dos dois vídeos e os comentários processados são apresentados na Tabela 3.

Após a etapa de pré-processamento, os comentários devem ser transformados para um formato numérico, permitindo a sua utilização por algoritmos de aprendizado de máquina. Esta transformação é essencial para a análise computacional, convertendo texto em linguagem natural para dados estruturados que representam as características ou "features" identificadas nos comentários.

Para realizar a vetorização dos comentários, passa-se por todos os vocábulos referente as características, e realiza-se a contagem de quantas vezes esse vocábulo apareceu nos comentários daquele vídeo com o determinado "ID". O vídeo "_QS0oy42bWg" apresenta os seguintes

Tabela 2 – Metodologia Rígida: Amostra dos comentários dos vídeos IDs "_QS0oy42bWg" e "eNA906t98LI"

IdVideo	Comentário	Class
_QS0oy42bWg	muito boa aula! parabéns!	1
_QS0oy42bWg	adorei a aula meu Deus , e alem disso ele é muito lindo né , por favor kk amei o jeito que ele fala aí meu Deus kk mais a aula tava maravilhosaa	1
_QS0oy42bWg	Melhor video aula serio	1
eNA906t98LI	Boa noite! tenho 11 anos e gosto muito de design, eu queria comprar um notebook para começar a aprender um pouco mais sobre, você me recomenda algum barato e que rode programas de edição?	0
eNA906t98LI	Obrigada! ajudou bastante.	0
eNA906t98LI	Quero aprender a trabalhar com ele. Tenho e quase não usei para artes...Tem uma playlist de como usar?	0

Tabela 3 – Metodologia Rígida: Amostra dos comentários dos vídeos IDs "_QS0oy42bWg" e "eNA906t98LI" processados.

IdVideo	Comentário	Class
_QS0oy42bWg	boa aul parab	1
_QS0oy42bWg	ador aul deu alem diss lind ne favor kk ame jeit fal ai deu kk aul tav maravilh	1
_QS0oy42bWg	melhor vide aul	1
eNA906t98LI	boa noit ano gost design quer compr notebook comec aprend pouc sobr recomend algum barat rod program edica	0
eNA906t98LI	obrig ajud bast	0
eNA906t98LI	quer aprend trabalh quas use art playlist us	0

comentários processados: "boa aul parab", "ador aul deu alem diss lind ne favor kk ame jeit fal ai deu kk aul tav maravilh" e "melhor vide aul". Para a representação do vídeo no exemplo, iremos contar a frequência das características. O termo "aul" se repete quatro vezes nesses três comentários, enquanto "profess" e "obrig", não são utilizados, com isso sua representação é de zero, o vocábulo "vid" aparece uma vez nos comentários, e os vocábulos "aprend", "pra", "ta", "vc", "faz", e "music" não foram mencionados, resultando em uma representação de zero para essas palavras. Assim, a representação do vídeo ID "_QS0oy42bWg" será representado termos de features como [4, 0, 0, 1, 0, 0, 0, 0, 0, 0]. Adicionalmente, o "ID" é representado no início e no final a classe 1, que faz referência a classe educacional.

Por outro lado, no vídeo com o ID "eNA906t98LI", os termos "aul", "profess", "vide", "pra", "ta", "vc", "faz", e "music" não foram citados nos comentários. Apenas os termos "obrig" e "aprend" foram utilizados, com "obrig" aparecendo uma vez e "aprend" duas vezes nos comentários. Dessa forma, a representação do vídeo ID "eNA906t98LI" em termos de features foi [0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0].

A Tabela 4 mostra como esses dois vídeos e seus comentários seriam representados para a metodologia rígida no dataset composto por 10 features.

A descrição anterior demonstra o processo desde a coleta dos comentários até a criação

Tabela 4 – Metodologia Rígida: Amostra do dataset dos vídeos IDs "_QS0oy42bWg" e "eNA906t98LI" considerando 10 features.

IdVideo	aul	profes	obrig	vide	aprend	pra	ta	vc	faz	music	Class
_QS0oy42bWg	4	0	0	1	0	0	0	0	0	0	1
eNA906t98LI	0	0	1	0	2	0	0	0	0	0	0

Tabela 5 – Amostra da base de dados considerando os 10 vocábulos mais frequentes

IdVideo	aul	profes	obrig	vide	aprend	pra	ta	vc	faz	music	Class
4g9JTQ2B6oo	21	28	4	3	0	5	1	5	1	0	yes
uWa2WLOveaQ	79	52	29	17	7	11	5	11	8	0	yes
ZlB6MZmpKls	27	13	8	2	1	1	1	3	2	0	yes
qaZ3fsUhBG8	5	9	3	8	1	3	1	2	2	0	yes
LSqOKMnakU4	33	38	17	23	10	18	4	6	16	2	yes
7NKlihwokyk	9	31	66	214	17	241	46	119	322	0	no
DSBHxRcMrzI	2	1	9	22	6	52	4	13	59	0	no
UkhSLsDgj4M	56	254	135	343	53	811	92	344	1266	29	no
CFvy6zSsOEc	0	1	0	51	20	0	0	0	0	0	no
PHigOIqh5SY	0	1	1	12	0	13	12	10	0	44	no

de um dataset estruturado. Utilizamos uma amostra com os 10 termos mais frequentes para exemplificar, mas a mesma metodologia foi aplicada para o desenvolvimento de outros datasets com diferentes números de termos (características).

Este procedimento detalha a conversão dos comentários em dados que podem ser interpretados por algoritmos de aprendizado de máquina. As características selecionadas, derivadas das palavras mais frequentes nos comentários, servem como indicadores cruciais para distinguir entre vídeos educacionais e não educacionais.

Uma amostra do dataset contendo 10 vocábulos como características é apresentado na Tabela 5. A tabela é organizada de maneira que o ID do vídeo é apresentado na primeira coluna, seguido pelas colunas dos termos selecionados e finalizando com a classificação do vídeo. Este método assegura que a modelagem seja eficaz e livre de duplicações, promovendo uma análise precisa e significativa dos dados.

A organização do conjunto de dados foi cuidadosamente projetada para que cada vídeo seja representado por uma única linha. Cada termo (ou característica) é mapeado pelo número de vezes que aparece nos comentários vinculados a um vídeo específico, que é identificado por um ID único. Dessa forma, o dataset inicial com 200 vídeos foi estruturado em 188 linhas, e o conjunto expandido com 500 vídeos, em 483 linhas, cada uma representando um vídeo específico.

Essa metodologia de modelagem, onde cada vídeo é representado por uma única linha independentemente do número de comentários, permite que os algoritmos de aprendizado de máquina realizem predições acerca da natureza educacional (ou não) dos vídeos com eficácia.

Com o dataset adequadamente organizado, podemos aplicar diversos algoritmos de aprendizado de máquina para realizar a classificação e recomendação de Objetos de Aprendizagem. Uma vez que o algoritmo é treinado, ele aprende a reconhecer os padrões usando esses dados. Para classificar um novo vídeo, seus comentários são coletados, passam pelo pré-processamento, são vetorizados seguindo a metodologia apresentada e, após esses processos, são enviados ao

algoritmo de aprendizado de máquina, que fornece a classificação do vídeo como educacional ou não educacional.

Inicialmente, priorizamos o uso de algoritmos interpretáveis, como as Árvores de Decisão e os baseados em Regras. Esta escolha foi motivada pela transparência desses modelos, que facilitam o entendimento de como as decisões de classificação são tomadas. Algoritmos baseados em árvores de decisão e regras foram escolhidos por sua clareza e facilidade de interpretação, permitindo não apenas classificar eficazmente os Objetos de Aprendizagem, mas também oferecer percepções valiosas sobre os padrões de interação dos usuários com o conteúdo dos vídeos. Essa abordagem contribui para uma compreensão mais profunda e prática dos dados analisados.

Com a implementação dessa metodologia, os algoritmos de aprendizado de máquina começaram a categorizar os objetos de aprendizagem como educacionais ou não, apresentando resultados promissores em termos de precisão na classificação. Os experimentos conduzidos revelaram que os algoritmos desenvolvidos são altamente eficazes para esse propósito. Contudo, durante o desenvolvimento do projeto, identificaram-se limitações relacionadas ao engajamento dos usuários em diferentes vídeos e restrições técnicas da *API* do YouTube, como a limitação de requisições e respostas de 50 vídeos por busca e o retorno de apenas 100 comentários por requisição.

Para superar essas limitações, considerou-se a introdução de evoluções na metodologia, permitindo que os objetos de aprendizagem fossem classificados de acordo com a classificação de cada um dos seus comentários e com diferentes "graus de certeza". Além disso, essa solução possibilita uma recomendação com diferentes quantidades de comentários e permite a realização de menos requisições. Dessa forma, os objetos de aprendizagem passaram a ser recomendados aos usuários acompanhados de uma indicação do grau de certeza sobre sua natureza educacional, enriquecendo a experiência de busca e seleção de conteúdos educativos.

4.6.2 Metodologia Flexível

Para alcançar um novo nível de classificação com “graus de certeza”, foi necessário realizar nova geração de datasets para serem utilizados pelos algoritmos de Aprendizado de Máquina. Apesar destas mudanças, a seleção dos vocábulos mais frequentes nos comentários permaneceu como a estratégia central para a classificação. Nesta variação, cada comentário é analisado individualmente, o que permite classificar o vídeo com diferentes níveis de certeza quanto à sua pertinência à categoria educacional.

O dataset desenvolvido para esta metodologia evoluída se diferencia daquele usado na classificação rígida, incluindo o “Id” do vídeo, as features derivadas dos vocábulos mais frequentes, e a classificação de cada comentário indicando se ele foi obtido através de um conteúdo educacional ou não. Foram preparados conjuntos de dados que abrangem 200 e 500 vídeos, respectivamente, e para cada tamanho de conjunto, foram desenvolvidos dois datasets específicos para a aplicação da metodologia flexível: um com 200 features e outro com 500 features. O conjunto de 200 vídeos incluiu 158.559 comentários, enquanto o conjunto maior, de 500 vídeos, compilou 738.653 comentários.

A Tabela 6 exemplifica como os vídeos são representados neste estudo. São apresentados

exemplos de comentários de dois vídeos distintos: um de natureza educacional e outro não educacional. Estes comentários já foram submetidos ao processo de pré-processamento, onde foram limpos e condensados em um formato estruturado adequado para análise.

Tabela 6 – Metodologia Flexível: Amostra dos comentários dos vídeos IDs "4g9JTQ2B6oo" e "s3Mb9qBEjO8" processados

Id	Comentários	Class
4g9JTQ2B6oo	car alucinadokkk melhor professorkkkkkkkkkkkk	1
4g9JTQ2B6oo	aul excel	1
4g9JTQ2B6oo	melhor profes hist	1
4g9JTQ2B6oo	ach nao kkkk	1
4g9JTQ2B6oo	coloqu veloc parec profes ta chap kkkkkkkkkkkkkkkkkkk	1
s3Mb9qBEjO8	absurd val ment nao ning atraz filh ta atraz mes var pesso tamb esta lut receb atraz	0
s3Mb9qBEjO8	tod noruc termin voz val diz ta tud bem dar direit trepl prq nao verdad val fal mid	0
s3Mb9qBEjO8	pesso morr send pouc temp	0
s3Mb9qBEjO8	gost notic	0
s3Mb9qBEjO8	not pra vinhet desan qualqu temp perd	0

Na tabela apresentada, cada linha representa um comentário específico, detalhando o "ID" do vídeo ao qual está associado, o conteúdo do comentário, e a classificação do vídeo ao qual pertence. Os comentários são simplificados para uma forma resumida que preserva informações essenciais enquanto remove elementos supérfluos e uniformiza o texto. Essa padronização é crucial para a eficácia dos algoritmos de aprendizado de máquina, facilitando uma avaliação consistente e precisa de cada comentário quanto à sua indicação de conteúdo educacional.

Após o pré-processamento, os comentários são submetidos a um processo de vetorização, sendo convertidos em formatos numéricos para classificação. Este processo é fundamental para preparar os dados para os algoritmos de aprendizado de máquina e é semelhante ao utilizado na metodologia rígida.

A principal diferença entre as metodologias reside na forma como os comentários são estruturados no conjunto de dados. Enquanto a metodologia rígida consolida todos os comentários de um vídeo em uma única linha, representando assim o vídeo de forma unificada, a metodologia flexível adota uma estratégia distinta, preservando a individualidade de cada comentário. Por exemplo, se um vídeo possui 10 comentários, o conjunto de dados na metodologia flexível refletirá isso através de 10 linhas distintas, cada uma correspondendo a um comentário vetorizado numericamente. Essa diferenciação é ilustrada na Tabela 7, que apresenta a vetorização da amostra mencionada anteriormente.

A transformação dos dados descrita anteriormente destaca a principal diferença entre as duas abordagens de modelagem. Na metodologia rígida, um vídeo com múltiplos comentários é condensado em uma única linha no dataset, como demonstrado pelo vídeo com o ID "_QS0oy42bWg", que tem 3 comentários e resulta em apenas uma linha no dataset final usado pelos algoritmos de aprendizado de máquina (AM). Em contraste, na metodologia flexível, cada

Tabela 7 – Metodologia Flexível: Amostra dos comentários dos vídeos IDs "4g9JTQ2B6oo" e "s3Mb9qBEjO8" considerando 10 features

IdVideo	aul	profes	obrig	vide	aprend	pra	ta	vc	faz	music	Class
4g9JTQ2B6oo	0	0	0	0	0	0	0	0	0	0	1
4g9JTQ2B6oo	1	0	0	0	0	0	0	0	0	0	1
4g9JTQ2B6oo	0	1	0	0	0	0	0	0	0	0	1
4g9JTQ2B6oo	0	0	0	0	0	0	0	0	0	0	1
4g9JTQ2B6oo	0	1	0	0	0	0	1	0	0	0	1
s3Mb9qBEjO8	0	0	0	0	0	0	1	0	0	0	0
s3Mb9qBEjO8	0	0	0	0	0	0	1	0	0	0	0
s3Mb9qBEjO8	0	0	0	0	0	0	0	0	0	0	0
s3Mb9qBEjO8	0	0	0	0	0	0	0	0	0	0	0
s3Mb9qBEjO8	0	0	0	0	0	1	0	0	0	0	0

comentário é tratado como uma entrada separada no dataset, mantendo a individualidade de cada feedback. Por exemplo, o vídeo "4g9JTQ2B6oo", que tem 5 comentários, continua a ser representado por 5 linhas no dataset.

Após a vetorização e conversão dos comentários para um formato numérico, os algoritmos de AM são aplicados para classificar os comentários individualmente. Diferente da metodologia rígida, que busca classificar um novo vídeo como educacional ou não baseando-se na agregação dos seus comentários, a metodologia flexível avalia cada comentário separadamente e retorna uma lista com a classificação de cada um. Isso requer o desenvolvimento de uma métrica específica para consolidar essas classificações individuais em uma classificação final para o vídeo, diferentemente da metodologia rígida, onde o retorno de um algoritmo de aprendizado de máquina já é a classificação direta de um vídeo como educacional ou não educacional.

Considerando a natureza individualizada do retorno de cada comentário na metodologia flexível, surge a necessidade de uma metodologia que agregue essas classificações para determinar a natureza educacional do vídeo como um todo. Para atender a essa necessidade, propõe-se o uso da equação 11 para calcular a probabilidade de um vídeo ser classificado como educacional.

A equação proposta para calcular $PEdu_{(vi)}$, a probabilidade de um vídeo *vi* ser educacional, leva em consideração os resultados da classificação de cada comentário associado ao vídeo.

$$PEdu_{(vi)} = \sum edu_{(vi)} / (\sum edu_{(vi)} + \sum non\_edu_{(vi)}) \tag{11}$$

onde:

- $\sum edu_{(vi)}$ é a soma de todos os comentários classificados como educacional no vídeo *vi*;
- $\sum non\_edu_{(vi)}$ é a soma de todos os comentários classificados como nao-educacional no vídeo *vi*.

Da mesma forma, pode-se calcular, usando 12, $PNaoEdu_{(vi)}$, que é a probabilidade de um vídeo *vi* ser não educacional,

$$PNaoEdu_{(vi)} = 1 - PEdu_{(vi)} \tag{12}$$

A Tabela 7, que apresenta comentários reais utilizados em nossa metodologia, mostra que alguns comentários são caracterizados por uma grande quantidade de zeros, indicando que poucas características (vocábulos identificados como relevantes) são encontradas nesses textos. Isso pode ocorrer devido à natureza breve ou fora de tópico de alguns comentários, que não engajam com o vocabulário central definido para a análise.

No entanto, quando consideramos o conjunto de todos os comentários associados a um vídeo, em combinação com a Fórmula 11 para o cálculo do Grau de Certeza, observamos uma excelente capacidade em classificar corretamente um vídeo como educacional. Esta metodologia tem mostrado resultados superiores à metodologia rígida, que consolida os comentários em uma única representação por vídeo.

A metodologia flexível apresenta vantagens significativas em termos de adaptabilidade e eficiência quando consideramos vídeos com poucos comentários ou a necessidade de limitar a quantidade de comentários utilizados na análise. Esta limitação é particularmente relevante em contextos onde a coleta de dados é custosa do ponto de vista computacional ou quando as restrições de API impõem limites ao número de solicitações que podem ser feitas ou quantidade de dados fornecidos, como é o caso da API do YouTube para coleta de comentários.

Limitar a quantidade de comentários analisados pode reduzir significativamente o custo computacional associado à busca e processamento de dados em grandes volumes. Isso é crucial em plataformas com grande fluxo de conteúdo e interações, onde a eficiência na gestão de recursos pode direcionar a escalabilidade do sistema de análise. Ao selecionar uma quantidade menor de comentários, a carga sobre o sistema de processamento de dados é reduzida. Isso não apenas acelera o processamento individual de cada vídeo, mas também otimiza o uso de recursos computacionais, permitindo que o sistema maneje um volume maior de dados de forma mais eficiente. A quantidade de dados que precisa ser baixada da API do YouTube é diretamente proporcional ao número de comentários solicitados. Reduzindo esse número, diminui-se também o número de requisições necessárias, o que é benéfico dado os limites de uso impostos pela API e potenciais custos associados a requisições excessivas.

Na metodologia rígida, um corte na quantidade de comentários utilizados exige não apenas um retreinamento dos algoritmos, mas também uma nova vetorização dos comentários. Isso deve-se ao fato de que a frequência e a representatividade dos vocábulos podem variar significativamente dependendo do volume de comentários. Algoritmos treinados para reconhecer padrões em datasets com um grande número de comentários podem não performar bem quando confrontados com uma quantidade muito menor de dados. Com menos dados, a frequência de certos termos tende a ser muito menor, o que influencia negativamente a corretude das classificações.

Há diferenças significativas entre os conjuntos de dados gerados pelas abordagens rígida e flexível. O conjunto da metodologia rígida é estruturado mapeando cada vídeo a uma linha única baseada nos vocábulos mais frequentes. Isso facilita o uso direto em algoritmos de aprendizado de máquina para classificar os vídeos como educacionais ou não, eliminando a necessidade de processamento adicional e possibilitando uma classificação rápida e eficaz com base nos padrões pré-identificados. Entretanto ela ainda necessita de passar por todos os comentários para verificar a contagem de um vocábulo característica.

Por outro lado, o conjunto de dados da metodologia flexível requer uma etapa adicional para

determinar a classificação de cada vídeo. Isso envolve classificar cada comentário individualmente e, em seguida, usar uma métrica ou fórmula específica para calcular a probabilidade de o vídeo ser educacional. Esse processo demanda uma análise mais detalhada e processamento adicional, aumentando o custo computacional.

No entanto, a metodologia flexível traz uma maior versatilidade e profundidade ao tratar cada comentário de forma individualizada. Isso permite que a classificação dos vídeos contemple nuances e diferentes graus de certeza quanto à sua natureza educacional, adaptando-se de forma mais eficaz a variados contextos e conteúdos de comentários. Essa metodologia enriquece a análise, possibilitando uma classificação mais flexível e evitando generalizações que poderiam comprometer a precisão dos resultados. Ademais, a metodologia flexível possui uma elevada capacidade de classificar vídeos com diferentes quantidades de comentários, o que é difícil para a metodologia rígida.

Neste estudo, para alcançar os objetivos propostos e superar as limitações técnicas, optou-se pela metodologia flexível no Sistema de Recomendação, LOIS. Essa abordagem se mostrou superior à metodologia rígida anteriormente utilizada e mais adequada para um contexto onde vídeos podem possuir grandes variações na quantidades de comentários dos usuários. Não apenas melhorou o desempenho em termos de Acurácia e F1-Score, mas também adicionou uma dimensão valiosa ao processo de seleção de materiais: o grau de certeza. Esse recurso informa o nível de confiança do sistema ao recomendar um vídeo, proporcionando aos usuários uma base mais sólida para tomar decisões informadas.

Importante ressaltar que a metodologia rígida foi fundamental no desenvolvimento do projeto. Servindo como ponto de partida, ela permitiu a evolução para a metodologia flexível, destacando áreas de melhoria e estabelecendo uma base sólida de conhecimento que potencializou as descobertas e inovações subsequentes.

A metodologia flexível oferece a vantagem de operar com uma quantidade reduzida de comentários, sem que a qualidade dos resultados seja comprometida, o que minimiza a dependência de dados extensivos da *API* e otimiza o uso do dataset completo. Além disso, ela permite recomendar vídeos independentemente da quantidade de comentários disponíveis, garantindo uma acurácia consistente tanto para vídeos com poucos quanto com muitos comentários. Essa abordagem individualizada na análise dos comentários reforça a eficiência da metodologia, tornando-a uma escolha estratégica para sistemas de recomendação que buscam equilibrar a otimização de recursos com a melhora da experiência do usuário.

4.7 Filtragem baseada em comentários

Após a modelagem e vetorização dos datasets, avançamos para a filtragem dos comentários visando à recomendação. Esta etapa crucial envolve a realização de experimentos para validar as abordagens propostas, onde diversos algoritmos de aprendizado de máquina são testados. Incluem-se nesta avaliação Árvores de Decisão, Random Forest, PART, JRIP, GenClust++, SVM e Redes Neurais — simples, profundas densas e convolucionais — para determinar qual oferece o melhor desempenho na classificação dos comentários.

Inicialmente, o Experimento 1 tinha como meta classificar vídeos educacionais. Para isso, empregamos a metodologia rígida e utilizamos algoritmos como JRIP, PART, J48, Random Forest e GenClust++. O propósito era entender o comportamento dos usuários frente a materiais educacionais e não educacionais, utilizando algoritmos baseados em Árvores e Regras. A inclusão do algoritmo de clusterização GenClust++ visava descobrir padrões ocultos nos dados.

O Experimento 2 focou em melhorar as classificações, mesmo que isso implicasse sacrificar a interpretabilidade dos modelos. Concentramos nossos esforços no desenvolvimento de algoritmos baseados em Redes Neurais — uma Rede Neural Simples, uma Rede Neural Profunda Densa e uma Rede Neural Convolucional — conhecidas por seu alto desempenho em certas condições.

No Experimento 3, buscávamos uma classificação mais flexível, usando comentários individualmente para determinar a categoria educacional dos vídeos. Essa abordagem levou à formulação da Fórmula de Classificação do Grau de Certeza, que utiliza a classificação de cada comentário para definir se um vídeo é educacional. Para o desenvolvimento inicial, escolhemos o Random Forest e o SVM.

O Experimento 4 foi criado para explorar e avaliar diferentes abordagens de vetorização dos comentários, comparando o método Bag of Words com nossa técnica que prioriza os vocábulos mais frequentes de cada classe. Para isso, utilizamos algoritmos como Random Forest, SVM e Redes Neurais, implementados via Scikit-Learn e TensorFlow.

Finalmente, o Experimento 5 comparou diretamente as abordagens Rígida e Flexível, utilizando o Random Forest, Rede Neural Simples, Rede Neural Profunda Densa e Rede Neural Convolucional. Este experimento foi mais abrangente e detalhado, revelando que a metodologia flexível tem potencial para superar a rígida, e destacou a Rede Neural Simples como a mais adequada para a recomendação de Objetos de Aprendizagem.

Os algoritmos foram testados com hiperparâmetros no modo padrão, exceto as Redes Neurais. As Redes Neurais, por sua vez, exigem um ajuste manual e contínuo de hiperparâmetros, um processo que não segue uma fórmula fixa e varia conforme o contexto de aplicação. Após vários testes, foram estabelecidos os modelos ideais para a Rede Neural Simples, a Rede Neural Profunda Densa e a Rede Neural Convolucional.

Para realizar a criação das Redes Neurais, foi utilizado o Tensorflow, sendo necessário definir a arquitetura da rede, que inclui a escolha do número de camadas, o número de neurônios em cada camada, a função de ativação e outros parâmetros.

A arquitetura da Rede Neural utilizada no experimento é demonstrada no código 4.1. A Rede Neural Profunda, utilizada no experimento é demonstrada no código 4.2. Por fim, a Rede Neural Convolucional é demonstrada no código 4.3

Listing 4.1 – Rede Neural

```
model = Sequential()
model.add(Dense(500, activation='relu', input_shape=(500,)))
model.add(Dense(1, activation='sigmoid'))
model.compile(optimizer='adam',
loss='binary_crossentropy', metrics=['accuracy'])
```

Essa é uma Rede Neural básica que contem duas camadas densas. A primeira camada densa, totalmente conectada, com 500 neurônios, com a função de ativação "relu" (Rectified Linear Unit) é aplicada a cada neurônio nesta camada, o que ajuda a introduzir não-linearidade no modelo. A segunda camada Dense é a camada de saída do modelo. Ela consiste em um único neurônio, pois a tarefa parece ser de classificação binária e sua função de ativação "sigmoid". Este modelo é usado para problemas de classificação binária, onde o objetivo é prever uma saída binária com base em um conjunto de características de entrada.

Listing 4.2 – Deep Neural Network

```
model = Sequential()
model.add(Dense(500, activation='relu', input_shape=(500,)))
model.add(Dropout(0.1))
model.add(Dense(10, activation='PReLU'))
model.add(Dense(10, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy',
optimizer='adam', metrics=['accuracy'])
```

Esta é uma Rede Neural Profunda com 5 camadas, incluindo camadas de dropout para evitar overfitting. A primeira camada é uma camada totalmente conectada com 500 unidades e uma função de ativação ReLU, seguida por uma camada de dropout com uma taxa de 0,1. As próximas 2 camadas também têm ativações PReLU e ReLU, e possuem 10 neurônios. Por fim, a camada final possui uma função de ativação sigmoide. O modelo é compilado com a função binária de perda de entropia cruzada e o otimizador Adam.

Listing 4.3 – Convolutional Neural Network

```
model = Sequential()
model.add(Conv1D(filters=128, kernel_size=1,
activation='relu', input_shape=(500,1)))
model.add(Dropout(0.2))
model.add(Conv1D(filters=64, kernel_size=1, activation='PReLU'))
model.add(Dropout(0.2))
model.add(Conv1D(filters=32, kernel_size=1, activation='relu'))
model.add(Dropout(0.2))
model.add(MaxPooling1D(pool_size=1))
model.add(Flatten())
model.add(Dense(200, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(20, activation='PReLU'))
model.add(Dropout(0.2))
model.add(Dense(10, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy',
```

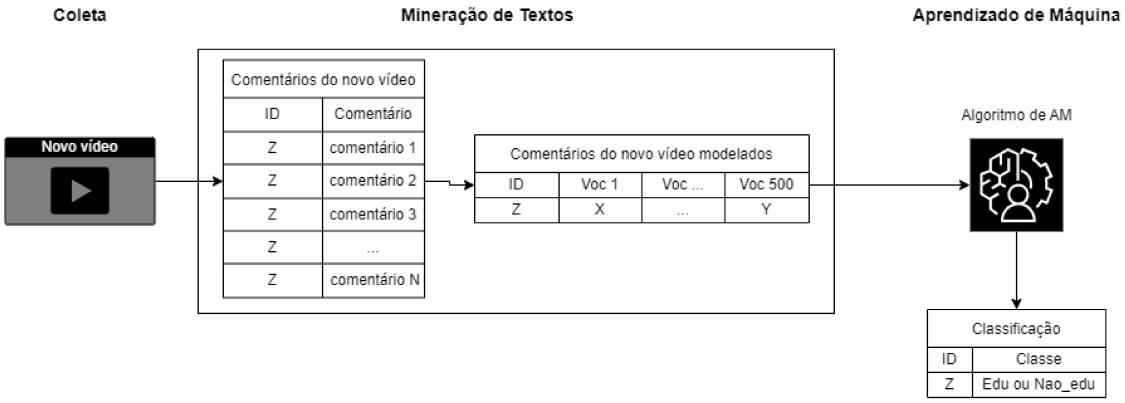


Figura 12 – Classificação de um novo vídeo segundo a metodologia “Rígida”.

```
optimizer='adam', metrics=['accuracy'])
```

Esta é uma rede neural convolucional projetada para dados 1D, com três camadas convolucionais seguidas por camadas de dropout e max-pooling. A primeira camada tem 128 filtros, um tamanho de kernel de 1 e uma função de ativação ReLU. A segunda camada tem 64 filtros, um tamanho de kernel de 1 e uma função de ativação Parametric ReLU. A terceira camada tem 32 filtros, um tamanho de kernel de 1 e uma função de ativação ReLU. Cada camada convolucional é seguida por uma camada dropout com uma taxa de 0,2. A camada de pooling máximo tem um tamanho de pool de 1, o que reduz o tamanho dos mapas de recursos. A saída achatada da camada de max-pooling é passada para duas camadas totalmente conectadas, com 200 e 10 unidades, respectivamente, e ambas com camadas de dropout. A camada de saída final tem uma função de ativação sigmoide. O modelo é compilado com a função binária de perda de entropia cruzada e o otimizador Adam.

A performance dos algoritmos é medida usando métricas padrão como Acurácia, Precisão, Recall e F1-Score, com especial atenção para Acurácia e F1-Score para assegurar a precisão e a generalidade dos resultados. A técnica de 10-folds cross validation é empregada para validar essas métricas, proporcionando uma análise robusta dos algoritmos em diferentes segmentos do dataset.

Durante os testes, foram foram realizados experimentos utilizados datasets configurados tanto para a metodologia rígida quanto para a flexível. Para ilustrar e demonstrar melhor como funcionaria a utilização das duas abordagens, as Figuras 12 e 13 ilustram de forma simplificada os processos de classificação de um novo vídeo para as abordagens “rígida” e “flexível”, respectivamente.

Quando um novo vídeo, identificado como “Z”, é selecionado, inicia-se a coleta de seus comentários. Após essa etapa, segue-se para a mineração de textos, que inclui o pré-processamento e a modelagem dos comentários para uso subsequente. Os comentários processados são então vetorizados, utilizando até 500 características distintas, conforme demonstrado anteriormente e ilustrado na figura. Essa é a quantidade de características definida para o projeto após o desenvolvimento e a realização de testes com outras quantidades.

Os comentários são vetorizados de modo que o vídeo seja representado por apenas uma linha

Figura 13 – Classificação de um novo vídeo segundo a metodologia “flexível”.

no dataset, onde cada valor nesta linha reflete a frequência com que cada vocábulo apareceu nos comentários coletados. Estes dados de frequência são então enviados ao algoritmo de aprendizado de máquina, que classifica o vídeo como educacional ou não educacional. Como o vídeo é representado por uma única linha, o algoritmo fornece uma classificação direta para esse vídeo, baseada na análise agregada de seus comentários.

Quando um novo vídeo, identificado como “Z”, é selecionado, o processo é similar com a coleta de comentários seguida pela mineração de textos, que envolve o pré-processamento e vetorização dos comentários.

Diferentemente da metodologia rígida, na metodologia flexível, cada comentário é convertido em uma linha no dataset. Por exemplo, se o vídeo tem 5 comentários, resultará em 5 linhas vetorizadas no dataset. Todos os comentários vetorizados são então enviados ao algoritmo de aprendizado de máquina, que classifica cada comentário individualmente como educacional ou não educacional. Assim, a saída do algoritmo é uma lista com a classificação de cada comentário.

Adicionalmente, e diferentemente da metodologia rígida, a metodologia flexível exige um passo adicional onde o “Grau de Certeza” é calculado usando a Fórmula 11. Com base nesse cálculo, o sistema determina se o vídeo é educacional ou não educacional. Este método proporciona uma análise mais detalhada e personalizada de cada vídeo, levando em conta a diversidade de opiniões expressas nos comentários.

Ao longo da execução do trabalho, optou-se pela metodologia flexível, implementando uma Rede Neural Simples no Sistema de Recomendação LOIS, que realiza recomendações em tempo real. O resultado final deste projeto é o desenvolvimento do sistema, ele sintetiza todas as etapas anteriores, desde a preparação dos dados até a aplicação dos modelos de aprendizado de máquina, resultando em uma ferramenta eficaz de recomendação.

Originalmente chamado de SysVidEduc, o sistema foi projetado para analisar os comentários dos vídeos do YouTube e identificar seu conteúdo educativo. Os usuários podem facilmente buscar por vídeos usando uma interface simples, que processa expressões de busca ou IDs de vídeos e retorna uma lista de vídeos classificados por sua relevância educacional. A primeira

etapa de desenvolvimento baseou-se no algoritmo Random Forest, escolhido por sua acurácia no primeiro experimento. O desenvolvimento foi realizado em Python, com suporte das bibliotecas string, unidecode, NLTK, e re para pré-processamento de texto, e sklearn, pandas e joblib para classificação de vídeos. Bibliotecas necessárias para a comunicação com a API do YouTube também foram integradas.

Durante o desenvolvimento, o sistema, que foi renomeado para LOIS (Learning Object Intelligent Search), evoluiu significativamente com a implementação de uma Rede Neural Simples em sua segunda versão. Este modelo não apenas superou o desempenho do algoritmo Random Forest utilizado inicialmente, mas também alcançou resultados comparáveis a modelos mais complexos, como a Rede Neural Profunda Densa e a Rede Neural Convolucional (CNN). A Rede Neural Simples destacou-se particularmente pela sua eficiência em termos de velocidade de treinamento e classificação, beneficiando-se de sua estrutura mais enxuta, que inclui apenas uma camada de processamento e uma de saída. A integração com o framework Tensorflow foi crucial, facilitando a implementação de modelos de Redes Neurais e mantendo a compatibilidade com as bibliotecas utilizadas anteriormente para outras funcionalidades do sistema.

Em termos operacionais, o funcionamento do LOIS está ilustrado na figura 14, com uma descrição detalhada dos passos envolvidos, apresentando-se como uma solução completa e integrada para a recomendação de conteúdo educativo em vídeo.

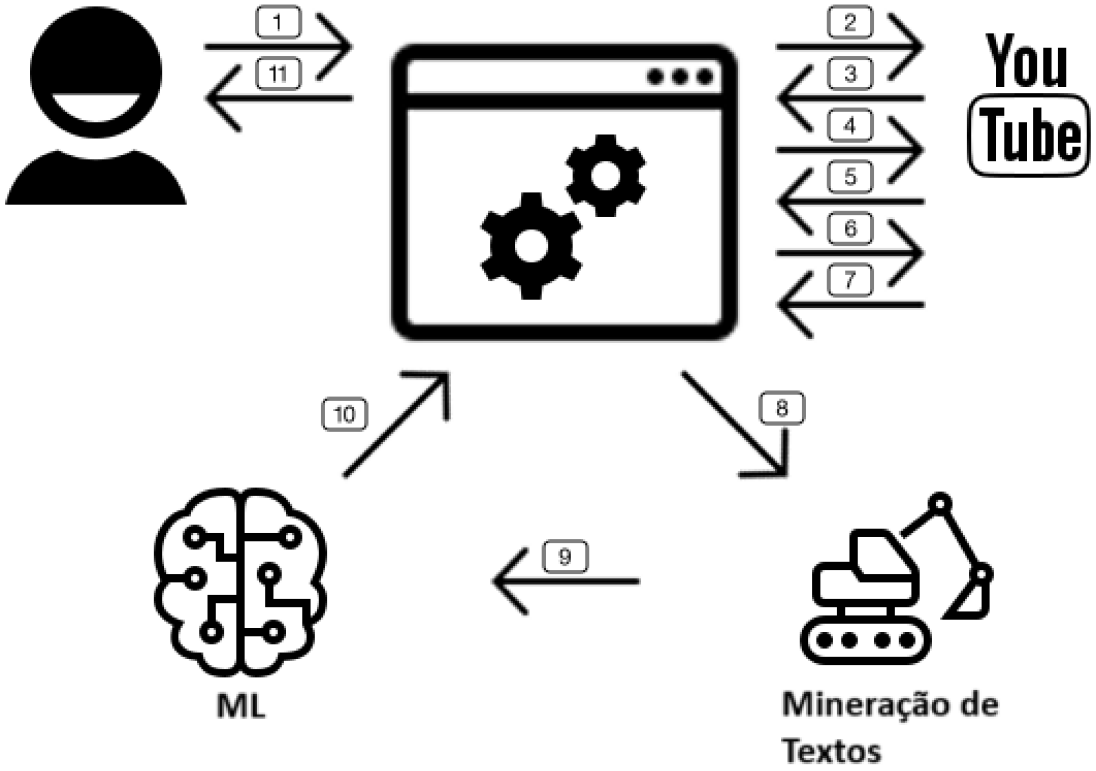


Figura 14 – LOIS: Modelo de funcionamento do sistema

1. Acesso ao sistema, acesso para busca pode ocorrer de duas maneiras: O usuário acessa o sistema e fornece a expressão de busca ou uma requisição, por um aplicativo externo,

através de um JSON;

2. O sistema realiza a busca na API do Youtube;
3. A API retorna os vídeos da busca (nomes dos vídeos e *ids*);
4. O sistema consulta novamente a API para obter os metadados dos vídeos;
5. A API responde com informações (categoria do vídeo, número de *like* e *dislike*, número de visualizações e comentários);
6. O sistema faz uma nova solicitação de aquisição de comentários. A API do Youtube possui uma restrição de retornar apenas 100 comentários a cada solicitação, devendo, dessa forma, realizar outras requisições para coletar o restante dos comentários;
7. API adquire comentários de vídeos;
8. Os comentários são enviados para o módulo de Mineração de Texto. Nesta etapa é realizado o pré-processamento dos dados, com etapas de retirada de pontuação, espaços, e-mails, links, etc. Adicionalmente, esta etapa prepara a base de dados a ser utilizada pelo módulo de Inteligência Artificial;
9. Os dados são enviados para o algoritmo de AM. O algoritmo supervisionado de Aprendizado de Máquina (um classificador que pode ser facilmente trocado) que desenvolverá um modelo de classificação de vídeo usando o banco de dados preparado no módulo Text Mining;
10. O sistema classifica os vídeos de acordo com a metodologia de classificação proposta;
11. Os vídeos, agora classificados, são apresentados ao usuário ou retornados via JSON;

Ao utilizar o sistema para buscar um termo específico na etapa 2, o processo realiza a busca segue para a obtenção de comentários relacionados a esse termo, nas etapas 2 a 7. Após coletar os comentários, eles são encaminhados para o módulo de mineração de textos, na etapa 8. Neste ponto, todos os comentários são processados e vetorizados com base nos 500 vocábulos mais frequentes identificados. Este preparo é essencial para a próxima fase, onde os vídeos processados são enviados ao algoritmo de Aprendizado de Máquina, na etapa 9.

Nessa etapa, um algoritmo baseado em uma rede neural simples analisa cada comentário que foi vetorizado em 500 características, correspondentes aos vocábulos mais frequentes. Este algoritmo classifica cada comentário como educacional ou não educacional. Baseado nesta classificação, o "Grau de Certeza" é calculado para cada vídeo. Se este grau atende ao limiar pré-definido de confiança, o vídeo é então categorizado como educacional e encaminhado de volta ao sistema, na etapa 10.

Finalmente, na etapa 11, os vídeos classificados como educacionais são retornados ao usuário. Este fluxo não apenas garante uma metodologia sistemática para filtrar e classificar conteúdo educativo, mas também otimiza a relevância e a precisão das recomendações oferecidas pelo sistema.

Para garantir ampla acessibilidade, o LOIS foi desenvolvido para funcionar via web, facilitando a integração com Ambientes Virtuais de Aprendizagem através de requisições web que retornam dados em formato JSON, incluindo metadados relevantes. Além disso, uma interface web foi disponibilizada para os usuários finais, permitindo que realizem buscas diretamente no sistema para encontrar Objetos de Aprendizagem de maneira intuitiva.

Para validar a eficácia do LOIS, estudantes voluntários foram convidados a avaliar o sistema por meio do questionário ResQue, uma ferramenta utilizada para avaliar a qualidade das recomendações fornecidas por sistemas de recomendação. O questionário, baseado no modelo proposto por Pu, Chen e Hu (2011), é composto por afirmações avaliadas em uma escala *Likert*, abordando aspectos como a adequação das recomendações aos interesses dos usuários, a usabilidade do sistema e a relevância das recomendações. Além das perguntas padrão, foi incluída uma questão sobre o tempo aceitável para receber uma recomendação e um campo para observações adicionais. O questionário foi aplicado a mestres e mestrandos do programa de Mestrado em Educação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, no período de 2019 a 2024.

Os feedbacks coletados através do ResQue forneceram ideias sobre a percepção dos usuários em relação ao LOIS, destacando pontos fortes do sistema e áreas para melhoria. A análise das respostas permitiu uma compreensão abrangente da experiência do usuário com o sistema de recomendação, contribuindo para o aprimoramento contínuo do LOIS.

Em suma, tanto o SysVidEduc quanto sua evolução, LOIS, representam avanços significativos na recomendação de conteúdo educacional online, oferecendo ferramentas poderosas para a filtragem e classificação de vídeos baseadas em comentários. A transição para um modelo de rede neural na segunda versão e a avaliação positiva do sistema pelos usuários reforçam o potencial do LOIS como uma solução eficaz no âmbito da educação digital.

Capítulo 5

Experimentos e Análise dos Resultados

No capítulo 4, apresentamos a abordagem desenvolvida de acordo com a execução do projeto e nesse capítulo serão detalhadas as etapas de Análise dos dados, Desenvolvimento das Abordagens, Experimentos realizados e o LOIS. Esse capítulo será dividido em seções e em cada seção apresentaremos os resultados e abordaremos as discussões relevantes.

Começaremos nossa exposição analisando os dados coletados, para depois discutirmos as metodologias desenvolvidas e detalhar a execução dos experimentos realizados. Concluiremos apresentando o sistema que foi desenvolvido, consolidando as percepções e as tecnologias empregadas ao longo do estudo. Esta estrutura oferece uma visão completa do processo de pesquisa, desde a concepção teórica e metodológica até a implementação prática e divulgação científica dos achados.

5.1 Análise dos dados

O projeto iniciou com a utilização de um dataset de 200 vídeos, distribuídos igualmente entre categorias educacionais e não educacionais. Posteriormente, para aprofundar a análise, o dataset foi expandido para abranger 500 vídeos, mantendo a mesma proporção de categorização.

No dataset inicial, que também foi o foco do primeiro experimento, havia uma paridade entre os vídeos educacionais e não educacionais, totalizando 200 vídeos. Destes, foram coletados 158.559 comentários, divididos entre 77.095 comentários de vídeos educacionais e 84.464 de vídeos não educacionais. A análise lexical desses comentários revelou um total de 2.262.903 vocábulos, sendo 1.002.070 associados aos vídeos educacionais e 1.260.833 aos não educacionais.

Com a expansão do dataset para 500 vídeos, igualmente divididos entre educacionais e não educacionais, o volume de comentários aumentou significativamente para 738.653, dos quais 201.792 foram classificados como educacionais e 536.861 como não educacionais. O número de vocábulos também cresceu, alcançando um total de 7.336.718, com 2.212.198 vocábulos derivados dos comentários dos vídeos educacionais e 5.124.520 dos vídeos não educacionais.

A Figura 15 ilustra de forma visual a quantidade de vídeos coletados e analisados, oferecendo uma perspectiva clara sobre a escala do dataset e a distribuição entre as categorias analisadas.

O Gráfico 15 ilustra a distribuição dos vídeos coletados para análise, enfatizando o esforço em

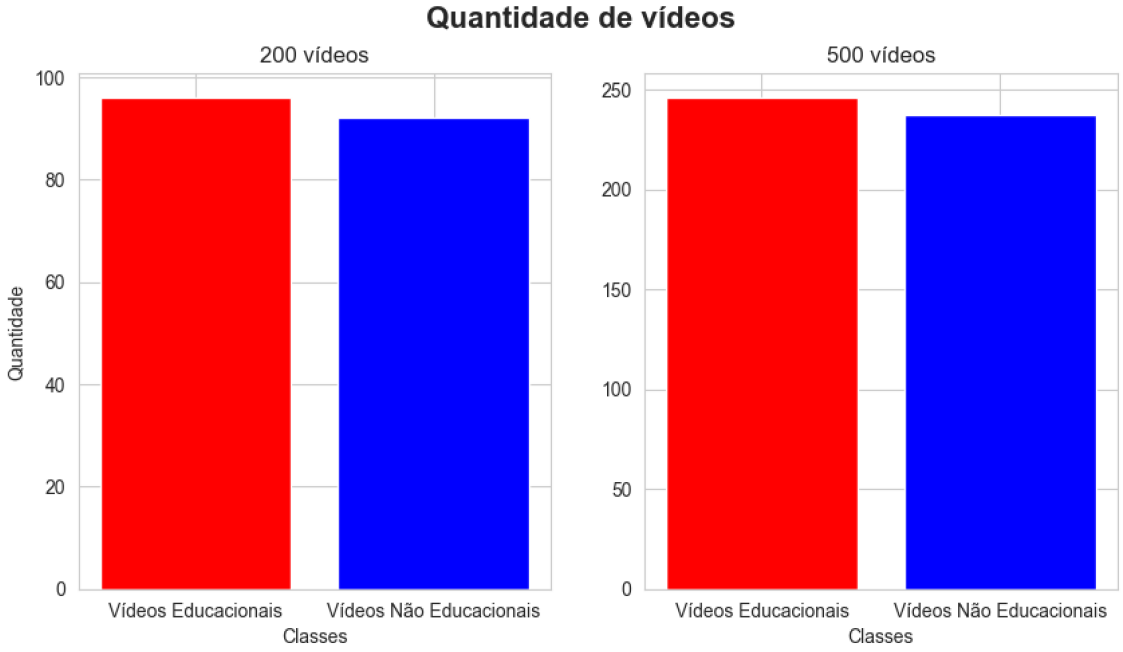


Figura 15 – Vídeos coletados

manter um equilíbrio entre as duas classes de vídeos: educacionais e não educacionais. Durante o processo de coleta, alguns vídeos não foram apresentados pois não foi possível obter os seus comentários, seja por falta de comentários ou comentários desabilitados. Esse cuidado na seleção assegura uma comparação justa e balanceada entre as categorias, refletindo a intenção de criar um dataset representativo de ambas as classes.

A tentativa de equilibrar a quantidade de vídeos de cada classe no dataset é uma prática importante para garantir a imparcialidade e a eficácia dos resultados da análise. A Figura 16 especificamente detalha o volume de comentários coletados por classe, além de apresentar a média de comentários por vídeo, baseada na coleta inicial de 200 vídeos. Essa informação é crucial para entender a interação dos usuários com os vídeos em cada categoria, oferecendo percepções valiosas sobre o engajamento e a receptividade do conteúdo por parte do público.

O Gráfico 16, apresenta a quantidade de comentários que foram obtidos em cada classe e o segundo apresenta a média de comentários por classe. Percebe-se um equilíbrio entre as quantidades de comentários de ambas as classes, propiciando um conjunto de dados bem balanceado. Este equilíbrio é crucial para garantir a imparcialidade na análise subsequente. Esse dataset possui 154.832 comentários, sendo 75.660 comentários educacionais e 79.172 comentários não educacionais, além disso, em média, cada vídeo educacional possui 788 comentários e cada vídeo não educacional possui 860 comentários. Percebe-se um bom valor com relação as médias entre os vídeos educacionais.

Entretanto, percebe-se que alguns, poucos vídeos, possuem um valor superior a essa média de comentários, um exemplo é o vídeo de Id “_kzTFOzf-_w” com o nome de “Orações Subordinadas Substantivas [Prof Noslen]” que na data da coleta possuía 11.272 comentários, e, atualmente, em Fevereiro de 2024 já apresenta 17.511 comentários. Adicionalmente, verificou-se a quantidade de vídeos que apresentam quantidade de comentários acima da média e apenas 15

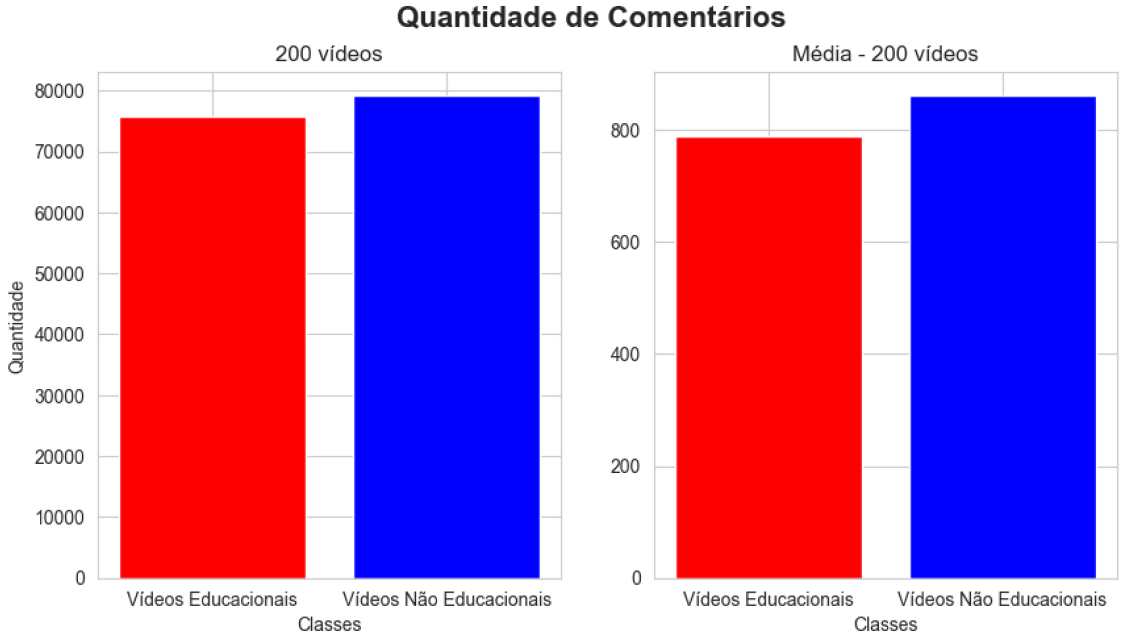


Figura 16 – Dados: Comentários obtidos nos 200 vídeos

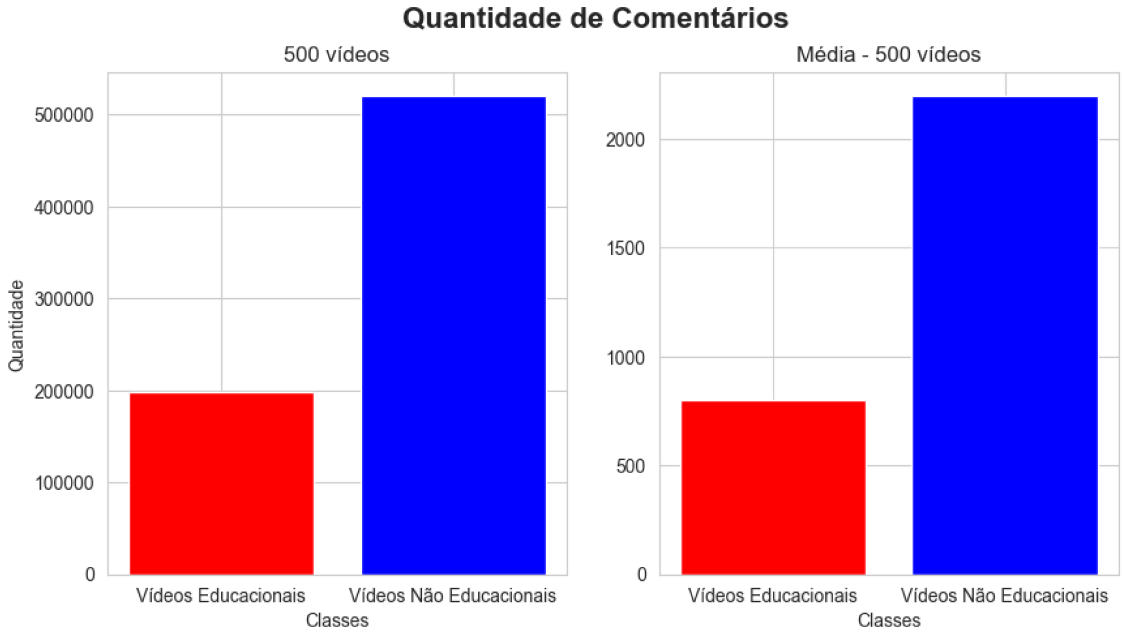


Figura 17 – Dados: Comentários obtidos nos 500 vídeos

vídeos possuem mais de 788 comentários e 82 vídeos apresentam menos que essa quantidade de comentários.

Ao expandir o conjunto de dados para incluir 500 vídeos, observou-se uma variação notável na quantidade de comentários entre as categorias de vídeos educacionais e não educacionais. Esta discrepância é ilustrada na Figura 17, que destaca as diferenças no engajamento dos usuários com os vídeos de cada categoria. A análise dessas variações é essencial para compreender o comportamento do usuário e para a classificação eficaz do conteúdo.

O Gráfico 17, demonstra uma grande diferença entre a quantidade de comentários entre ambas as classes, enquanto os comentários educacionais são compostos por 197.673 comentários, os vídeos não educacionais possuem 520.375. Isso indica que, em média, cada vídeo educacional recebeu cerca de 804 comentários, enquanto cada vídeo não educacional recebeu, em média, 2.196 comentários. Esta diferença substancial na quantidade de comentários reflete distintos níveis de interação e engajamento dos usuários com os vídeos de cada categoria, fornecendo percepções valiosas para a análise comportamental e metodologias de classificação do conteúdo.

A diferença marcante observada reforça a ideia de que vídeos educacionais costumam receber menos interações que vídeos não educacionais, uma tendência que sublinha um dos desafios centrais enfrentados neste estudo: a necessidade de lidar com dados reais influenciados diretamente pelo comportamento variável dos usuários. No entanto, este obstáculo é contornado pelas metodologias implementadas, em particular pela escolha cuidadosa dos termos mais frequentes identificados em cada categoria. Esta estrategia permite uma classificação mais precisa, superando as dificuldades trazidas pela variabilidade nas interações dos usuários.

Devido a expansão do dataset, tornou-se necessário compreender se o comportamento dos usuários perante a vídeos educacionais e não educacionais mantinha-se constante ou se apresentava grandes variações. Esta investigação é crucial para determinar se os padrões observados permanecem consistentes e, portanto, aplicáveis em futuras fases de classificação do conteúdo.

Levando em consideração o dataset inicial, os resultados dos comentários mais frequentes nos vídeos são apresentados na tabela 8. Por sua vez, a tabela 9 apresenta os vocábulos mais frequentes nos comentários dos vídeos.

Tabela 8 – Dados: Comentários mais frequentes nos vídeos (200 vídeos).

Termos mais frequentes			
Educacional		Não Educacional	
Termo	Quantidade	Termo	Quantidade
muito bom	572	top	152
melhor professor	291	oi	117
ótima aula	138	primeiro	97
obrigado	127	ah é	95
amei	126	muito bom	91
eu	113	eu	61
show	103	salve	58

Fonte: Obtido de (CARVALHO et al., 2020b)

Observa-se que os vídeos não educacionais apresentam termos como risadas, que não foram incluídos na tabela 8 por serem irrelevantes ao estudo, inicialmente. Neste sentido, apontam-se que os termos “kkkkk” (cinco k’s) e “kkkkkk” (seis k’s), figuram, respectivamente, 65 e 59 vezes, e ocupariam, respectivamente, a sexta e oitava posição.

Ao analisar a tabela 8, percebem-se as diferenças entre os termos mais frequentes. O único termo utilizado nos comentários de ambas as classes de vídeos é “muito bom”, sendo utilizado 572 vezes nos cometários dos vídeos educacionais, e 91 vezes nos comentários dos vídeos não educacionais. Apesar de estar presente nos comentários de ambas as classes de vídeos, o número

Tabela 9 – Dados: Vocábulos mais frequentes nos comentários dos vídeos (200 vídeos).

Vocábulos mais frequentes			
Educacional		Não Educacional	
Vocábulo	Quantidade	Vocábulo	Quantidade
profes	16.421	pra	6.363
aul	11.828	faz	5.666
obrig	7.544	músic	4.020
melhor	6.183	vai	3.723
ajud	6.087	tod	3.658
vc	5.073	melhor	3.411
aprend	4.974	vc	3.365

Fonte: Obtido de (CARVALHO et al., 2020b)

de vezes em que este termo foi utilizado é bastante diferente: mais de 6 vezes nos comentários dos vídeos educacionais quando comparada à utilização nos comentários dos vídeos não educacionais.

Outras palavras como “amei”, “eu” e “show”, apesar de figurarem mais frequentemente nos comentários dos vídeos educacionais, podem não ser muito indicadas para a utilização para a classificação, por serem muito genéricas. Já os termos como “melhor professor” e “ótima aula”, por serem mais específicos ao contexto educacional, podem ser indicados para a categorização de um vídeo.

Na tabela 9, percebe-se que as únicas palavras que figuram tanto nos comentários dos vídeos educacionais quanto nos cometários dos vídeos não educacionais são “melhor” e “vc”. Contudo, mesmo presente em ambas as classes de vídeos, o número de vezes em que figuram é muito diferente: “melhor” foi utilizado 6183 vezes, e “vc” 5073 vezes nos comentários dos vídeos educacionais, em comparação a, respectivamente, 3411 vezes e 3365 vezes nos comentários dos vídeos não educacionais.

Os radicais “profes”, “aul”, “obrig”, “ajud”, e “aprend” figuram apenas nos comentários dos vídeos educacionais, sendo, portanto, potencialmente adequados para serem utilizados para a classificação de um vídeo.

Levando em consideração o dataset expandido, 500 vídeos, os resultados dos comentários mais frequentes nos vídeos são apresentados na tabela 10.

Tabela 10 – Dados: Comentários mais frequentes nos vídeos (500 vídeos).

Termos mais frequentes			
Educacional		Não Educacional	
Termo	Quantidade	Termo	Quantidade
muito bom	1.500	amei	1.526
curso de excel	742	oi	1.153
melhor professor	735	pocahontas	1.023
amei	561	up	933
otima aula	433	top	677
eu	392	muito bom	607
obrigado	379	maravilhosa	561

Fonte: Elaborado pelo autor, com base na pesquisa realizada.

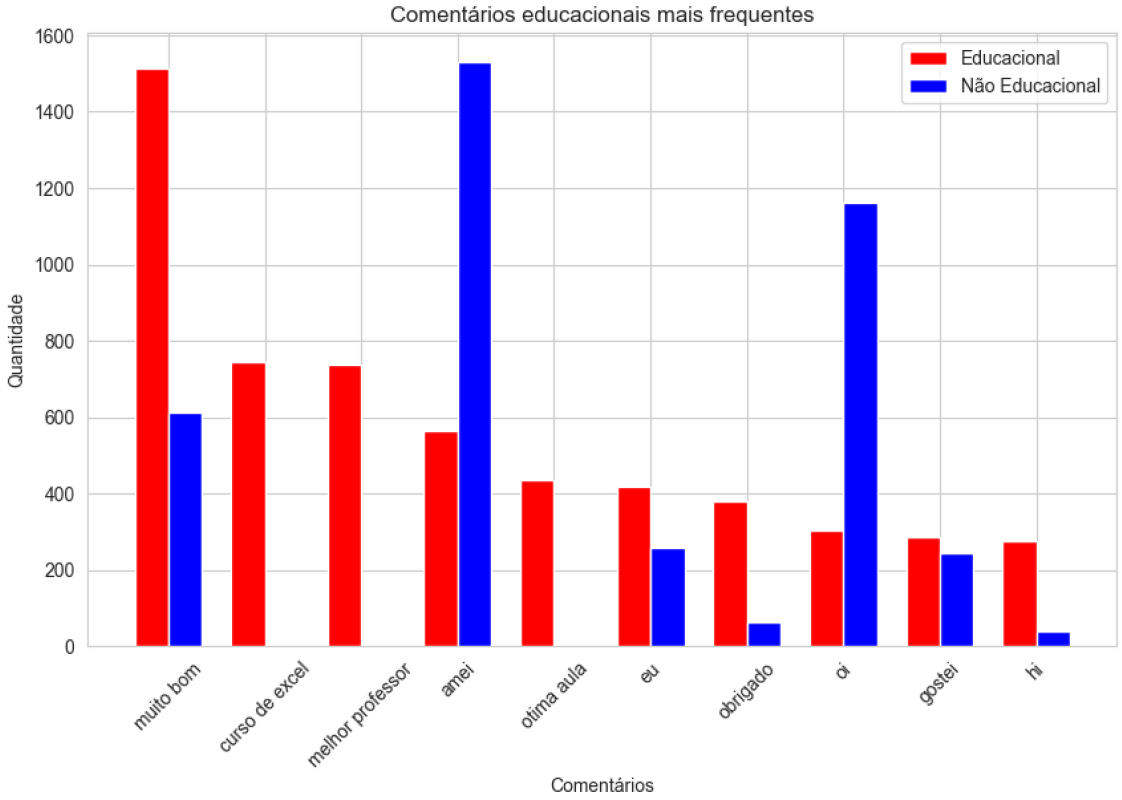


Figura 18 – Dados: Frequência dos comentários educacionais (500 vídeos)

Observa-se que os vídeos não educacionais apresentam termos como risadas, que não foram incluídos na tabela 10 por serem irrelevantes ao estudo, inicialmente. Neste sentido, apontam-se que os termos “kkkkk” (cinco k’s), “kkkk” (quatro k’s), “kkkkkk” (seis k’s) e “kkkkkkk” (sete k’s), figuram, respectivamente, 787, 603, 601 e 540 vezes, e ocupariam, respectivamente, a quinta, sétima, oitava e decima posição.

As Figuras 18 e 19 apresentam, graficamente, a frequência dos comentários mais presentes nos vídeos educacionais e não educacionais, respectivamente.

Observa-se uma alta frequência do comentário “melhor professor” nos vídeos educacionais, sendo que ele aparece uma vez em um vídeo não educacional. O vídeo com o ID UkhSLsDgj4M, intitulado “Você REALMENTE quer MEDICINA?”, foi criado por Paulo Jubilut e tem como objetivo fornecer uma reflexão sobre o curso de medicina, auxiliando as pessoas a avaliar seu interesse e vontade de seguir essa carreira. Embora esse vídeo não tenha uma abordagem estritamente educacional, ele apresenta uma ideia de reflexão. O comentário “melhor professor!” ocorre uma vez nesse vídeo e as palavras “melhor professor” também são encontradas em outros comentários, como por exemplo:

- Você é o melhor professor! Eu fiz Ciências Contábeis e Administração, agora vou fazer Medicina
- Melhor professor de biologia. Conselhos maravilhosos
- Jubi vc é, de longe, o melhor professor que já vi, por favor não pare.

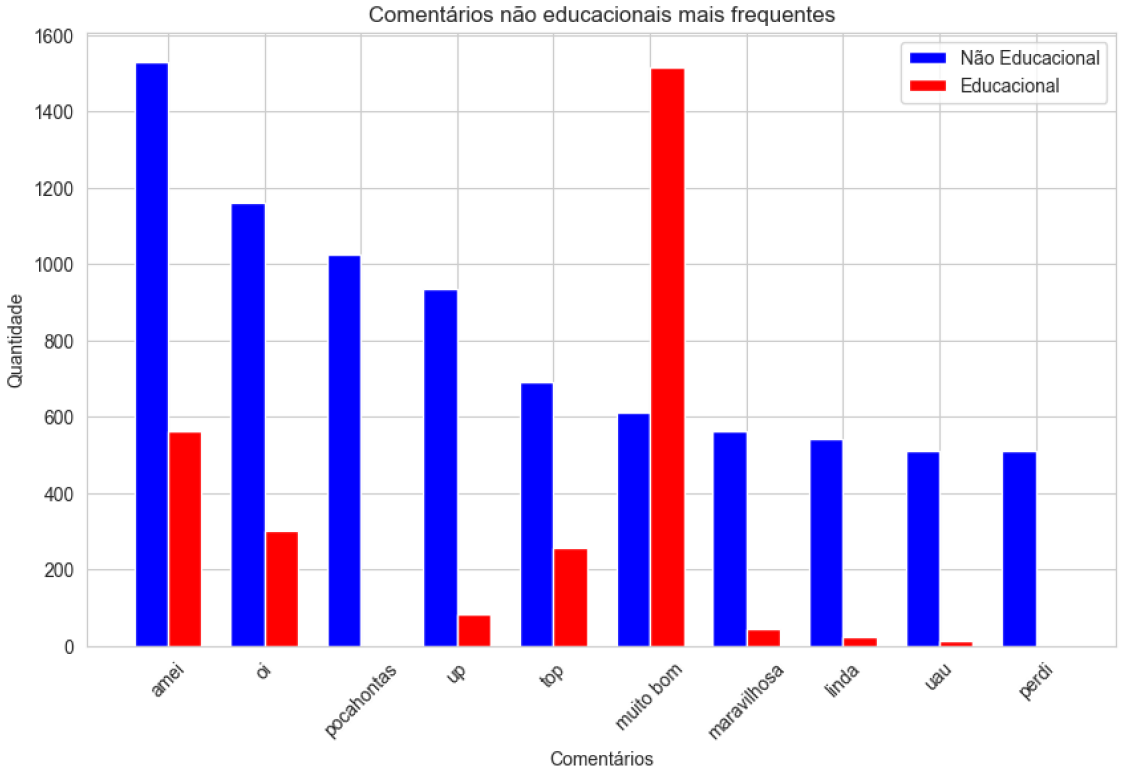


Figura 19 – Dados: Frequência dos comentários não educacionais (500 vídeos)

No entanto, é importante utilizar com cautela essa abordagem de análise que considera todo o comentário, devido às limitações de classificação e modelagem de texto. É fundamental ponderar sobre as melhores maneiras de utilizar os textos e considerar outras formas de modelagem para obter resultados mais confiáveis.

Durante as análises, percebeu-se que termos contendo “risadas” (kkk’s) são bem frequentes em vídeos não educacionais. Inicialmente, tivemos dúvidas quanto a utilização desses termos, porém, percebe-se que os vídeos não educacionais tendem a ser mais “descontraídos” e isso possibilita que as pessoas se expressem mais livremente assim como comentar apenas com risadas. A Figura 20 apresenta como as risadas aparecem nos vídeos educacionais e não educacionais.

A Tabela 11 apresenta os vocábulos mais frequentes nos comentários dos vídeos.

Observa-se que apesar do aumento do número de vídeos e comentários presentes neste trabalho em relação ao Carvalho et al. (2020b), os principais resultados de ambas as pesquisas são semelhantes. Nesse sentido, os vocábulos referentes aos vídeos educacionais, “profes”, “aul”, “obrig”, “aprend”, “ajud”, e “vc” estão presentes nos dois estudos, enquanto o vocábulo “vide” apareceu apenas na presente pesquisa. Por sua vez, os vocábulos “pra”, “faz”, “music”, “vai”, “vc”, figuram como vocábulos mais frequentes dos vídeos não educacionais em ambos os estudos. Nota-se que, apesar de alguns termos estarem presentes em ambas as classes de vídeos, a exemplo de “vide”, a diferença entre a frequência da utilização nos comentários é considerável (15.549 vezes em vídeos educacionais e 27.556 vezes em vídeos não educacionais).

As Figuras 21 e 22 apresentam, graficamente, a frequência dos vocábulos mais frequentes nos vídeos educacionais e não educacionais, respectivamente.

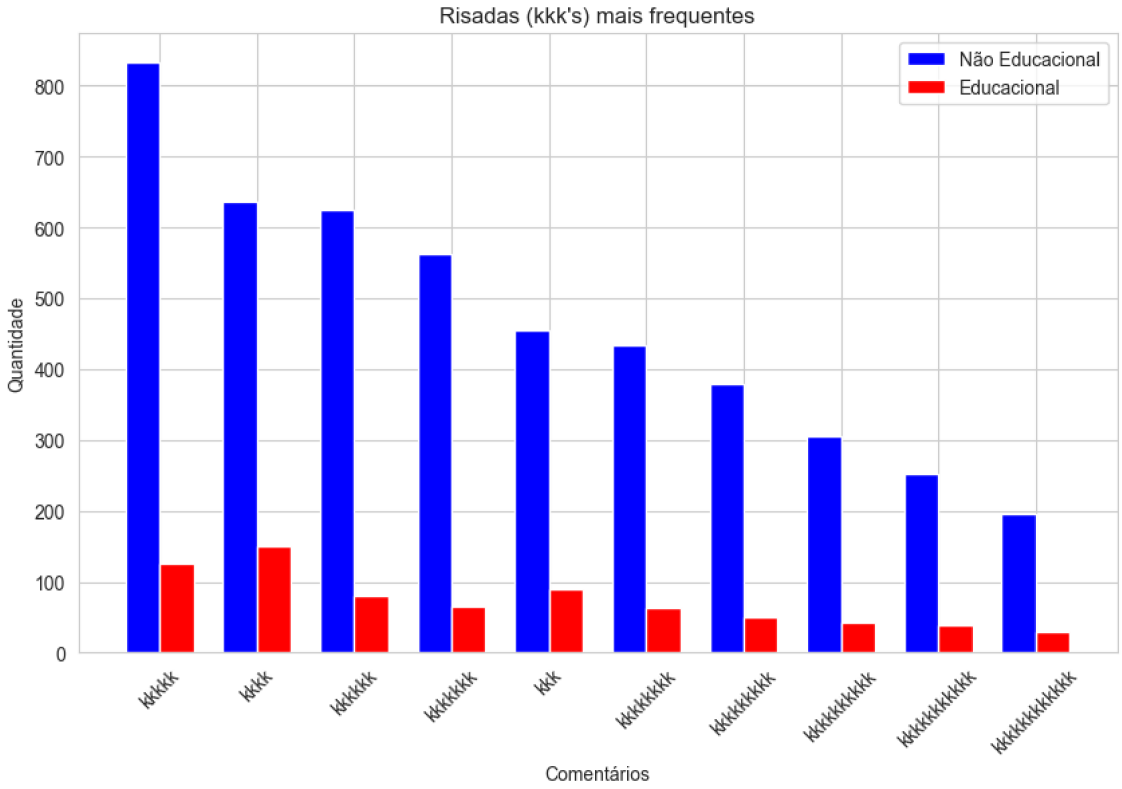


Figura 20 – Dados: Frequência das risadas (500 vídeos)

Tabela 11 – Dados: Vocábulos mais frequentes nos comentários dos vídeos (500 vídeos).

Vocábulos mais frequentes			
Educacional		**Não Educacional**	
Vocábulo	**Quantidade**	**Vocábulo**	**Quantidade**
aul	28.748	pra	37.049
profes	28.369	ta	31.133
obrig	17.119	vc	30.530
vide	15.549	faz	29.076
aprend	14.761	vide	27.556
ajud	12.550	music	25.634
vc	11.791	vai	21.357

Fonte: Elaborado pelo autor, com base na pesquisa realizada.

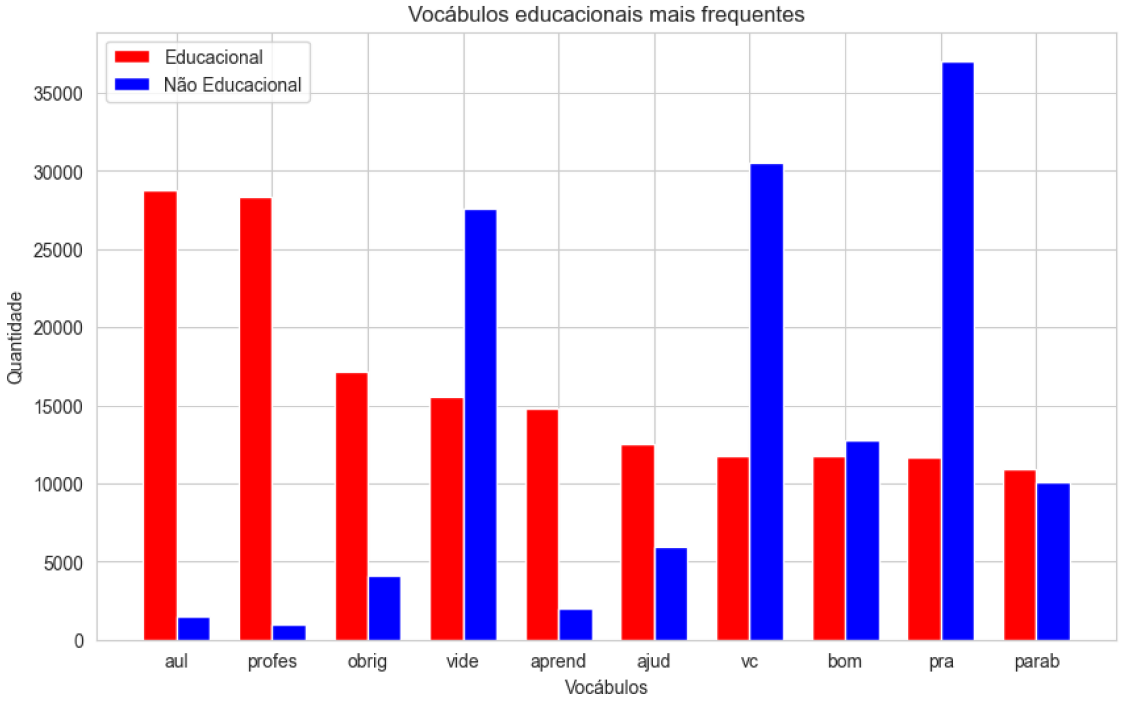


Figura 21 – Dados: Frequência dos vocábulos educacionais (500 vídeos)

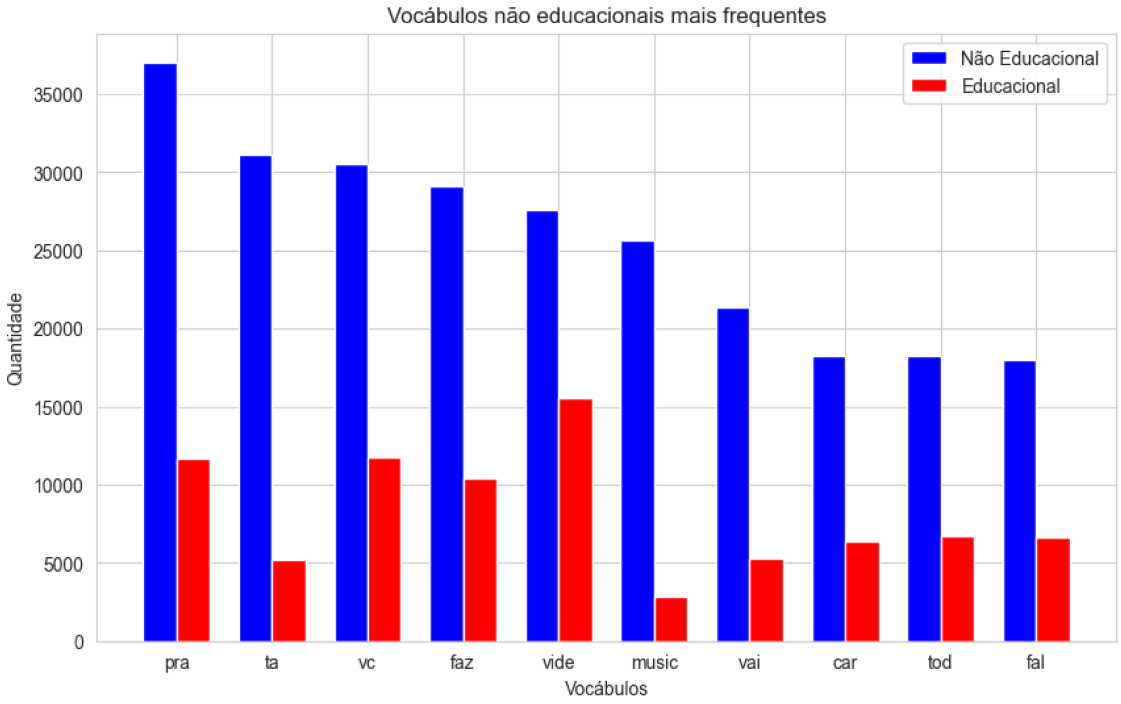


Figura 22 – Dados: Frequência dos vocábulos não educacionais (500 vídeos)

Ao analisar as figuras, percebe-se que os vocábulos são bem representativos em cada classe. Os vocábulos “profes”, “aul”, “obrig”, “aprend” e “ajud” são bem marcantes em vídeos educacionais e, principalmente, os dois primeiros apresentam uma grande diferença na frequência,

podendo ser bem importante para classificar os vídeos educacionais. Já os vocábulos mais frequentes nos vídeos não educacionais, apresentam uma diferença considerável entre os vídeos educacionais e não educacionais. Percebe-se que essa diferença também pode estar relacionada a elevada interação dos usuários aos vídeos não educacionais.

Devido as diferenças identificadas nas frequências dos vocábulos, iniciou-se os experimentos para classificar os vídeos através dos vocábulos mais frequentes.

5.2 Experimentos

O projeto teve início com o objetivo de classificar vídeos de maneira categórica, submetendo dados de um possível novo vídeo ao algoritmo, que então determinaria a qual classe ele pertence, seguindo a metodologia rígida. Além disso, visava-se entender melhor o comportamento dos usuários em ambas as categorias de vídeos, com o intuito de descobrir as formas mais eficazes de identificar conteúdo educacional por meio dos comentários. Nesse contexto, optou-se por algoritmos de aprendizado de máquina que oferecessem clareza em suas classificações, permitindo a identificação fácil dos melhores vocábulos para a classificação dos vídeos educacionais.

Para alcançar esses objetivos, selecionaram-se algoritmos baseados em árvores, como o Random Forest e o J48, além de algoritmos baseados em regras, como o JRIP e o PART. Estes foram escolhidos por seus bons resultados em tarefas de classificação e pela interpretabilidade de suas classificações. Além desses algoritmos supervisionados, empregou-se também o GenClust++, um algoritmo de clusterização não supervisionado, para explorar nuances não evidentes na análise supervisionada. O GenClust++ foi escolhido por ser uma versão avançada do K-Means, um dos algoritmos de clustering mais reconhecidos.

Motivados pelos resultados do primeiro experimento, buscamos aprimorar a classificação e explorar técnicas adicionais de aprendizado de máquina, incluindo redes neurais artificiais, com o objetivo de desenvolver um modelo mais acurado. Assim, iniciamos o segundo experimento, ampliando o dataset para incluir 500 vocábulos (features) e realizando comparações com o melhor modelo identificado anteriormente. Utilizamos o Random Forest e aplicamos redes neurais, incluindo a Rede Neural Simples, a Rede Neural Profunda Densa e a Rede Neural Convolucional, devido ao seu excelente desempenho em tarefas de classificação. Neste estágio, procuramos alcançar uma classificação superior, mesmo que isso afetasse a interpretabilidade dos algoritmos. Embora o uso de redes neurais reduza a compreensão das recomendações, a clareza das decisões do modelo pode não ser a prioridade em um sistema de recomendação.

Inspirados pelos resultados dos dois experimentos anteriores e com o objetivo de ultrapassar os desafios identificados, avançamos para o terceiro experimento. Esta etapa inovou ao propor a classificação dos comentários de forma individualizada, introduzindo a possibilidade de avaliar os vídeos com “graus de certeza” variáveis em relação ao seu conteúdo educacional. Essa variação oferece a flexibilidade de selecionar vídeos com base em um determinado nível de confiança na sua classificação. A intenção original era empregar o dataset de 500 vídeos para esta análise; no entanto, restrições de dados e limitações computacionais nos levaram a optar por uma amostra de 200 vídeos, distribuídos de maneira equitativa entre categorias educacionais e não educacionais.

Para a formulação inicial desta abordagem flexível, recorreu-se ao uso de apenas dois algoritmos: o Random Forest e o SVM (Support Vector Machine). O Random Forest foi escolhido devido ao seu desempenho satisfatório nos experimentos anteriores, enquanto o SVM foi selecionado por sua capacidade de generalização em uma variedade de problemas, frequentemente sendo destacado por sua eficácia comparável à das redes neurais.

Após finalizar o terceiro experimento, percebeu-se a importância de investigar mais detalhadamente as técnicas de conversão textual. Surgiu, portanto, o interesse em avaliar a eficiência de uma técnica de contagem simples de frequência dos vocábulos em comparação à estratégia até então utilizada, que prioriza os termos mais frequentes identificados em cada grupo. Essa investigação também visa determinar o limiar ótimo para o número de características a serem incluídas na análise. Portanto, no quarto experimento, realizou-se uma comparação entre a seleção dos vocábulos mais frequentes por classe e a contagem simples de vocábulos utilizando o CountVectorizer, uma implementação da técnica Bag of Words (BoW). Adicionalmente, o estudo explorou se o aumento no número de vocábulos considerados como características poderia influenciar positivamente os resultados da análise.

No quarto experimento, foram empregados cinco algoritmos, incluindo três variantes de Redes Neurais: duas Redes Neurais Simples, sendo uma desenvolvida com o framework Tensorflow e a outra com a biblioteca Scikit-learn, além de uma Rede Neural Convolucional. Complementando a análise, o Random Forest e o SVM, ambos implementados com o uso do Scikit-learn, também foram utilizados.

Finalmente, o quinto experimento foca na comparação entre as metodologias rígida e flexível, detalhando uma análise aprofundada dos resultados gerados pelos classificadores. Este experimento visa esclarecer quaisquer dúvidas sobre o desempenho comparativo das duas abordagens, determinando o modelo mais eficaz para ser implementado no LOIS.

Para realizar as avaliações das classificações dos modelos, algumas métricas podem ser utilizadas, entre elas podemos utilizar a Acurácia, Precisão, Recall e F1-Score.

A acurácia destaca-se por medir a eficácia geral do modelo em suas classificações, expressando a razão entre o número de previsões corretas e o total de previsões realizadas. Essa métrica é essencial para avaliar a capacidade do modelo de diferenciar corretamente entre as categorias envolvidas.

Em relação à precisão, esta métrica mede a proporção de previsões corretas da classe Positiva em relação a todas as previsões Positivas feitas pelo modelo. É uma métrica fundamental em contextos onde os falsos positivos (casos negativos classificados erroneamente como positivos) têm um custo elevado. A precisão é intuitivamente a capacidade do classificador de não rotular como positiva uma amostra negativa, ou seja, é a capacidade do classificador não rotular um vídeo não educacional (negativo) como educacional (positivo).

Quanto ao recall, conhecido também como sensibilidade ou revocação, esta métrica avalia a capacidade do modelo de identificar todas as instâncias reais da classe Positiva (educacional). Ele faz isso medindo a fração de Verdadeiros Positivos em relação ao número total de casos que pertencem à classe Positiva. O recall é intuitivamente a capacidade do classificador de encontrar todas as amostras da classe Positiva.

Por fim, o F1-Score representa uma métrica composta que integra precisão e recall em uma

única medida, através da média harmônica entre estas duas. Essa estratégia busca criar um equilíbrio entre a precisão e o recall, revelando-se particularmente valiosa em situações que exigem uma ponderação cuidadosa entre a redução de falsos positivos quando a identificação correta de todos os positivos reais. Considerando ambas as métricas, o F1-Score proporciona uma visão abrangente da performance do modelo, revelando-se crucial em ambientes onde tanto a precisão quanto o recall são fundamentais.

A acurácia é definida como a proporção de previsões corretas (tanto positivas quanto negativas) em relação ao total de previsões feitas. Matematicamente, pode ser expressa através da Equação 13.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

onde:

- TP representa os verdadeiros positivos,
- TN os verdadeiros negativos,
- FP os falsos positivos, e
- FN os falsos negativos.

Já a precisão foca nas previsões positivas corretas e é calculada pela razão dos verdadeiros positivos pelo total de positivos previstos pelo modelo e é apresentada na Equação 14. Já o Recall mede a proporção de positivos corretamente identificados pelo modelo em relação a todos os casos positivos reais, apresentado matematicamente pela Equação 15. Por fim, o F1-Score é a média harmônica entre precisão e recall, oferecendo um equilíbrio entre estas duas métricas e é representado pela Equação 16

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

$$\text{F1-Score} = \frac{2 * TP}{2 * TP + FP + FN} \tag{16}$$

Estas métricas oferecem resultados numéricos para avaliar o desempenho de modelos de classificação, facilitando uma análise mais profunda e informada sobre suas habilidades de previsão.

5.2.1 Experimento 1

Para o primeiro experimento, realizou-se uma expansão do dataset de 200 vídeos, 100 educacionais e 100 não educacionais, para 500 vídeos, 250 educacionais e 250 não educacionais. Após o aumento e coleta dos comentários, foi necessário pré-processar os comentários e gerar os datasets para o experimento. Inicialmente foram gerados 8 datasets, contendo 10, 20, 40, 60, 80, 100, 200 e 6 “melhores vocábulos” selecionados em cima do dataset contendo 200 vocábulos. Nesse

experimento, a categorização dos datasets foi realizada através dos algoritmos implementados no *framework* Weka, que fornece uma vasta coleção de ferramentas para classificação, regressão, seleção de atributos, dentre outros (FRANK; HALL; WITTEN, 2016).

Para os experimentos realizados, foram utilizadas técnicas de Aprendizado de Máquina supervisionadas e não supervisionadas. Para os experimentos "supervisionados", foram utilizados quatro classificadores, a saber: JRip , PART, J48 e Random Forest. Para os experimentos "não supervisionados", foi utilizado o GenClust++. Todas as técnicas utilizadas foram configuradas com os valores *default* dos seus parâmetros. Foram utilizados os datasets contendo 10, 20, 40, 60, 80, 100, 200 e os vocábulos selecionados através do avaliador de atributos "CfsSubsetEval".

Como já pontuado, foram geradas, inicialmente, 8 bases de dados, sendo que a diferença entre elas se faz pelo número de vocábulos mais frequentes nos comentários dos vídeos. Nesse sentido, o dataset #1 utiliza os 10 vocábulos mais frequentes, a base de dados #2 utiliza os 20 vocábulos mais frequentes e assim sucessivamente, para 40, 60, 80, 100 e 200 vocábulos mais frequentes nos comentários dos vídeos selecionados. Os vocábulos da base de dados #8 foram selecionados da base de dados #7 (que usa os 200 vocábulos mais frequentes) utilizando-se o algoritmo de seleção de atributos do Weka (FRANK; HALL; WITTEN, 2016) com o avaliador de atributos "CfsSubsetEval" e método de busca "BestFirst", ambos com os parâmetros *default* do *framework*. Tal processo resultou numa base de dados contendo apenas seis vocábulos, a saber: "aul", "profes", "prof", "clip", "jog" e "compr". Essa base de dados foi construída com o objetivo de se verificar o comportamento dos classificadores quando se utiliza como forma de pré-processamento de dados um algoritmo de seleção de atributos. Observa-se que essa configuração foi utilizada no segundo experimento, e, posteriormente, o terceiro experimento foi adicionado com o dataset contendo 500 vocábulos.

A título de esclarecimento, aponta-se, por exemplo, que o dataset #1, que possui os 10 vocábulos mais frequentes nos comentários dos vídeos, contém os 5 primeiros vocábulos mais frequentes nos comentários dos vídeos educacionais, e os 5 vocábulos restantes são preenchidos com os 5 primeiros vocábulos mais frequentes nos comentários dos vídeos não educacionais, que ainda não estão presentes. Dessa forma, um vocábulo mais frequente não aparecerá repetidamente. De forma análoga, a base de dados #2, que possui os 20 vocábulos mais frequentes nos comentários dos vídeos, contém os 10 primeiros vocábulos mais frequentes nos comentários dos vídeos educacionais, e os 10 primeiros vocábulos mais frequentes nos comentários dos vídeos não educacionais, que ainda não foram utilizados. Esse modelo de estruturação foi utilizado para a formação das sete bases de dados contendo, como apontado, os 10, 20, 40, 60, 80, 100, 200 e 500 vocábulos mais frequentes nos comentários dos vídeos. Por sua vez, a base de dados #8 foi construída conforme previamente apresentado.

Em relação aos elementos das bases de dados, cada exemplo ou instância da mesma representa um vídeo. Cada vídeo é descrito pelo número de vezes que cada vocábulo mais frequente figurou em seus comentários, e por sua classe (educacional, descrito como "*yes*" ou não educacional, descrito por "*no*"). Para exemplificar a estrutura das bases de dados, a Tabela 5 apresenta uma amostra da base de dados #1 que armazena os 10 vocábulos mais frequentes e 10 vídeos, sendo 5 educacionais e 5 não educacionais. A primeira coluna da tabela representa o "*Id*" de cada vídeo, e as colunas seguintes representam o número de vezes que cada vocábulo figurou

no comentário do mesmo. Por sua vez, a última coluna da tabela representa a classe do vídeo, sendo "*yes*" para os vídeos educacionais e "*no*" para os vídeos não educacionais. Neste sentido, o vídeo de "*Id*" = "4g9JTQ2B6oo", apresenta em seus comentários, 21 repetições do vocábulo "aul", 28 repetições do vocábulo "profes", 4 repetições do vocábulo "obrig", 3 repetições do vocábulo "vide", 0 repetições do vocábulo "aprend", 5 repetições do vocábulo "pra", 1 repetição do vocábulo "ta", 5 repetições do vocábulo "vc", 1 repetição do vocábulo "faz", 0 repetições do vocábulo "music", e a classe educacional "yes". Todas as bases de dados deste trabalho seguem essa mesma estrutura, com a diferença do número de vocábulos mais frequentes em cada uma delas.

Ressalta-se, que nesse experimento optou-se pela utilização de algoritmos baseados em regras e em árvores devido interpretabilidade dos modelos gerados, i.e, devido à facilidade em se abordar e se interpretar as classificações realizadas por esta "categoria" de algoritmos e também para explorar melhor os comportamentos dos usuários perante vídeos educacionais e não educacionais.

Durante a seleção dos algoritmos a serem utilizados, atentou-se também em se verificar o possível desempenho de algoritmos de clusterização. Neste sentido, selecionou-se o algoritmo GenClust++, uma vez que esta técnica apresenta, de forma geral, resultados superiores quando analisada em relação a outras técnicas de clusterização para fins de classificação. Apesar disso, é importante destacar que esta técnica obteve resultados inferiores às técnicas supervisionadas adotadas no trabalho e, que, portanto, optou-se por não se investigar mais profundamente o uso de técnicas não supervisionadas no presente trabalho.

No total, realizaram-se 8 experimentos para cada técnica avaliada, a saber: JRIP, PART, J48, Random Forest e GenClust++. Os experimentos foram numerados de #1, #2, #3, #4, #5, #6, e #7, e #8 e utilizaram as bases de dados numeradas e descritas anteriormente. Todas as oito bases de dados utilizadas na presente proposta possuem 250 vídeos educacionais e 250 vídeos não educacionais. Os resultados experimentais foram obtidos utilizando-se a metodologia de validação cruzada de 10 folds (*10-fold cross-validation*).

A Tabela 12 apresenta as acurácia dos métodos utilizados para cada base de dados seguindo a metodologia experimental previamente descrita. Dessa forma, a coluna "Exp." descreve o número do experimento, a coluna "Vocábulos" aponta a quantidade de vocábulos mais frequentes utilizada no experimento, a coluna "Média" apresenta a média das acurácias obtidas para as técnicas considerando isoladamente cada base de dados, e os classificadores Random Forest e GenClust++ foram abreviados para RF e GC++, respectivamente, é o Por fim, a Tabela 12 destaca, em negrito, os melhores resultados para cada experimento, e apresenta, em sua última linha, a média dos resultados de cada método considerando todas as bases de dados.

Nota-se que o PART apresentou o melhor resultado para o experimento #8 (6 vocábulos selecionados) com acurácia de 89,03%. O J48 apresentou os melhores resultados para os experimentos #1 (10 vocábulos) e #2 (20 vocábulos), com acurácias, respectivamente, de 89,86% e 89,65%. Por sua vez, o Random Forest apresentou os melhores resultados para os experimentos #3 (40 vocábulos) com acurácia de 90,68%, #4 (60 vocábulos) com acurácia de 89,03%, #5 (80 vocábulos) com acurácia de 90,06%, #6 (100 vocábulos) com acurácia de 90,68%, e #7 (200 vocábulos) com acurácia de 91,30%. As técnicas JRIP e GenClust++ não obtiveram os melhores resultados em nenhum dos experimentos realizados. Aponta-se, ademais, que o Random

Tabela 12 – Experimento 1: Resultados dos experimentos

Exp.	Vocábulos	JRIP	PART	J48	RF	GC++	Média
#1	10	86,75%	87,78%	**89,86%**	86,96%	51,35%	80,54%
#2	20	87,37%	87,37%	**89,65%**	89,44%	51,35%	81,04%
#3	40	87,37%	87,16%	86,96%	**90,68%**	50,72%	80,58%
#4	60	86,96%	86,54%	87,58%	**89,03%**	51,14%	80,25%
#5	80	86,13%	87,37%	86,75%	**90,06%**	57,14%	81,49%
#6	100	87,16%	86,13%	85,92%	**90,68%**	56,94%	81,37%
#7	200	86,75%	85,92%	85,92%	**91,30%**	60,46%	**82,07%**
#8	"seleção"	87,16%	**89,03%**	87,37%	83,02%	51,14%	79,54%
Média	64.5	86,96%	87,16%	87,50%	**88,90%**	53,78%	

Fonte: Elaborado pelo autor, com base na pesquisa realizada

Forest obteve a melhor média de acurácia considerando todas as técnicas, e que a maior média de acurácia, considerando todas as bases de dados, foi obtida para o experimento #7.

Observam-se que todos os experimentos em que foram utilizadas técnicas supervisionadas, i.e., JRIP, PART, J48 e Random Forest, apresentam elevada acurácia, ou seja, elevado número de acertos ao se classificar um vídeo entre educacional ou não. Dentre os experimentos realizados utilizando-se técnicas supervisionadas, a menor acurácia, 83,02%, foi obtida pelo Random Forest no experimento #8. Não obstante, como já apontado, o Random Forest também obteve a maior acurácia entre todos os testes, com a taxa de 91,30% no experimento #7. Destaca-se que neste mesmo experimento obteve-se a maior média de acurácia considerando todas as técnicas utilizadas.

O classificador JRIP obteve seus melhores resultados nos experimentos #2, e #3, com acurácia de 87,37%. A título de ilustração, a Figura 23 apresenta as regras de classificação geradas pelo JRIP para o experimento #2.

```
JRIP rules:
===========

(aul <= 4) and (ta >= 3) => Class=no (133.0/4.0)
(aul <= 0) and (tod >= 1) => Class=no (27.0/5.0)
(profes <= 0) and (aul <= 0) => Class=no (70.0/22.0)
(fal >= 64) and (profes <= 37) => Class=no (23.0/1.0)
(fal >= 56) and (parab <= 58) => Class=no (5.0/0.0)
 => Class=yes (225.0/11.0)

Number of Rules : 6
```

Figura 23 – Experimento 1: Regras de classificação do JRIP para o experimento 1: #2

A técnica utiliza um total de 6 regras de classificação para o experimento #2. Observam-se que tais regras são facilmente interpretáveis. Nesse sentido, a primeira regra, "(aul <= 4) and (ta => 3) => Class=no", informa que, se nos comentários de um vídeo figurarem os vocábulos "aul", 4 ou menos vezes, e "ta", 3 ou mais vezes, então este vídeo é classificado como não educacional. Todas as outras regras geradas são interpretadas dessa mesma maneira. É importante notar que embora o JRIP não tenha obtido a melhor acurácia dentre todas as técnicas em nenhum experimento, seu modelo de classificação é bastante simples e intuitivo.

```
PART decision list
------------------

aul > 4 AND
clip <= 0 AND
jog <= 3 AND
jog <= 1: yes (120.0/1.0)

profes > 46 AND
aul > 59: yes (39.0)

clip <= 0 AND
compr > 0 AND
prof <= 0 AND
compr > 3: no (63.0)

clip <= 0 AND
aul <= 0 AND
prof <= 0 AND
jog <= 1 AND
profes <= 0: no (96.0/23.0)

clip > 0: no (52.0)

compr <= 0 AND
jog <= 8 AND
profes <= 1 AND
jog <= 0: yes (31.0/3.0)

compr <= 0 AND
profes > 1: yes (21.0)

: no (61.0/16.0)

Number of Rules  :      8
```

Figura 24 – Experimento 1: Regras de classificação do PART para o experimento 1: #8

O classificador PART obteve o melhor resultado, dentre todas as técnicas, no experimento #8, com acurácia de 89,03%. A título de ilustração, a Figura 24 apresenta as regras de classificação geradas pelo PART para o experimento #8.

A técnica utiliza um total de 8 regras de classificação para o experimento #8. Tais regras são também facilmente interpretáveis. Nesse sentido, a primeira regra, "(aul > 4) AND (clip <= 0) AND (jog <= 3) AND (jog <= 1) : yes", informa que, se nos comentários de um vídeo figurarem os vocábulos "aul", 5 vezes ou mais, "clip", nenhuma vez, e "jog" uma ou nenhuma vez, então este vídeo é classificado como educacional. Todas as outras regras geradas são interpretadas dessa mesma maneira. É importante notar que o PART obteve o melhor resultado dentre todas as técnicas para o experimento #8, contudo, suas regras são mais complexas quando comparadas às regras obtidas pelo JRIP. Apesar disso, ainda pode-se afirmar que as regras do PART são de simples interpretação.

O classificador J48 obteve os melhores resultados, dentre todas as técnicas, nos experimentos #1, e #2, com acurácias de 89,86% e 89,65%, respectivamente. A título de ilustração, a Figura 25 apresenta a árvore de decisão gerada para o experimento #1.

```
J48 pruned tree
------------------

aul <= 4
|   ta <= 2
|   |   aul <= 0
|   |   |   profes <= 0
|   |   |   |   vc <= 2
|   |   |   |   |   vide <= 0: no (59.0/15.0)
|   |   |   |   |   vide > 0
|   |   |   |   |   |   vc <= 0
|   |   |   |   |   |   |   aprend <= 0: yes (12.0/4.0)
|   |   |   |   |   |   |   aprend > 0: no (2.0)
|   |   |   |   |   |   vc > 0
|   |   |   |   |   |   |   vide <= 6: no (9.0)
|   |   |   |   |   |   |   vide > 6: yes (2.0)
|   |   |   |   vc > 2: no (8.0)
|   |   |   profes > 0
|   |   |   |   aprend <= 2: yes (10.0/1.0)
|   |   |   |   aprend > 2: no (3.0)
|   |   aul > 0
|   |   |   pra <= 9: yes (34.0/3.0)
|   |   |   pra > 9: no (4.0/1.0)
|   ta > 2: no (133.0/4.0)
aul > 4
|   ta <= 40
|   |   aul <= 15
|   |   |   vc <= 55: yes (52.0/2.0)
|   |   |   vc > 55: no (3.0)
|   |   aul > 15: yes (105.0)
|   ta > 40
|   |   profes <= 51: no (26.0/1.0)
|   |   profes > 51: yes (21.0/1.0)

Number of Leaves  :     16

Size of the tree :      31
```

Figura 25 – Experimento 1: Árvore de decisão do J48 para o experimento 1: #1

A técnica utiliza um total de 16 regras de classificação para o experimento #1. Cada regra é formada pelo caminho partindo do nó raiz até se chegar a um nó folha. À exemplo, considerando que o vocábulo do nó raiz, "aul", figure 4 ou menos vezes nos comentários do vídeo, "ta" figure 2 ou menos vezes, "aul" não figure, "profes" não figure, "vc" figure 2 ou menos vezes, e "vide"

não figure, então o vídeo é classificado como não educacional (nó folha "*no*"). Todas as outras regras geradas são interpretadas dessa mesma maneira. É importante notar que o J48 obteve o melhor resultado dentre todas as técnicas para os experimentos #1 e #2, contudo, suas regras são bem mais complexas quando comparadas às regras obtidas pelo JRIP e PART. Apesar disso, o modelo de classificação do J48 se mostrou interpretável.

O classificador GenClust++ obteve os piores resultados para os experimentos realizados e não se mostrou viável para resolver o problema em questão.

Por fim, o Random Forest obteve os melhores resultados, dentre todas as técnicas, nos experimentos #3, #4, #5, #6, e #7, com acurácias, respectivamente, de 90,68%, 89,03%, 90,06%, 90,68%, e 91,30%. Apesar desta técnica obter os melhores resultados na maioria das bases de dados utilizadas, é importante ressaltar que o Random Forest gera um modelo de classificação bastante complexo, utilizando-se de várias árvores de decisão para realizar sua classificação. À exemplo, no experimento #3, a quantidade de regras das árvores de decisão geradas foi de 150, um número bastante elevado. Devido à complexidade do modelo gerado, optou-se por não apresentá-lo.

Neste experimento foi apresentado e testado uma proposta de aplicação de Técnicas de Aprendizagem de Máquina no intuito de classificar vídeos educacionais do Youtube por meio de seus comentários. Aponta-se que a proposta foi capaz de diferenciar, com elevada acurácia 91,30%, vídeos educacionais de vídeos não educacionais. Percebe-se, também, que a técnica de Aprendizagem de Máquina não supervisionada utilizada não performou bem nos problemas abordados.

Os resultados desse experimento demonstraram a conquista do objetivo proposto em termos de acurácia através da utilização do dataset contendo 200 vocábulos e o algoritmo Random Forest. Contudo, esses resultados despertaram o interesse de aprimorar ainda mais a acurácia alcançada, aplicar outros algoritmos de aprendizagem (Redes Neurais) e de desenvolver um Sistema de Recomendação capaz de performar bem a recomendação de vídeos educacionais.

5.2.2 Experimento 2

O segundo experimento foi realizado com o objetivo de melhorar a acurácia obtida no experimento anterior e desenvolver um modelo de aprendizado adequado para o Sistema de Recomendação. Devido a proposta de obter um modelo com melhor acurácia, não preocupou-se com a interpretabilidade do classificador, pois o comportamento de usuários que buscam vídeos educacionais e que procuram outros tipos de vídeos foi explorado anteriormente.

Para alcançar esse objetivo, utilizou-se um dataset inicial contendo 200 vocábulos, esse dataset foi o que apresentou melhor acurácia no experimento anterior, e, em seguida, foi gerado um novo dataset com 500 vocábulos, seguindo a mesma metodologia empregada anteriormente. Esse experimento buscou explorar um conjunto de dados expandido, o qual poderia fornecer informações mais abrangentes e precisas para o modelo de aprendizado. A adição de mais vocábulos ao dataset permite considerar um espectro maior de características e nuances na classificação dos OA.

Tabela 13 – Experimento 2: Acurácia para 200 vocábulos

Execução	Random Forest	NN	DNN	CNN
1	92,60%	94,18%	92,02%	94,15%
2	92,54%	94,73%	95,76%	93,63%
3	93,60%	93,65%	90,44%	93,07%
4	92,49%	91,46%	93,13%	93,54%
5	92,57%	93,07%	94,09%	93,01%
6	93,63%	93,13%	93,54%	94,18%
7	94,15%	92,46%	94,65%	95,26%
8	92,05%	94,15%	93,63%	93,63%
9	93,63%	94,65%	91,49%	94,71%
10	92,05%	92,04%	93,60%	93,65%
Média	92,93%	93,35%	93,23%	93,88%

Ao utilizar a mesma metodologia anterior, as etapas de pré-processamento, classificação e geração dos conjuntos de dados foram aplicadas novamente. No entanto, o novo dataset com 500 vocábulos ofereceu um dataset mais abrangente, com mais características, para serem utilizadas pelos classificadores na etapa de treino e validação.

Para atingir as metas propostas para o experimento, houve uma alteração na utilização dos aplicativos anteriores, priorizando o Python como linguagem principal do projeto. Essa escolha também visa facilitar o desenvolvimento do Sistema de Recomendação. No experimento, a linguagem de programação Python foi utilizada em conjunto com o framework Tensorflow para construir as Redes Neurais, além da biblioteca Scikit-learn para comparação com o Random Forest, que se destacou como o melhor algoritmo empregado no segundo experimento. A opção pelo Python deve-se à sua flexibilidade e à disponibilidade de diversas bibliotecas e frameworks que auxiliam no campo de Aprendizagem de Máquina.

Os classificadores Random Forest, Rede Neural, Rede Neural Profunda e Rede Neural Convolucional foram utilizados nesse experimento para classificar os vídeos como educacionais ou não educacionais. A biblioteca Scikit-learn foi utilizada para gerar o Random Forest e o framework Tensorflow foi utilizado para gerar as Redes Neurais.

Para realizar a criação das Redes Neurais, foi utilizado o Tensorflow, sendo necessário definir a arquitetura da rede, que inclui a escolha do número de camadas, o número de neurônios em cada camada, a função de ativação e outros parâmetros.

A arquitetura da Rede Neural utilizada no experimento é demonstrada no código 4.1. A Rede Neural Profunda, utilizada no experimento é demonstrada no código 4.2. Por fim, a Rede Neural Convolucional é demonstrada no código 4.3

O algoritmo Random Forest apresentou valores de acurácia ligeiramente superiores aos alcançados no experimento anterior. É importante destacar que a arquitetura simples da Rede Neural desenvolvida também mostrou desempenho notável, atingindo uma acurácia média de 93,35%. Essa rede neural simples foi superada apenas pela Rede Neural Convolucional (CNN), indicando a eficácia das abordagens baseadas em redes neurais para essa aplicação.

A tabela 14 apresenta as acurácias obtidas no experimento para o dataset contendo 500 vocábulos.

Tabela 14 – Experimento 2: Acurácia para 500 vocábulos

Execução	Random Forest	NN	DNN	CNN
1	92,08%	93,10%	93,65%	95,18%
2	93,10%	93,51%	95,73%	95,58%
3	90,96%	94,65%	94,15%	96,20%
4	91,49%	93,54%	94,15%	95,09%
5	93,13%	92,57%	94,65%	95,67%
6	92,54%	93,57%	94,09%	96,26%
7	92,02%	93,65%	94,18%	94,01%
8	93,10%	94,12%	96,23%	95,12%
9	91,43%	94,62%	92,60%	94,59%
10	91,46%	95,23%	95,23%	95,67%
Média	92,13%	93,86%	94,46%	95,34%

Ao analisar resultados, observa-se que, de maneira geral, a maioria dos algoritmos apresentaram melhores desempenhos com o dataset maior, contendo 500 palavras. Apenas o Random Forest obteve resultados levemente inferiores, uma diferença de 0,8%. Os resultados demonstram que a expansão do dataset para 500 palavras apresentou melhorias para as redes neurais mais complexas, como a Rede Neural Profunda Densa e a Rede Neural Convolucional. A variedade maior de palavras permitiu que os algoritmos reconhecessem melhor as diferenças entre as duas classes dos vídeos através dos comentários, o que possibilitou maior acurácia.

A Rede Neural Convolucional demonstrou desempenho excepcional, ultrapassando os demais algoritmos em ambos os conjuntos de dados. Com uma precisão média de 95,34%, a Rede Neural Convolucional demonstra elevada capacidade de identificar os padrões e distinguir os vídeos educacionais e não educacionais. Ressalta-se que os resultados são expressivos, principalmente, pela natureza dos dados e a dificuldade em interpretar os comentários.

A excelente performance da Rede Neural Convolucional ressalta a importância de uma metodologia bem montada no desenvolvimento do dataset. Os resultados mostram que as diferenças entre as classes educacional e não educacional foram capturadas de forma eficiente. Isso sugere que o processo de coleta e pré-processamento dos dados foi criterioso e bem-sucedido.

Após alcançar excelentes resultados neste experimento, e enfrentar dificuldades técnicas relacionadas à capacidade de classificação em tempo real, surgiu o interesse em desenvolver uma abordagem mais flexível e adaptável para a classificação dos Objetos de Aprendizagem (OA). Assim, a proposta inicial, metodologia rígida, foi aprimorada, resultando na criação de uma variação "flexível". Esta nova metodologia foi implementada no experimento 3, que permite uma melhor adaptação às incertezas presentes nos comentários e oferece a possibilidade de recomendar os OA com diferentes graus de certeza, uma funcionalidade não disponível na abordagem rígida anterior.

5.2.3 Experimento 3

O terceiro experimento foi realizado através da segunda abordagem, abordagem flexível para a classificação dos OA. Foi utilizado o dataset contendo 200 vídeos, sendo 100 vídeos

educacionais e 100 vídeos não educacionais. Este dataset é composto por 158.559 comentários, tornando-se uma boa amostra para a análise. Para a classificação dos comentários foram utilizados os classificadores Random Forest e SVM. O experimento foi realizado em Python e com a utilização da biblioteca Scikit-learn.

É importante destacar que, devido à complexidade computacional envolvida, a opção de utilizar um dataset menor foi utilizada. Isso se deve ao fato de que o dataset expandido, contendo 500 vídeos, demandaria recursos computacionais mais elevados, tornando inviável a sua utilização nesse primeiro momento. Essa limitação fez com que um dataset menor fosse utilizado em um primeiro momento, porém estamos procurando alternativas para utilizar o dataset expandido.

A complexidade computacional está diretamente relacionada a necessidade de transformar os textos (comentários), em formas de serem interpretadas pelos algoritmos de AM. Nesse projeto, em vetorizar todos os comentários para serem representados por números. A critério de comparação o dataset contendo 500 vídeos possui 735.517 comentários, levando em consideração o pre-processamento e, também, a exclusão de comentários "vazios", ou seja, que foram processados e não apresentam conteúdo, e, isso gera um dataset que possui 718.048 linhas e um determinado número de colunas, características ou vocábulos, que representam o texto.

Para a transformação do dataset utilizado, foi utilizado a técnica "CountVectorizer", essa técnica realiza a contagem das palavras e transforma cada comentário em uma lista onde cada coluna representa a ocorrência de determinada palavra em cada comentário. Para utilizar a técnica foi utilizado os parâmetros max_features = 3000, min_df = 2, max_df = 0.7 e stop_words = 'portuguese'.

O parâmetro max_features definiu que cada comentário será convertido em um vetor de 3 mil posições, o min_df significa que as palavras que poderão ser utilizadas para gerar nossa matriz deverão ter uma frequência de pelo menos 2, ou seja, em todos os meus comentários as palavras devem ter sido utilizadas pelo menos 2 vezes, o max_df=0.7 especifica a frequência máxima que uma palavra pode ocorrer nos comentários, nesse caso, palavras que aparecem em mais de 70% dos documentos são excluídas e, por fim, o stop_words='portuguese' remove os stopwords, no projeto esse comando é redundante pois os dados já haviam sido pré-processados e os stopwords removidos.

A abordagem proposta utiliza a equação 11 para calcular $PEdu_{vi}$, que é a probabilidade de um vídeo *vi* ser educacional.

Dessa forma, os algoritmos de aprendizagem utilizam cada comentário individualmente e o classificam, dessa maneira conseguimos pegar cada classificação e aplicar as equações anteriores para determinar a probabilidade do vídeo ser educacional. Desenvolvemos um sistema para realizar essa classificação probabilística. O sistema foi codificado em Python e utilizou as bibliotecas *unidecode*, *regex*, *nltk*, *string* e o corpus stopwords em português.

A Tabela 15 mostra exemplos de comentários pré-processados de ambas as classes, educacional (edu) e não educacional (nao_edu), utilizados para treinamento dos classificadores.

A Tabela 16 mostra os resultados de acurácia alcançados pelos classificadores usando o método de validação cruzada de 10 folds, cada linha representa a uma execução do 10-folds. Observa-se que resultados de elevada acurácia são alcançados tanto pelo Random Forest quanto

Tabela 15 – Experimento 3: Exemplo do dataset utilizado.

Comentarios	Class
car alucinadokkk melhor professorkkkkkkkkkkkk	edu
aul excel	edu
melhor profes hist	edu
ach kkkk	edu
coloqu veloc parec profes ta chap kkkkkkkkkkkkkkkkkkk	edu
sucess carr sol bom heranc esper vc alcanc sucess aind mand cd ai kkkkkkkkkkkk	nao_edu
faz temp ouv music heranc desd sai ms procur alg grup ano menos agor ach ire ouv	nao_edu
temp procur dvd top demal	nao_edu
guau busqu dvd grup herenc much tiemp per fin lo pud ver vay en verdad gen principi fin heranc dvd complet en la voz jaim juni	nao_edu
hist nao	nao_edu

Fonte: Elaborado pelos autores.

Tabela 16 – Experimento 3: Acurácia da classificação dos comentários.

Execução	Random Forest	SVM
#1	83,97%	84,56%
#2	83,78%	84,53%
#3	83,82%	84,63%
#4	83,70%	84,56%
#5	83,99%	84,65%
#6	83,84%	84,55%
#7	84,02%	84,67%
#8	83,85%	84,62%
#9	83,83%	84,66%
#10	83,89%	84,66%
Média	**83,87%**	**84,61%**

Fonte: Elaborado pelo autor.

pelo SVM ao classificar os comentários como educacionais ou não educacionais. Nesse sentido, ressalta-se que a pequena melhora nos resultados de acurácia do SVM, em relação ao Random Forest, não é estatisticamente relevante, portanto, ambos os classificadores, considerando os resultados de acurácia, são equivalentes neste conjunto de dados.

Uma vez que os classificadores se mostraram eficazes em classificar os comentários em educacionais e não educacionais, o passo seguinte foi utilizá-los para realizar a classificação probabilística dos vídeos. Nesse sentido, selecionamos dois vídeos (removidos para revisão I), a saber, “Herança Autossômica” e “Engenharia de Software - Aula 01 - Modelos de processo de software e atividades de software”, para serem classificados probabilisticamente como educacional ou não educacional.

A Tabela 17 apresenta os comentários sobre o vídeo “Herança Autossômica”. Além disso,

Tabela 17 – Experimento 3: Classificação dos comentários no vídeo "Herança Autossômica"

Comentário	Comentário processado	Class
7:47 Falha na realidade bem ali.	falh real bem ali	edu
Muito boa a aula ajudou bastante!	boa aul ajud bast	edu
Ótima aula	otim aul	edu
Suas aulas são muito boas!!! <3	aul boa	edu
Muito bom, parabéns! Custei achar um vídeo que tratasse desse assunto de uma forma mais fácil e dinâmica de entender. Obrigada pela aula!	bom parab cust ach vide trat dess assunt facil dinam entend obrig aul	edu
Obrigada pela aula!	obrig aul	edu

Fonte: Elaborado pelo autor.

apresenta os comentários pré-processados e suas aulas (edu ou nao_edu) de acordo com o modelo de classificação do SVM. Nota-se que todos os 6 comentários pré-processados são classificados como educacionais, assim, utilizando a equação 11, o sistema calcula $PEdu_{vi} = 6/(6+0) = 1$. Assim, o vídeo é classificado como educacional com probabilidade igual a 1.

A Tabela 18 apresenta os comentários do vídeo "Engenharia de Software - Aula 01 - Modelos de processo de software e atividades de software". Além disso, apresenta os comentários pré-processados e suas aulas (edu ou nao_edu) de acordo com o modelo de classificação do SVM. Pode-se notar que de todos os 5 comentários pré-processados, 3 são classificados como educacionais e 2 são classificados como não educacionais, assim, usando a equação 11, o sistema calcula $PEdu_{vi} = 3/(3+2) = 0{,}6$. Assim, o vídeo é classificado como educativo com probabilidade igual a 0,6.

Aqui, é importante destacar a flexibilidade da interpretação da classificação probabilística realizada pela metodologia proposta. Em relação ao primeiro vídeo, 100% de seus comentários são classificados como educacionais e, portanto, conclui-se que este vídeo tem alta probabilidade de ser educativo. Em relação ao segundo vídeo, apenas 60% dos seus comentários são classificados como educacionais e, portanto, a classificação deste vídeo como educacional deve ser interpretada com alguma parcimônia.

Nesse sentido, um Sistema Tutor Inteligente pode considerar um limite probabilístico para recomendar um determinado vídeo. Como exemplo, o sistema pode ser parametrizado para recomendar vídeos considerados educativos com probabilidade acima de 80%. Para vídeos classificados como educativos com probabilidade inferior a esse limite, o sistema pode tentar considerar outros metadados do vídeo, como título, número de likes e dislikes, descrição etc., a fim de verificar se o vídeo se enquadra na categoria educacional.

A abordagem proposta considera a frequência de comentários educacionais em um vídeo para classificá-lo probabilisticamente como educacional ou não educacional. Neste sentido, esta metodologia permite facilmente testar e utilizar outras técnicas e algoritmos com o mesmo objetivo. Portanto, é importante ressaltar que a abordagem proposta é bastante flexível, pois permite que o classificador seja facilmente alterado sem interferir na arquitetura do sistema desenvolvido, proporcionando uma maneira fácil de produzir novos experimentos e avanços.

Tabela 18 – Experimento 3: Classificação para os comentários no vídeo “Engenharia de Software - Aula 01 - Modelos de processo de software e atividades de software”.

Comentário	Comentário processado	Class
O dia é 6 e eu pretendo entrar em engenharia da computação.	dia pret engenh computaca	edu
Grato pela aula	grat aul	edu
Assistindo essas aulas percebo que “antigamente” os professores da Univesp davam aulas mesmo, bem preparadas, com comentários relevantes e demonstrando profundo conhecimento. As aulas atuais (2021) são uma chatice, com o professor passando e lendo os slides. Lamentável.	assist aul perceb antig profes univesp dav aul bem prepar comentari relev demonstr profund conhec aul atual chat profes pass lend slid lamenta	edu
Experiência e prática: Uma boa forma de reduzir o tempo de produção de um software. Algo que depende do engenheiro. :)	experienc pra boa reduz temp produca softw alg depend engenh	nao_edu
Amando cada vez mais a Engenharia de Software. Esse curso tem tudo o que eu quero seguir profissionalmente. #VemUFC	am cad vez engenh softw curs tud quer segu profiss vemufc	nao_edu

Fonte: Elaborado pelo autor.

Nesse experimento, é importante destacar algumas particularidades. Inicialmente, a coluna “Id” não foi utilizada, pois o objetivo era processar os comentários individualmente e permitir que os algoritmos classificassem cada um deles, obtendo assim, uma classificação mais flexível. A análise das acurácias foi baseada na classificação dos comentários e apresentou duas classificações realizadas em vídeos.

Contudo, é necessário ampliar a análise, focando em verificar a acurácia ao classificar diretamente um vídeo como educacional ou não educacional. Para isso, podemos “limitar” um pouco a proposta, definindo uma probabilidade mínima para considerar um vídeo como educacional ou não. Nessa análise obteremos um resultado mais sólido quanto a abordagem e conseguiremos identificar melhor a aplicação prática da abordagem, conseguindo até comparar as duas metodologias desenvolvidas, possibilitando utilizar determinada abordagem em cada caso.

É importante ressaltar que, devido a restrições computacionais e a necessidade de explorar a abordagem flexível, optou-se por utilizar o dataset com 200 vídeos. Para fins de comparação, o dataset com 500 vídeos, sem restrições no número de palavras através do parâmetro max_features e utilizando min_df = 2, geraria um dataset no formato (718.048, 45.712), ou seja, 718.048 linhas e 45.712 colunas. Esse tamanho de dataset requer aproximadamente 248 GB de memória RAM para ser gerado, o que torna inviável para muitos computadores.

5.2.4 Experimento 4

O Experimento 4 foi concebido com o objetivo de comparar duas técnicas distintas para a vetorização dos comentários: a abordagem empregada no trabalho, utilizando os vocábulos

mais frequentes em cada categoria e a aplicação de uma métrica de contagem de palavras, o CountVectorizer, uma implementação da técnica Bag of Words (BoW). O propósito era identificar qual abordagem resultaria em uma classificação mais eficaz dos Objetos de Aprendizagem (OA).

Para este experimento, utilizou-se um dataset que incluía 500 vídeos, divididos igualmente entre conteúdos educacionais (250 vídeos) e não educacionais (250 vídeos), acompanhados por um total de 718.048 comentários. Desenvolveu-se um dataset para esse experimento, contendo os 500 vocábulos mais frequentes, sendo 250 deles extraídos dos vídeos educacionais e os outros 250 dos vídeos não educacionais. Estes vocábulos foram os mesmos utilizados nos Experimentos 1 e 2 e, sendo utilizada a mesma metodologia empregada nesses experimentos.

Na classificação dos comentários, empregaram-se diferentes classificadores: Random Forest, SVM, Rede Neural Simples (MultiLayer Perceptron (MLP)) através da biblioteca Scikit-learn, além de uma Rede Neural Simples e uma Rede Neural Convolucional utilizando o framework TensorFlow. O experimento foi conduzido na linguagem de programação Python, com o auxílio da biblioteca Scikit-learn e do framework TensorFlow.

Destaca-se a importância da transformação de textos em formatos numéricos para o processamento por algoritmos de aprendizado de máquina. Neste experimento, foram exploradas duas técnicas de representação textual para comparação. A primeira técnica selecionou os 500 vocábulos mais frequentes divididos entre vídeos educacionais e não educacionais, utilizada nos Experimentos 1 e 2. Aqui, cada comentário foi convertido em um vetor de 500 elementos, com cada elemento indicando a frequência de ocorrência de um vocábulo específico no comentário.

A segunda técnica empregou o CountVectorizer, que realiza a contagem da frequência das palavras presentes nos comentários, distinguindo-se da primeira abordagem por considerar os termos mais comuns em todo o dataset, ao invés de focar nos mais frequentes de cada categoria de vídeo. Essa característica possibilita que a técnica de vocábulos mais frequentes por categoria dê mais importância aos termos importantes para cada grupo, evitando que a classe com maior volume de comentários (não educacional) tenha uma vantagem injusta devido à maior presença de termos específicos.

Para implementar o CountVectorizer e adaptar o dataset, foram testados diferentes valores para o parâmetro max_features, especificamente 500, 1000 e 3000 features, ajustando os parâmetros min_df=2, max_df=0.7 e excluindo as stop_words em português, a fim de observar o impacto dessas variações no desempenho da classificação.

O parâmetro max_features definiu que cada comentário será convertido em um vetor de 500, 1 mil ou 3 mil posições, o min_df significa que as palavras que poderão ser utilizadas para gerar nossa matriz deverão ter uma frequência de pelo menos 2, ou seja, em todos os meus comentários as palavras devem ter sido utilizadas pelo menos 2 vezes, o max_df=0.7 especifica a frequência máxima que uma palavra pode ocorrer nos comentários, nesse caso, palavras que aparecem em mais de 70% dos documentos são excluídas e, por fim, o stop_words='portuguese' remove os stopwords, no projeto esse comando é redundante pois os dados já haviam sido pré-processados e os stopwords removidos.

Na realização dos experimentos mencionados, a divisão do dataset foi feita seguindo a proporção de 80% para treinamento e 20% para teste. Essa divisão é uma prática comum em

aprendizagem de máquina, onde uma maior parte do dataset é usada para treinar o modelo e uma menor parte é reservada para avaliar sua performance em dados não vistos durante o treinamento.

Um elemento importante na configuração do experimento foi a utilização de um "random state" fixo, especificamente o valor 42. O random state é um parâmetro usado em algoritmos que envolvem aleatoriedade, como a divisão de datasets em conjuntos de treino e teste. Ao fixar esse parâmetro, garante-se a replicabilidade dos resultados. Isso significa que cada vez que o processo de divisão do dataset é realizado com o mesmo random state, a distribuição dos dados nos conjuntos de treino e teste é idêntica. Essa consistência é de grande importância para comparar o desempenho de diferentes algoritmos de maneira justa e confiável, pois assegura que todos eles são testados e treinados com exatamente os mesmos dados.

Ao manter a divisão do dataset constante, é possível fazer uma avaliação mais precisa de qual algoritmo de aprendizagem de máquina apresenta o melhor desempenho para o conjunto de dados específico, contribuindo para a escolha da abordagem mais adequada para resolver o problema em questão.

Realizaram-se quatro experimentos para avaliar a acurácia de diversos algoritmos de aprendizado de máquina em diferentes datasets. Os algoritmos testados foram Rede Neural Simples (Neural Network), Rede Neural Convolucional (Convolutional Neural Network), Random Forest, SVM (Máquina de Vetores de Suporte) e MLP (Perceptron Multicamadas, uma Rede Neural Simples mas com esse nome no Scikit-learn). Quanto aos datasets, foram empregadas distintas configurações de features: o primeiro dataset continha a metodologia empregada nos outros experimentos, com os 500 vocábulos mais frequentes; o segundo, 500 features obtidas por meio do CountVectorizer; o terceiro, 1000 features também extraídas através do CountVectorizer; e, por último, o quarto dataset incorporou 3000 features, igualmente derivadas utilizando o CountVectorizer.

A Figura 26 exibe o gráfico das acurácias obtidas pelos algoritmos de aprendizado de máquina, que utilizaram os 500 vocábulos mais frequentes, distribuídos de forma equitativa entre as duas classes, educacional e não educacional. Este gráfico foi desenvolvido com base na abordagem apresentada anteriormente no estudo. Já a Figura 27 apresenta as acurácias para o CountVectorizer com 500 features. Na Figura 28 é apresentado as acurácias para o CountVectorizer com 1000 features. Por fim, na Figura 29 é apresentado as acurácias para o CountVectorizer com 3000 features.

A comparação entre diferentes algoritmos de aprendizado de máquina neste experimento destacou a superioridade da Rede Neural Convolucional (CNN) em todos os cenários avaliados. Notavelmente, a CNN alcançou uma acurácia de até 87,67% ao utilizar um dataset vetorizado com 1000 vocábulos através do CountVectorizer. Contudo, é essencial observar que a diferença de desempenho em relação à nossa abordagem, que prioriza os vocábulos mais frequentes, foi mínima, situando-se em apenas 0,3%. Essa leve discrepância sublinha que, apesar de a configuração do CountVectorizer com 1000 features ter obtido a maior acurácia, o custo em termos de recursos computacionais para processar um conjunto tão extenso de features pode não ser justificável.

A metodologia focada na seleção dos vocábulos mais frequentes provou ser não apenas com-

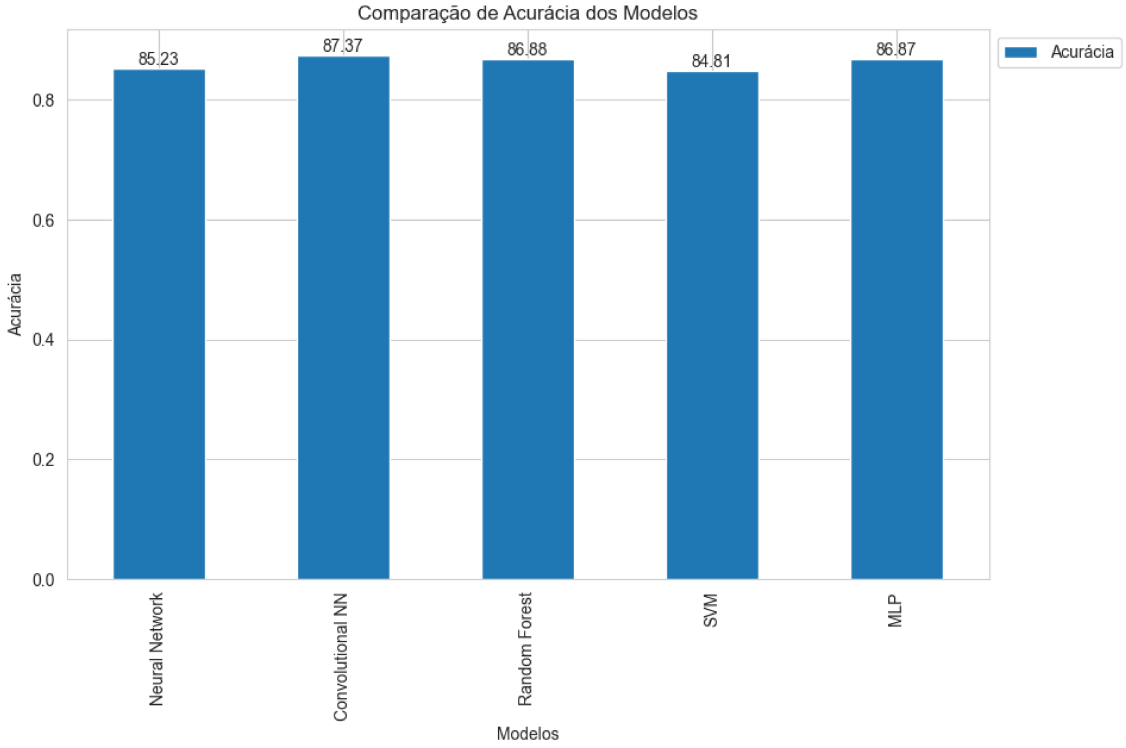


Figura 26 – Experimento 4: Acurácias com utilização dos 500 vocábulos mais frequentes

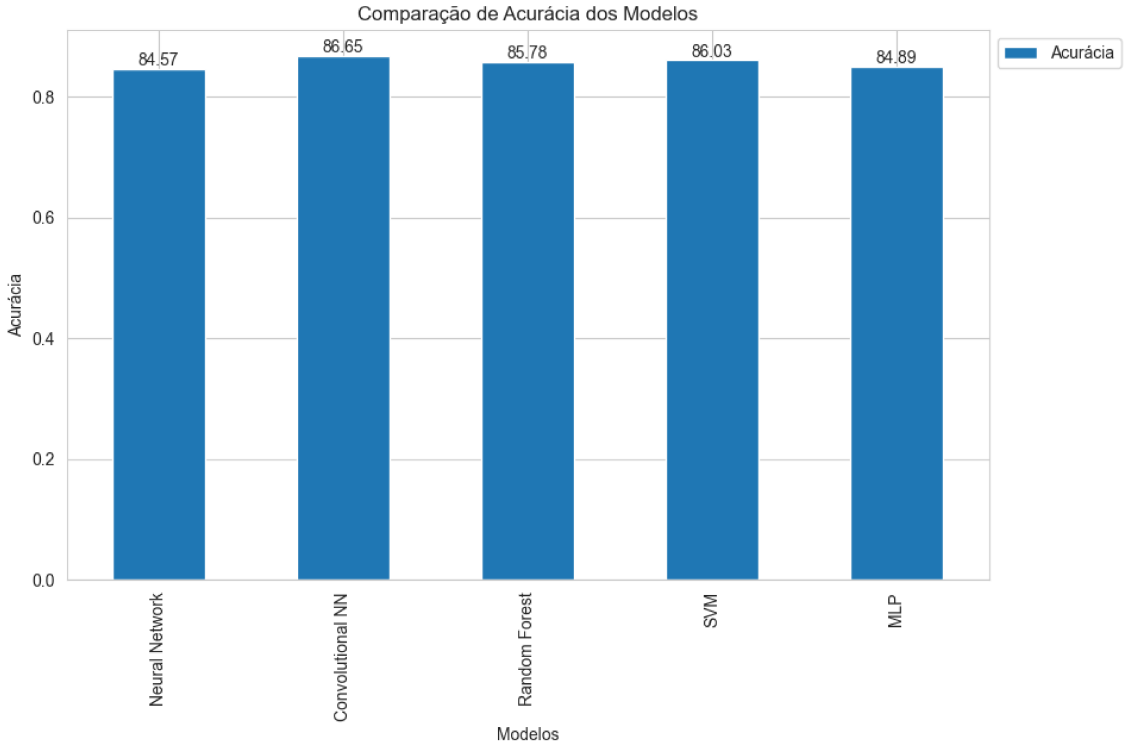


Figura 27 – Experimento 4: Acurácias com utilização dos 500 features CountVectorizer

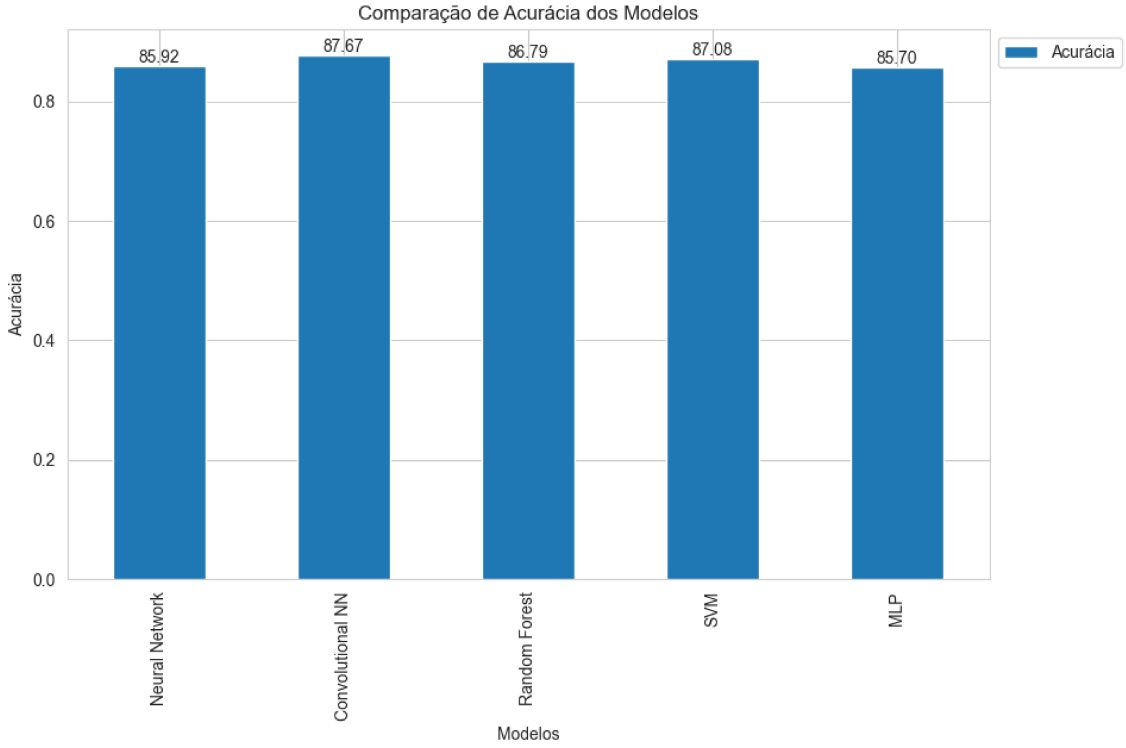


Figura 28 – Experimento 4: Acurácias com utilização dos 1000 features CountVectorizer

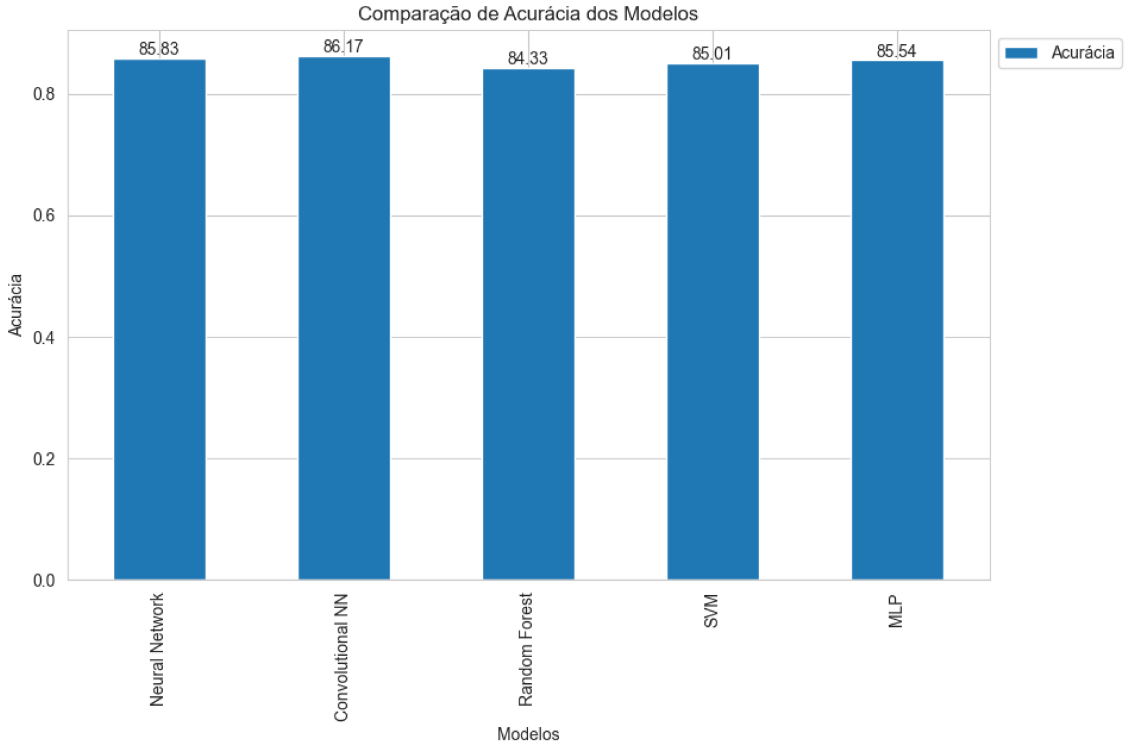


Figura 29 – Experimento 4: Acurácias com utilização dos 3000 features CountVectorizer

parável em termos de eficácia, mas também mais eficiente do ponto de vista computacional, especialmente em contraste com o uso do CountVectorizer configurado para 500 e 3000 features. Esse resultado indica uma otimização no uso de recursos computacionais, sem reduzir a qualidade da classificação.

Notadamente, a estratégia que utiliza 500 features (vocábulos) se mostrou a mais eficaz para o problema, oferecendo acurácia competitiva e vantagens no tempo de processamento. A diminuição no número de iterações necessárias para treinamento e predição, graças ao menor número de features, resulta em um desempenho mais rápido, otimizando tanto o tempo quanto os custos computacionais envolvidos.

A análise detalhada também indicou que a utilização de 3000 features não resultou em melhorias marcantes na acurácia, sugerindo que um aumento para além de 1000 features pode não compensar devido à queda nas taxas de acurácia. Esse achado sugere um ponto de equilíbrio na relação entre a quantidade de features e a eficiência da classificação, onde um incremento no número de features pode introduzir uma complexidade desnecessária ao modelo, sem melhorar a precisão e aumentando o custo computacional.

5.2.5 Experimento 5

O Experimento 5 foi projetado para comparar as duas variações apresentadas para a classificação dos vídeos como educacionais ou não educacionais: rígida e flexível.

Na variação rígida, todos os comentários de um vídeo são utilizados conjuntamente para classificar o vídeo como educacional ou não educacional. Essa metodologia toma uma decisão categórica, recomendando apenas vídeos que o algoritmo tem "certeza" que são educacionais.

Por outro lado, a variação flexível adota uma estratégia mais maleável. Cada comentário é classificado individualmente como educacional ou não educacional e após a classificação de todos os comentários, é aplicada uma fórmula específica (Equação 11) para calcular o "grau de certeza" do vídeo ser educacional. Esta metodologia permite uma maior flexibilidade na seleção de materiais educacionais, pois não se restringe apenas aos vídeos onde tem a "certeza" de serem educacionais. Além disso, a possibilidade de recomendar vídeos com base nos seus comentários permite que a abordagem flexível se ajuste melhor às particularidades de cada vídeo e ao engajamento do seu público. Também proporciona aos usuários uma compreensão da motivação por trás da recomendação de determinado vídeo, oferecendo uma métrica adicional para a seleção de vídeos, além dos metadados, que é o grau de certeza.

O experimento objetiva determinar quando um vídeo é considerado educacional com base nos comentários e buscar a resposta paras as questões: "Qual grau de certeza deve ser utilizado para que a metodologia flexível apresente resultados equivalentes ou melhores que a metodologia rígida?", "A classificação de cada comentários pode fornecer melhor classificação?" e "Qual performa melhor para diferentes quantidades de comentários?". Além dessas perguntas, o experimento busca responder qual modelo será utilizado no Sistema de Recomendação, LOIS.

Este estudo foi conduzido utilizando os algoritmos de aprendizado de máquina, Random Forest, Rede Neural Simples, Rede Neural Profunda Densa e Rede Neural Convolucional. Para

a implementação do Random Forest, utilizou-se a biblioteca Scikit-Learn, ao passo que as Redes Neurais foram desenvolvidas com o auxílio do TensorFlow.

Este estudo deu particular atenção às métricas de Acurácia e F1-Score, ainda que outras métricas, tais como Recall e Precisão, tenham sido apresentadas. A preferência pelas métricas Acurácia e F1-Score deve-se ao fato de que a Acurácia indica a capacidade do classificador em reconhecer corretamente novos dados, enquanto o F1-Score fornece um panorama do desempenho do modelo, cobrindo tanto as classificações previstas quanto as classificações reais na categoria Educacional.

O dataset utilizado consiste em 200 vídeos, divididos igualmente entre educacionais e não educacionais, representando uma amostra do dataset de 500 vídeos. Este conjunto de dados foi usado para testar ambas as abordagens, rígida e flexível. Na abordagem flexível, foram experimentados diferentes valores para a equação de probabilidade - Grau de Certeza - (Equação 11), entre 10% até 100%, em incrementos de 10%, para classificar um vídeo como educacional. O objetivo foi examinar o impacto de diferentes Graus de Certeza ("confiança") na classificação.

O experimento apresentou desafios significativos, principalmente na transformação e avaliação do dataset sob a perspectiva da variação flexível. As etapas críticas desse processo incluíram:

1. **Divisão do Dataset:** Inicialmente, foi essencial dividir os IDs dos vídeos em conjuntos de treinamento e teste. Esta divisão cuidadosa assegurou que ambos os conjuntos fossem representativos do dataset total, garantindo a validade dos resultados obtidos e permitindo uma comparação eficaz da abordagem flexível. Utilizou-se a técnica de validação cruzada de 10-folds para a validação do experimento. Nessa divisão, é separado apenas os IDs, que serão utilizados para treinamento e para teste.

2. **Carregamento de Comentários:** Após a divisão, os comentários associados a cada ID de vídeo são separados. Esta etapa foi crucial para manter uma correspondência precisa entre os vídeos e seus respectivos comentários, sendo fundamental para a subsequente análise. Nesse momento, os IDs que serão utilizados para treinamento tem os seus comentários e classes carregados no dataset para treinamento. Esses comentários, juntamente com a classe de onde foram extraídos, foram submetidos ao algoritmo para o processo de treinamento.

3. **Classificação dos Vídeos:** Esta fase consistiu na seleção e validação dos vídeos de teste e compreendeu três subetapas:

    a) *Identificação de Cada Vídeo de Teste:* Nessa etapa, cada ID dos vídeos selecionados para o treinamento serão pegos individualmente para serem classificados. Ao pegar um ID os comentários obtidos e separados para serem classificados.

    b) *Classificação de Comentários:* Em seguida, aplica-se um algoritmo de aprendizado de máquina para prever a classe de cada comentário do ID selecionado.

    c) *Cálculo de Graus de Certeza:* Com as predições em mãos, calculou-se o grau de certeza de cada vídeo ser classificado como educacional, usando a fórmula 11. Se o grau de certeza calculado atingisse um limiar predefinido, o vídeo era classificado

como educacional; do contrário, como não educacional. Foram testados os limiares entre 10% e 100%, de 10 em 10.

Essa sequência de passos é realizada para ID de vídeo que vai ser utilizado para teste. Dessa forma, cada ID do conjunto de dados de teste, passará por todas essas 3 etapas e após o cálculo do Grau de Certeza, temos a classificação do vídeo como Educacional ou Não.

4. **Verificação da Classificação:** A última etapa consistiu na validação das classificações realizadas com base no grau de certeza. Foi essencial avaliar as classificações, comparando-as com a natureza real (educacional ou não) dos vídeos, para determinar a eficácia da abordagem probabilística em comparação com a determinística.

Essas etapas, interconectadas e complexas, ressaltaram a natureza desafiadora do experimento. Além disso, a precisão no alinhamento dos comentários com os vídeos correspondentes e a análise dos graus de certeza adicionaram uma camada de complexidade ao estudo.

Foram realizados 10 experimentos para a avaliação da abordagem rígida. Cada experimento representa a média de uma execução do 10-folds cross-validation. A tabela 19 apresenta a Acurácia dos classificadores para a abordagem rígida. Já a tabela 20 apresenta a F1-Score obtida nos experimentos.

Tabela 19 – Experimento 5: Acurácia para a Abordagem Rígida

Exp.	**Random Forest**	**NN**	**DNN**	**CNN**
1	92,08%	93,10%	93,65%	95,18%
2	93,10%	93,51%	95,73%	95,58%
3	90,96%	94,65%	94,15%	96,20%
4	91,49%	93,54%	94,15%	95,09%
5	93,13%	92,57%	94,65%	95,67%
6	92,54%	93,57%	94,09%	96,26%
7	92,02%	93,65%	94,18%	94,01%
8	93,10%	94,12%	96,23%	95,12%
9	91,43%	94,62%	92,60%	94,59%
10	91,46%	95,23%	95,23%	95,67%
Média	92,13%	93,86%	94,46%	95,34%

Os resultados deste experimento evidenciam a superioridade das Redes Neurais em comparação ao Random Forest, com um aumento de mais de 1,73% em termos de acurácia ao empregar uma Rede Neural Simples. Além disso, observa-se um desempenhos notáveis com o uso das Redes Neurais, alcançando acurácias de até 96,26% e um F1-Score de até 96,12%. Ademais, percebe-se que a Rede Neural Convolucional apresenta melhores valores médios para a Acurácia e para o F1-Score.

Tais resultados são particularmente notáveis considerando a complexidade e a diversidade do universo de estudo, composto por comentários de usuários em vídeos educacionais e não educacionais. A elevada acurácia e precisão alcançadas sublinham a capacidade das Redes

Tabela 20 – Experimento 5: F1-Score para a Abordagem Rígida

Exp.	**Random Forest**	**NN**	**DNN**	**CNN**
1	91,19%	92,26%	93,24%	94,80%
2	92,93%	93,36%	95,63%	95,31%
3	90,65%	94,31%	93,86%	96,03%
4	91,20%	93,46%	93,79%	94,96%
5	93,06%	92,22%	94,48%	95,61%
6	92,07%	93,34%	93,85%	96,07%
7	91,71%	93,50%	93,92%	93,75%
8	92,30%	94,04%	96,12%	94,80%
9	91,16%	94,35%	92,36%	94,38%
10	91,29%	95,04%	94,53%	95,52%
Média	91,76%	93,59%	94,18%	95,12%

Neurais em distinguir eficazmente entre as categorias definidas, demonstrando a robustez dos modelos em contextos de classificação detalhada.

As matrizes de confusão para os algoritmos Random Forest, Rede Neural Simples, Rede Neural Profunda e Rede Neural Convolucional são apresentadas, respectivamente, nas Figuras 30, 31, 32 e 33. Essas matrizes foram elaboradas com base em todas as classificações realizadas nos 10 experimentos realizados. Adicionalmente, a Tabela 21 detalha os valores obtidos para as métricas de Acurácia, Precisão, Recall e F1-Score para cada um dos algoritmos mencionados.

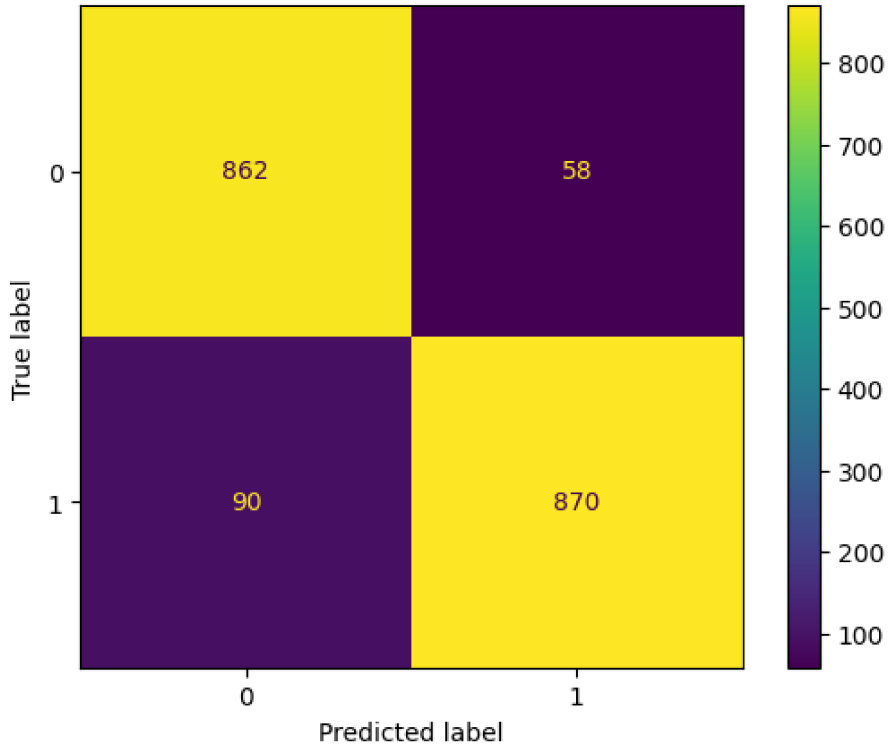


Figura 30 – Experimento 5: Matriz de confusão para o Random Forest na Abordagem Rígida

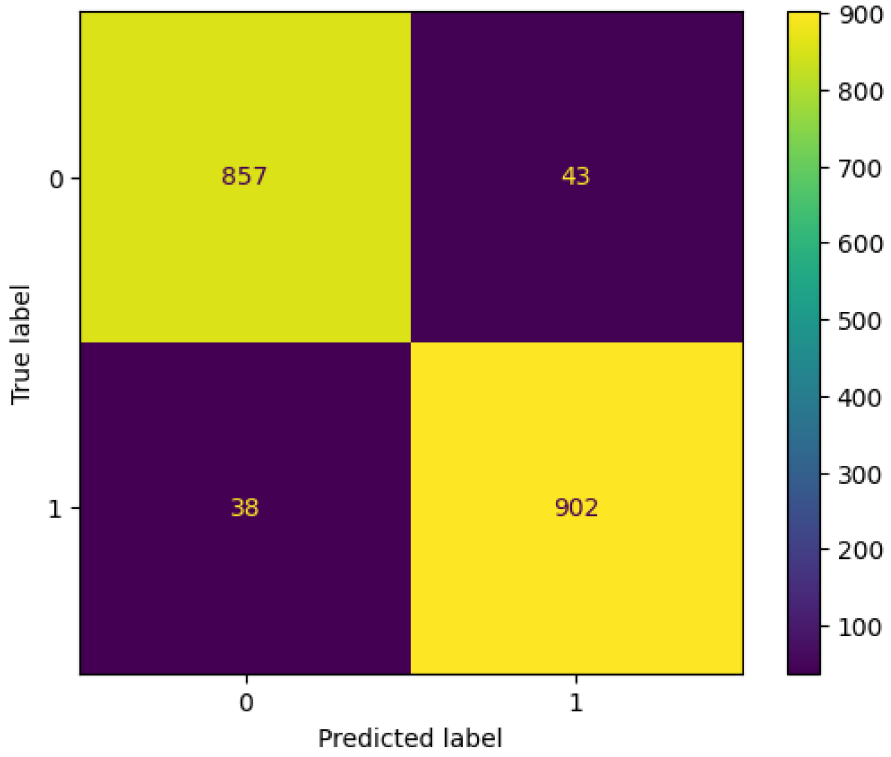


Figura 31 – Experimento 5: Matriz de confusão para a Rede Neural na Abordagem Rígida

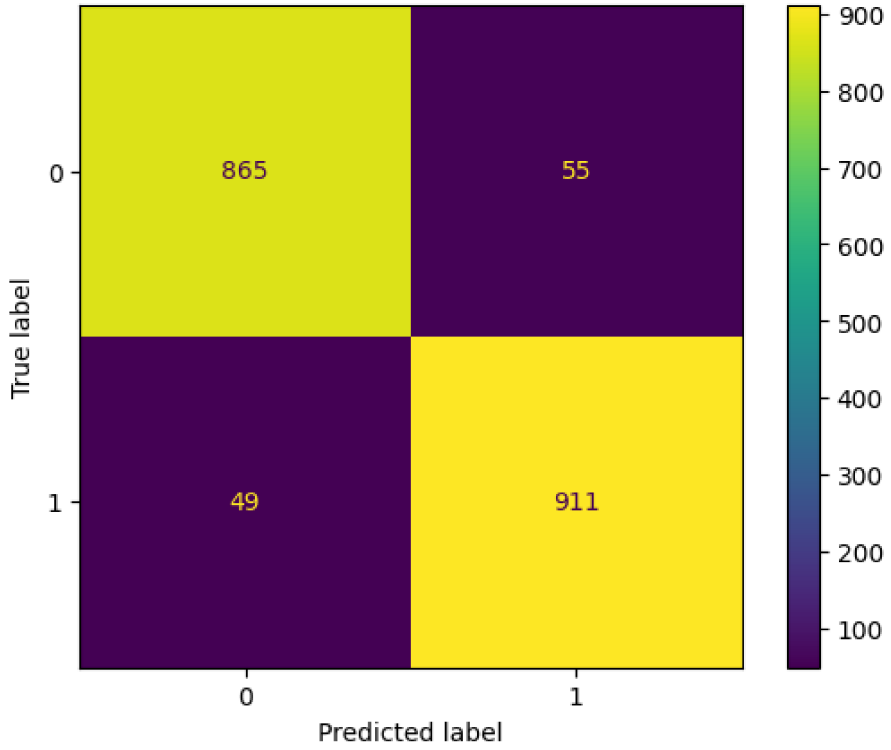


Figura 32 – Experimento 5: Matriz de confusão para a Rede Neural Profunda Densa na Abordagem Rígida

Tabela 21 – Experimento 5: Valores das Métricas para a Abordagem Rígida

Métrica	**Random Forest**	**NN**	**DNN**	**CNN**
Acurácia	92,13%	93,86%	94,47%	95,34%
Precisão	92,17%	93,82%	94,66%	95,35%
Recall	92,16%	94,11%	94,37%	95,44%
F1-Score	91,76%	93,59%	94,18%	95,12%

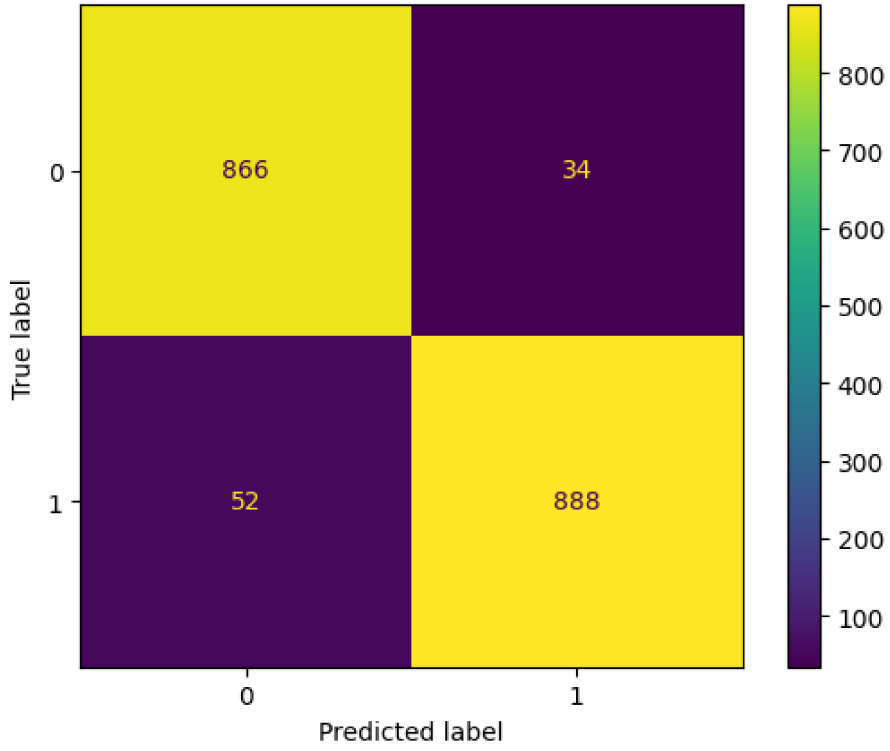


Figura 33 – Experimento 5: Matriz de confusão para a Rede Neural Convolucional na Abordagem Rígida

Percebe-se que de modo geral, a CNN performou melhor que os outros algoritmos, apresentando 95% de acertos para todas as métricas, Acurácia, Precisão, Recall, F1-Score.

A Tabela 22 mostra os resultados de acurácia alcançados utilizando a abordagem flexível para classificar os vídeos como educacionais ou não educacionais. Cada valor apresentado é a média de três execuções do método de validação cruzada de 10-folds, aplicadas para um específico grau de certeza, exceto para o grau de certeza de 40%, onde foi aplicada a média de 10 execuções da validação cruzada. Por outro lado, a Tabela 23 detalha os valores de F1-Score obtidos nos mesmos experimentos.

Foi observado que os resultados mais expressivos para todos os algoritmos foram alcançados ao definir o "grau de certeza" em 40%, isto é, um vídeo é classificado como educacional se mais de 40% dos seus comentários forem considerados educacionais. Notavelmente, o valor mais baixo de acurácia registrado em todos os experimentos foi de 95,14%, usando o algoritmo Random Forest com o "grau de certeza" ajustado para 40%. Adicionalmente, foi evidenciado que os algoritmos de Redes Neurais apresentaram um desempenho ligeiramente superior em termos de

Tabela 22 – Experimento 5: Acurácia para a Abordagem Flexível

Exp.	Certeza	Random Forest	NN	DNN	CNN
1	**10%**	60,90%	57,52%	58,05%	60,51%
2	**20%**	83,67%	81,94%	82,14%	82,80%
3	**30%**	94,67%	92,88%	92,90%	93,95%
4	**40%**	95,14%	95,66%	95,18%	95,70%
5	**50%**	93,74%	94,64%	94,28%	95,17%
6	**60%**	83,75%	88,04%	88,04%	87,49%
7	**70%**	73,44%	77,38%	77,19%	77,17%
8	**80%**	58,58%	61,14%	61,50%	62,07%
9	**90%**	49,91%	50,09%	49,91%	50,09%
10	**100%**	49,39%	49,39%	49,39%	49,39%

Tabela 23 – Experimento 5: F1-Score para a Abordagem Flexível

Exp.	Certeza	Random Forest	NN	DNN	CNN
1	**10%**	72,31%	70,52%	70,84%	72,13%
2	**20%**	86,30%	85,13%	85,42%	85,81%
3	**30%**	95,03%	93,69%	93,79%	94,55%
4	**40%**	95,42%	95,93%	95,47%	96,07%
5	**50%**	93,79%	94,76%	94,42%	95,25%
6	**60%**	80,73%	86,75%	86,86%	85,80%
7	**70%**	65,10%	71,13%	70,88%	70,75%
8	**80%**	30,16%	36,04%	37,85%	38,93%
9	**90%**	3,67%	4,27%	6,67%	4,27%
10	**100%**	1,67%	1,67%	1,67%	1,67%

acurácia, com a Rede Neural Convolucional alcançando 95,70% e tanto a Rede Neural Simples quanto a Rede Neural Profunda atingindo acurácias superiores a 95%, todos sob o limiar de certeza de 40%.

Os melhores resultados, tanto em acurácia quanto em F1-Score, foram identificados precisamente com o grau de certeza definido em 40%. Adicionalmente, observou-se que a abordagem flexível, com "grau de certeza" de 40%, fez com que todos os algoritmos superassem suas próprias médias obtidas pela Abordagem Rígida, reforçando a eficácia dessa metodologia flexível na classificação de vídeos educacionais. Além disso, ao definir a certeza em 40% e utilizar a Rede Neural Simples, os resultados superaram os obtidos com a CNN na Abordagem Rígida, evidenciando a superioridade da abordagem flexível.

Vale ressaltar que a Rede Neural Simples possui uma complexidade de operações muito menor do que uma Rede Neural Convolucional, resultando em um processamento mais veloz, o que é melhor aproveitado quando desejamos um sistema com respostas rápidas.

A Figura 34 ilustra um gráfico detalhando como a acurácia varia de acordo com diferentes "Graus de Certeza". Este gráfico oferece uma visão clara da relação entre o nível de certeza estabelecido para a classificação dos vídeos e a acurácia alcançada pelos algoritmos na identificação correta de vídeos educacionais. Em sequência, a Figura 35 exibe os resultados do F1-Score alcançados, proporcionando uma análise complementar sobre a eficácia dos modelos em ter-

mos de precisão e recall, fundamentais para avaliar o equilíbrio entre a identificação correta de verdadeiros positivos e a minimização de falsos positivos e falsos negativos.

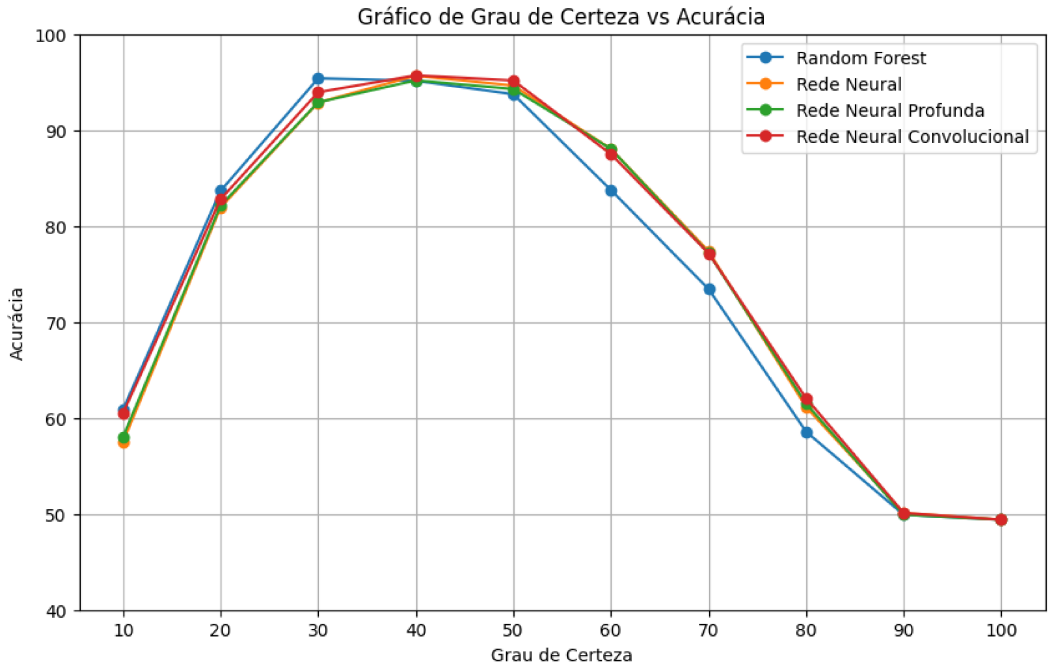


Figura 34 – Experimento 5: Gráfico de Acurácia por Grau de Certeza

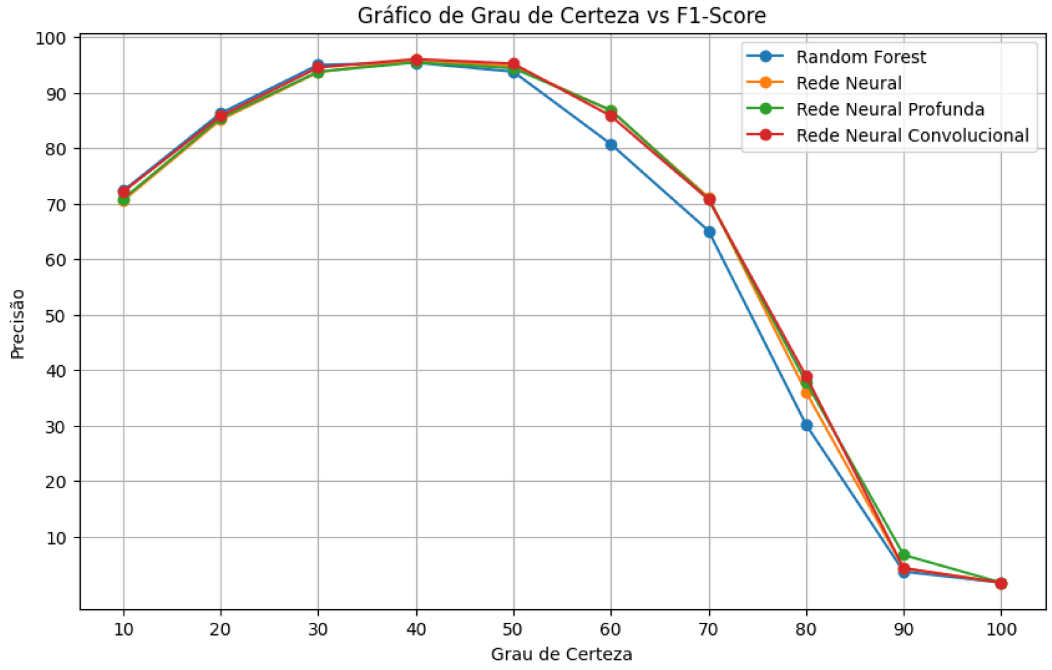


Figura 35 – Experimento 5: Gráfico de F1-Score por Grau de Certeza

Os gráficos revelam uma tendência parabólica na classificação dos vídeos, iniciando com acurácias mais baixas, na faixa de 60%, atingindo um ápice de desempenho quando o grau de

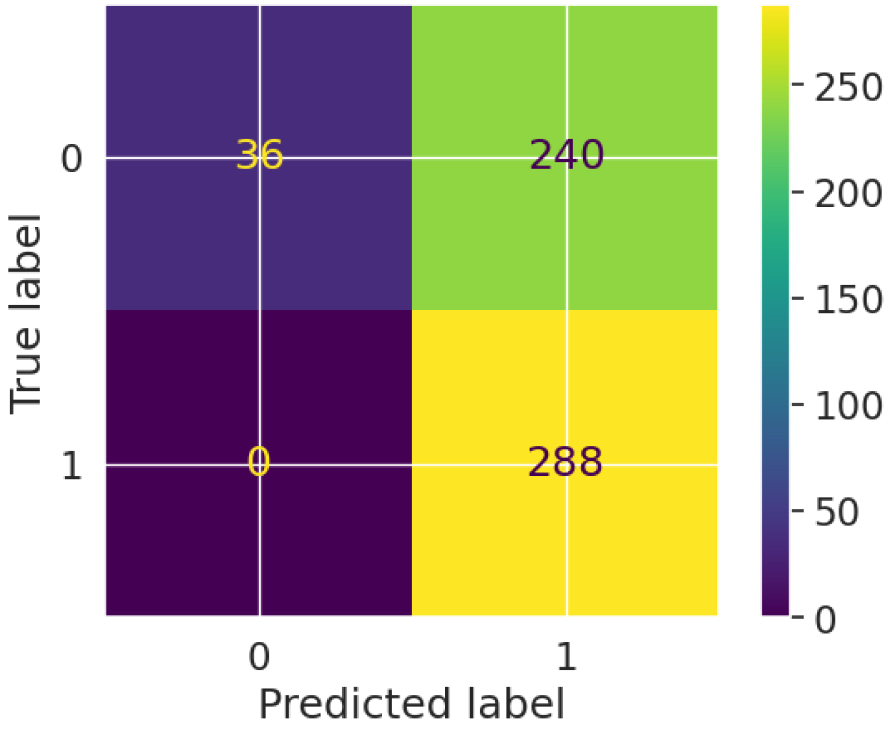


Figura 36 – Experimento 5: Matriz de Confusão para a Rede Neural Simples com 10% de Grau de Certeza

certeza se aproxima de 40%, marcando acurácias superiores a 95%, e posteriormente declinando, apresentando os valores mais baixos em graus de certeza mais elevados, por volta de 50%.

Definindo o grau de certeza em 10%, o que implica que um vídeo é considerado educacional se mais de 10% dos seus comentários forem classificados como tal, observa-se uma acurácia relativamente baixa. Esse declínio na precisão do modelo pode ser atribuído à classificação incorreta de uma grande quantidade de vídeos não educacionais como educacionais, comprometendo a habilidade do modelo em identificar com precisão os conteúdos verdadeiramente educacionais. A Figura 36 demonstra a matriz de confusão resultante da aplicação da Rede Neural Simples para a classificação de vídeos com um grau de certeza de 10%.

A matriz de confusão destaca um alto número de Falsos Positivos, indicando que vídeos não educacionais foram erroneamente categorizados como educacionais, alcançando um total de 240 casos. No entanto, a análise também revela uma quantidade baixa de Verdadeiros Negativos, com 36 vídeos sendo corretamente reconhecidos como não educacionais. Interessantemente, nesse patamar de certeza, o classificador conseguiu identificar corretamente todos os vídeos educacionais.

Além disso, observou-se que, ao estipular o grau de certeza em 90% ou 100%, tanto a acurácia quanto o F1-Score apresentaram declínios acentuados, aproximando-se de 50% para acurácia e de 5% para F1-Score. Essa queda nos indicadores sugere que níveis extremamente altos de certeza limitam drasticamente a capacidade do modelo de classificar adequadamente os vídeos, resultando em uma performance geral bastante reduzida. As matrizes de confusão para os graus de certeza de 90% e 100% são ilustradas nas Figuras 37 e 38, respectivamente, evidenciando os

desafios enfrentados pelo classificador nesses limiares de decisão.

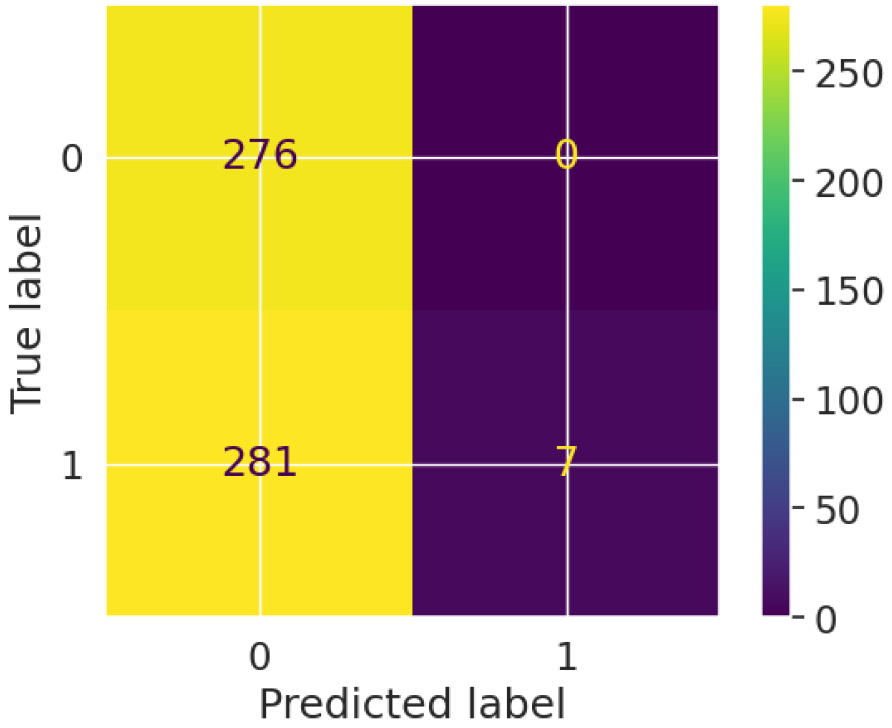


Figura 37 – Experimento 5: Matriz de Confusão para a Rede Neural Simples com 90% de Grau de Certeza

Nas duas matrizes de confusão apresentadas, observa-se zero casos de Falsos Positivos, indicando que o algoritmo, sob estas configurações específicas, é eficaz em corretamente identificar vídeos não educacionais como Verdadeiros Negativos. Contudo, é igualmente notável o elevado número de Falsos Negativos, o que sugere que o algoritmo tem dificuldades em reconhecer corretamente os Verdadeiros Positivos, ou seja, os vídeos efetivamente educacionais.

Essa dinâmica mostra que, ao elevar o grau de certeza necessário para classificar um vídeo como educacional, há uma diminuição na quantidade de Verdadeiros Positivos. Por consequência, apesar de um aumento no grau de certeza poder reduzir o número de Falsos Positivos(vídeos não educacionais classificados como educacionais), isso também resulta em uma queda acentuada no número de Verdadeiros Positivos. Tal cenário pode limitar severamente a aplicabilidade dos algoritmos, uma vez que a restrição excessiva na classificação de vídeos como educacionais poderia reduzir drasticamente as recomendações de conteúdo, comprometendo sua utilidade prática.

Os melhores resultados, tanto em acurácia quanto em F1-Score, foram alcançados ao se estabelecer um grau de certeza de 40%. Esse nível se mostrou o mais eficiente para todos os algoritmos avaliados. Entretanto, as avaliações subsequentes focarão na Rede Neural Simples, devido ao seu desempenho notável. Esta alcançou uma acurácia apenas 0,04% inferior à da Rede Neural Convolucional, que registrou a maior acurácia dentre os modelos testados. Além disso, a simplicidade operacional da Rede Neural Simples favorece uma execução mais rápida e demanda menos recursos computacionais, razões que reforçam sua seleção para o projeto LOIS.

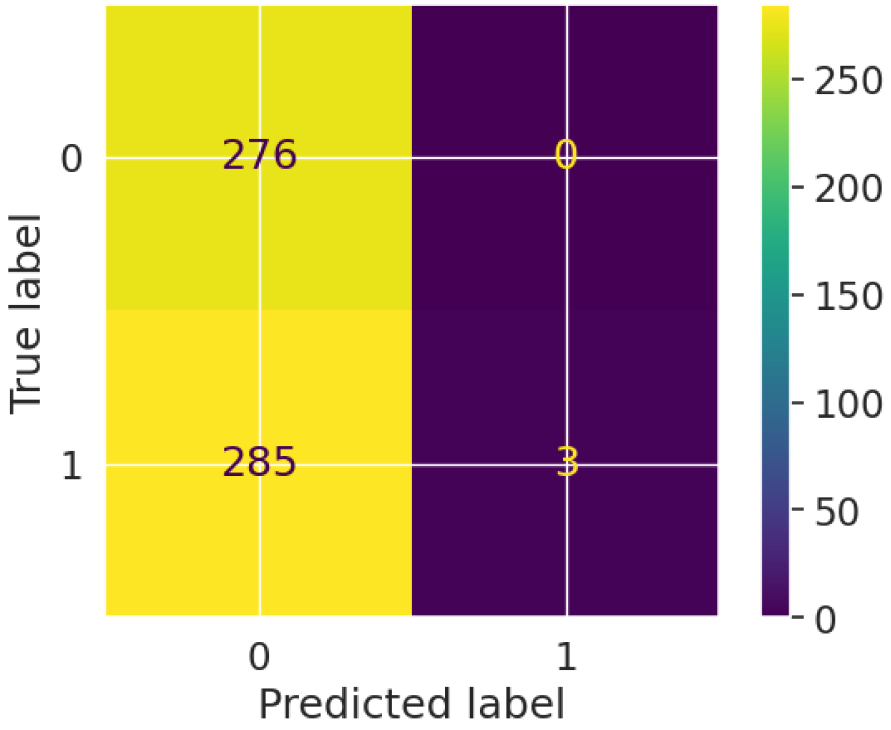


Figura 38 – Matriz de Confusão para a Rede Neural Simples com 100% de Grau de Certeza

A Tabela 24 detalha o tempo de processamento necessário para cada Rede Neural durante um único fold de uma validação cruzada de 10-folds, oferecendo informações sobre a eficiência de cada modelo em termos de tempo.

Tabela 24 – Experimento 5: Tempo de treinamento e teste das Redes Neurais

	200 Vídeos		
	NN	**DNN**	**CNN**
Treinamento	12,38 min	15,03 min	316,02 min / 5h e 26 min
Predição	1,08 seg	1,20 seg	5,44 seg
	500 Vídeos		
	NN	**DNN**	**CNN**
Treinamento	56,41 min	67,23 min / 1h e 7min e 23seg	1501 min / 25h e 01 min
Predição	4,55 seg	5,20 seg	28,53 seg

Observa-se que a Rede Neural Simples demanda menos tempo para treinamento e teste em comparação à Rede Neural Profunda e, de forma mais acentuada, em relação à Rede Neural Convolucional. Destaca-se também que, até mesmo para a tarefa de predição, a Rede Neural Simples prova ser cerca de 5 vezes mais rápida. Esta eficiência é crucial para aplicações em tempo real, onde a rapidez na resposta do modelo é fundamental para a usabilidade e eficácia do sistema.

A Tabela 25 apresenta os resultados de Acurácia, Precisão, Recall e F1-Score para a Rede Neural Simples quando utilizamos os graus de certeza de 30%, 40%, 50%, 60%, 70% e 80%, e ao utilizar todas as classificações realizadas.

Tabela 25 – Experimento 5: Resultados das Métricas para a Rede Neural Simples

Métrica.	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**
Acurácia	92,91%	**95,69%**	94,68%	88,12%	77,48%	61,34%
Precisão	88,27%	95,92%	97,43%	**98,68%**	98,20%	98,61%
Recall	**99,31%**	95,63%	92,01%	77,78%	56,94%	24,65%
F1-Score	93,46%	**95,77%**	94,64%	86,99%	72,09%	39,44%

Percebe-se, através da tabela, que os melhores resultados para Acurácia e F1-Score são obtidos quando o grau de certeza é definido como 40%, para Precisão quando o grau de certeza é de 60% e para Recall quando é 30%. As Figuras 39, 40, 41, 42, 43 e 44 apresentam as matrizes de confusão para os graus de certeza de 30%, 40%, 50%, 60%, 70% e 80%, reforça-se que a matriz de confusão representa todas as classificações realizadas. Para apresentar a mesma quantidade de classificações, padronizou-se os resultados também da Figura 40.

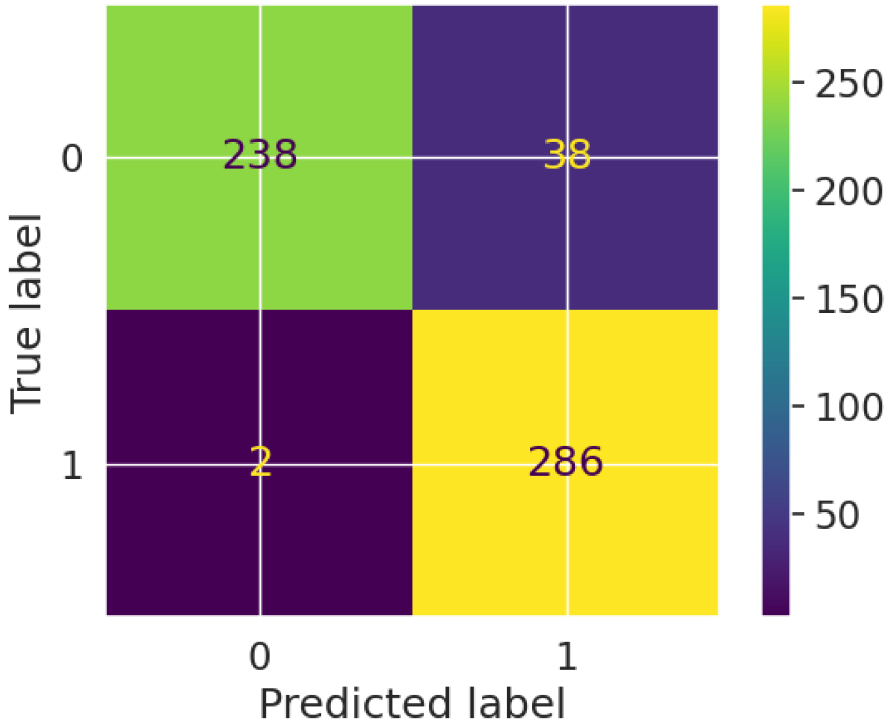


Figura 39 – Experimento 5: Matriz de Confusão para a Rede Neural Simples com 30% de Grau de Certeza

Conforme o grau de certeza exigido é elevado, observa-se uma diminuição nos casos de Verdadeiros Positivos (vídeos educacionais identificados corretamente como tal) e Falsos Positivos (vídeos não educacionais erroneamente classificados como educacionais), enquanto os Falsos Negativos (vídeos educacionais classificados como não educacionais) e Verdadeiros Negativos (vídeos não educacionais identificados corretamente) tendem a crescer. Isso mostra que, a cada aumento no grau de certeza, o número de classificações como educacional começa a diminuir.

Interessantemente, até um Grau de Certeza de 30%, o classificador consegue manter um Recall de 100%, identificando todos os vídeos educacionais como tal e classificando muitos não

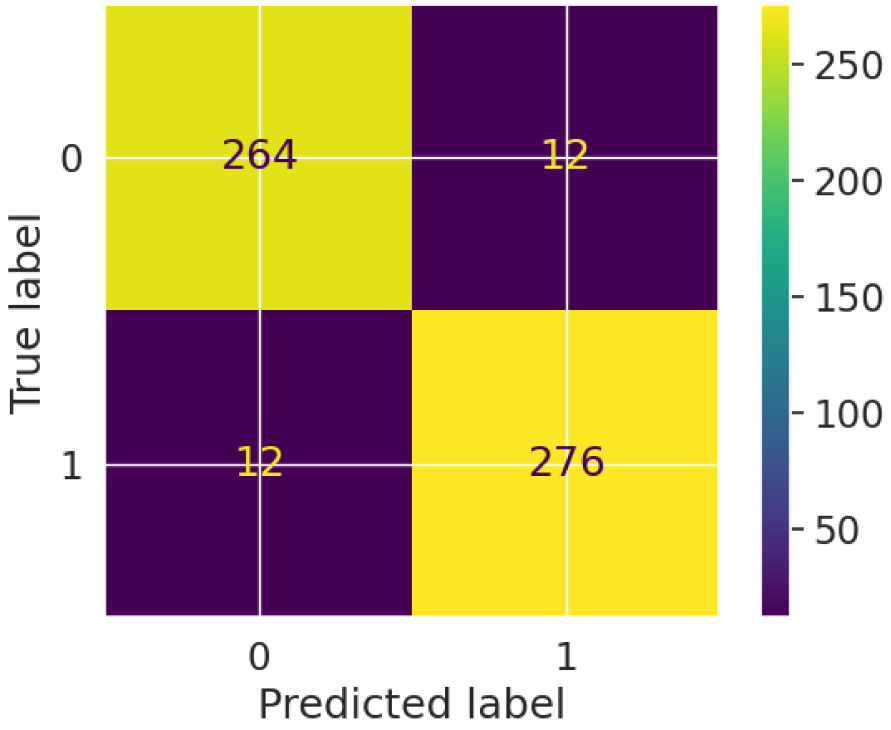


Figura 40 – Experimento 5: Matriz de Confusão para a Rede Neural Simples com 40% de Grau de Certeza - para 3 Folds

educacionais como educacionais. No entanto, ao ajustar o Grau de Certeza para 30%, o Recall cai para 99,31%, com o modelo classificando incorretamente 2 vídeos educacionais. Já com o Grau de Certeza ajustado para 80%, o modelo alcança uma Precisão de 98,61%, indicando que apenas 1 vídeo não educacional é identificados incorretamente como educacional, mas apenas 71 vídeos educacionais são reconhecidos como tal, resultando em 217 vídeos educacionais classificados como não educacionais. Isso leva a um Recall de 24,65% e um F1-Score de 39,44%.

Pode-se dizer que, quanto menor o grau de certeza, mais fácil é definir "tudo" como educacional, e conforme o grau de certeza aumenta, mais difícil fica selecionar um vídeo para colocá-lo no grupo dos educacionais. No entanto, conforme o grau de certeza aumenta, até 80%, a certeza de que um vídeo classificado como educacional é efetivamente educacional também aumenta. Entretanto, considerar apenas esse fator, a precisão, pode ser extremamente prejudicial, pois isso demonstra que as classificações educacionais diminuem consideravelmente.

Além disso, é importante notar que o processo de diferenciar vídeos educacionais dos não educacionais apresenta uma complexidade notável. Essa complexidade é largamente influenciada pela variabilidade dos comentários postados pelos usuários, bem como pelas características demográficas, culturais e comportamentais desses usuários que fornecem tais dados. A diversidade encontrada nos comentários, tanto em termos de sua pertinência quanto do contexto em que são inseridos, não só intensifica os desafios enfrentados durante a classificação, mas também demonstra a relevância dos resultados alcançados.

Ademais, para comparar os resultados de novos vídeos, foram selecionados 50 vídeos, seguindo os mesmos critérios para vídeos educacionais adotados anteriormente, e classificados de

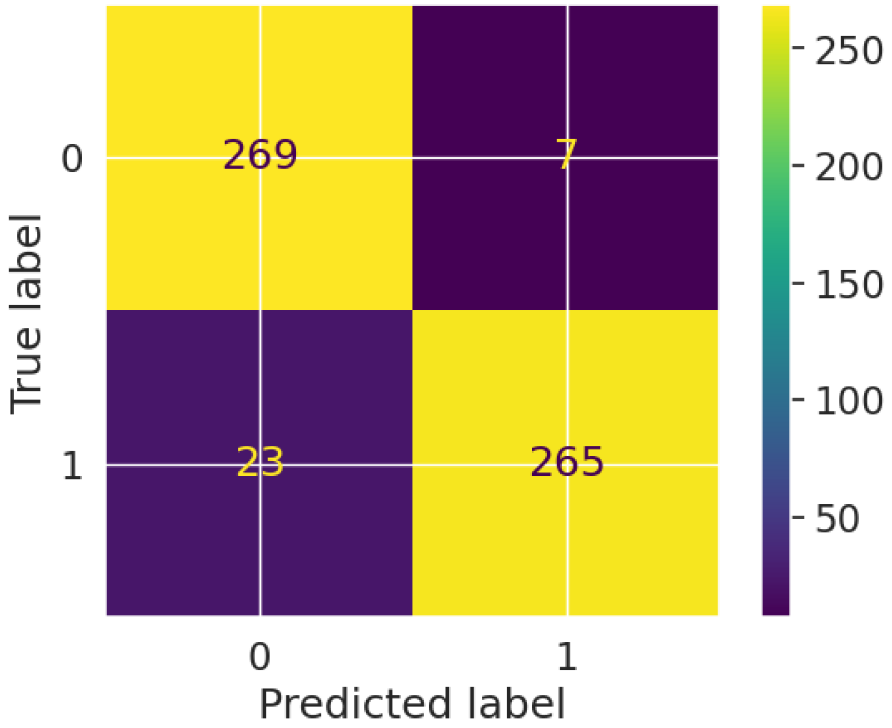


Figura 41 – Experimento 5: Matriz de Confusão para a Rede Neural Simples com 50% de Grau de Certeza

Tabela 26 – Experimento 5: Comparação de novos vídeos

Quantidade de Comentários	Abordagem Rígida	Abordagem Flexível
10	88%	94%
20	92%	92%
30	90%	98%
40	92%	96%
50	92%	94%
60	92%	94%
70	92%	94%
80	92%	94%
90	92%	94%
100	92%	96%
Todos	94%	94%
Média	91,64%	**94,55%**

acordo com ambas as abordagens. Os vídeos foram selecionados mantendo uma quantidade de comentários entre 700 e 2000. Além disso, foram comparadas diferentes quantidades de comentários, demonstrando como o comportamento do classificador varia conforme a quantidade de comentários em um vídeo.

F Figura 45 apresenta os vídeos selecionados e a Tabela 26 apresenta os resultados de acurácia para o melhor classificador de ambas as metodologias.

Percebe-se que a variação flexível obteve os melhores resultados, com uma média de 94,55%. Além disso, ambas as abordagens atingiram 94% de acurácia ao classificar os novos vídeos,

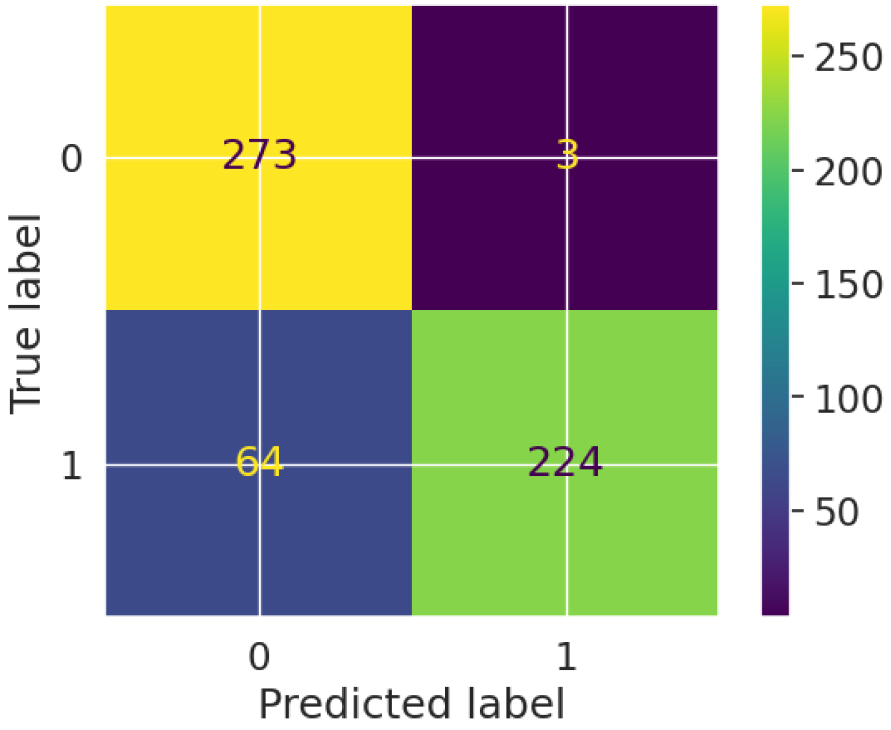


Figura 42 – Experimento 5: Matriz de Confusão para a Rede Neural Simples com 60% de Grau de Certeza

entretanto, quando não se utiliza todos os comentários, a variação rígida apresenta resultados menores, obtendo resultados até 6 pontos inferiores, o que é elevado e demonstra a necessidade do classificador ter uma grande quantidade de comentários para representar um vídeo e classificá-lo corretamente, o que é demonstrado quando se utiliza apenas 10 comentários, onde a acurácia cai de 94% para 88%.

Diferentemente, a metodologia flexível é mais maleável e adaptável, conseguindo manter uma acurácia de 94% com o dataset completo ou com 10 comentários. Além disso, a metodologia flexível performou melhor com quase todas as quantidades de comentários testadas e apenas com 20 comentários que ela obteve uma acurácia de 92%, igual a metodologia rígida para a mesma quantidade de comentários.

Uma forma eficaz de comparar experimentos é por meio da análise estatística dos dados, utilizando dois testes estatísticos: o teste de Kolmogorov-Smirnov, que verifica se a amostra segue uma distribuição normal, e, caso não siga, o teste U de Mann-Whitney, que compara amostras de distribuições não normais para avaliar se há diferença significativa entre elas.

O teste de Kolmogorov-Smirnov avalia duas hipóteses concorrentes: a hipótese nula (H0), que assume que os dados seguem uma distribuição normal, e a hipótese alternativa (H1), que assume que os dados não seguem uma distribuição normal. O teste gera um valor-p que, se inferior a um determinado nível de significância (geralmente 0,05), leva à rejeição da hipótese nula, indicando que os dados não seguem a distribuição normal. No presente caso, o valor-p obtido foi 0.000000, confirmando que a amostra não segue uma distribuição normal.

Após esse resultado, utilizou-se o teste U de Mann-Whitney, indicado para comparar dois

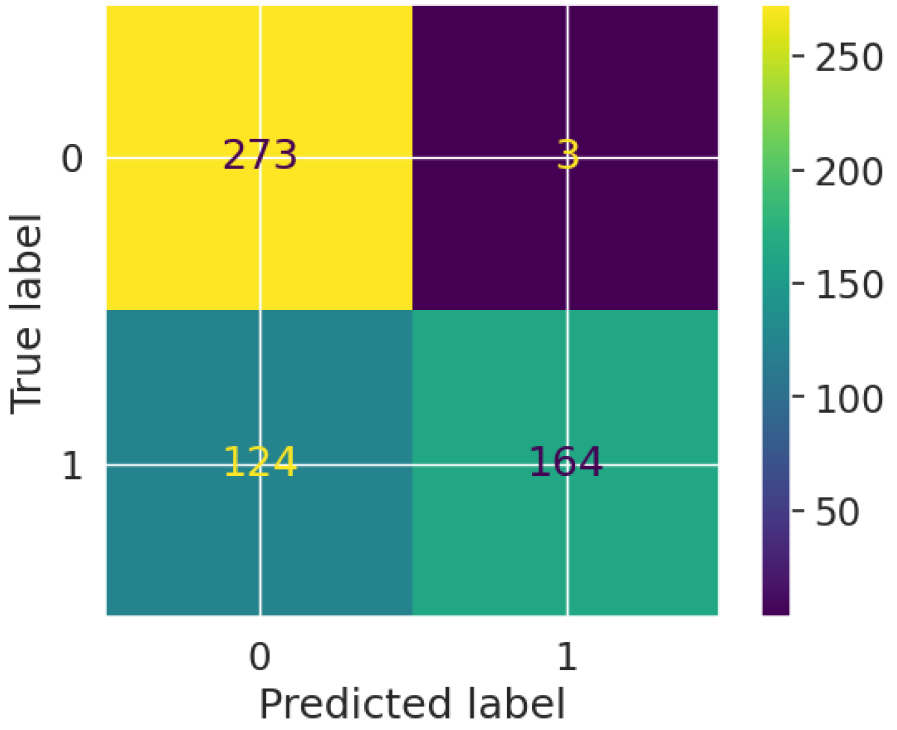


Figura 43 – Experimento 5: Matriz de Confusão para a Rede Neural Simples com 70% de Grau de Certeza

grupos independentes e verificar se pertencem ou não à mesma população, ou seja, se existem diferenças estatisticamente significativas entre as distribuições dos atributos dos grupos. Esse teste também avalia duas hipóteses concorrentes: a hipótese nula (H0), que assume que ambos os grupos são amostras da mesma população, e a hipótese alternativa (H1), que assume que os grupos pertencem a populações diferentes. Um valor-p inferior ao nível de significância (geralmente 0,05) indica que a diferença observada entre os grupos não é fruto do acaso.

Ao aplicar o teste U de Mann-Whitney nas acurácias das variações rígida e flexível, obteve-se um valor-p de 0.000323. Isso indica uma diferença estatisticamente significativa entre as duas amostras. Assim, há evidências suficientes para concluir que a metodologia flexível apresenta um desempenho significativamente diferente, sugerindo um desempenho superior em relação à metodologia rígida.

Ao concluir o experimento, conseguimos responder aos questionamentos iniciais: “Qual grau de certeza deve ser utilizado para que a metodologia flexível apresente resultados equivalentes ou melhores que a metodologia rígida?”, “A classificação individual de cada comentário pode fornecer uma melhor classificação?” e “Qual metodologia performa melhor para diferentes quantidades de comentários?”. Observou-se que ao utilizar um grau de certeza de 40%, a abordagem flexível supera a abordagem rígida em termos de resultados. Além disso, ficou evidente que a abordagem flexível permite uma classificação mais detalhada e eficaz de vídeos que apresentam variadas quantidades de comentários.

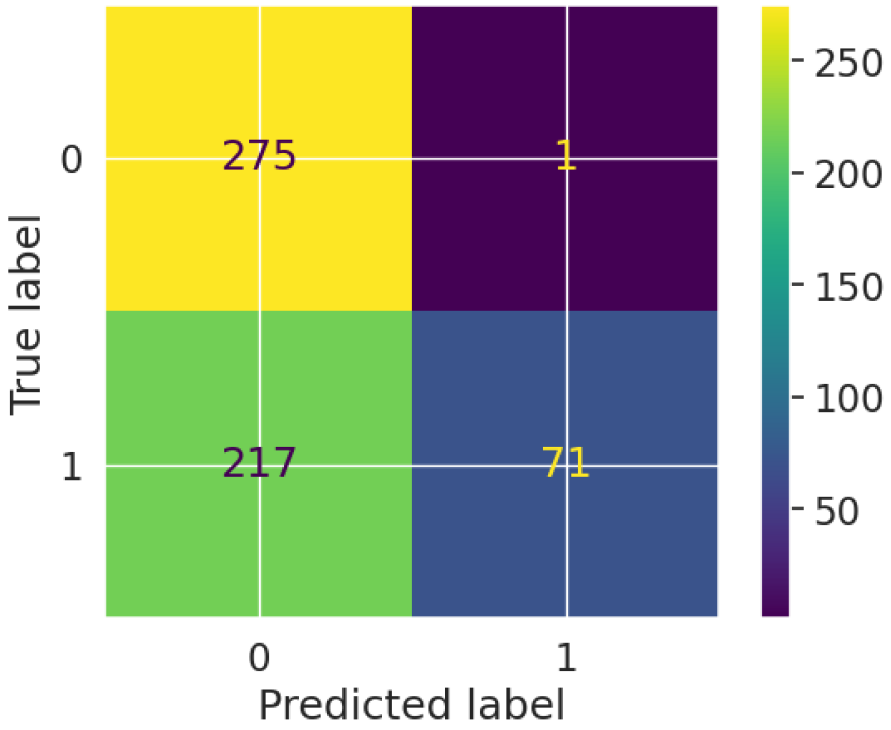


Figura 44 – Experimento 5: Matriz de Confusão para a Rede Neural Simples com 80% de Grau de Certeza

ID	Título	Classe
0TiqW54oos4	MITOS E CURIOSIDADES DOS BRASILEIROS SOBRE PARIS	Nao_edu
1QjHD7JEZHo	Tipos de Alunos em Uma Aula On-Line	Nao_edu
1kEMhX0-Vos	250km de TESLA DIRIGINDO SOZINHO no BRASIL! (Autopilot)	Nao_edu
3rnhyGqxXWo	PRIMEIRO DIA DE AULA NA ESCOLA DOS EUA - Família Brancoala	Nao_edu
553T10R-O_M	A PANDEMIA NAS FAVELAS | Cortes do Flow	Nao_edu
9F-mKTl1njg	REVELANDO AS MINHAS GAMBIARRAS!	Nao_edu
9I2wMm6ZFe4	Literatura - Aula 5: Romantismo (introdução)	Edu
Ao7Sz3gdmNY	DICA 27 - MMC e MDC Simples e Fácil (mínimo múltiplo comum e máximo divisor comum)	Edu
COS-76i-atc	Cálculo de Medicamento de um jeito fácil?	Edu
DL9llZP4a6k	ERA VARGAS - Resumo Desenhado	Edu
DfaOQlnf91U	Veja como a exposição nas redes sociais pode colocar sua privacidade em risco	Nao_edu
EDELn_QOvAE	Válvula Tesla | A física completa	Edu
I2taMQ3j6qo	AULÃO DE EXCEL 2024 | Nível Básico 01 | Curso de Excel	Edu
IP4MOFa9YPM	FUI ASSALTADO! Não Confie nas pessoas erradas!	Nao_edu
KMA24zhJBQ4	43 Sabores de KIT KAT que Voce Nunca Viu	Nao_edu
Kux5Utu-FOM	Aula de ballet para crianças - conexão TCM	Edu
LlzhuO8Mg5Q	Aula de espanhol #01: Cumprimentos e apresentações	Edu
NHFt7QTyUfs	Aula de Inglês #01: Aprenda as datas e os números	Edu
NmFtSgMpZm0	A pizza do SUBWAY	Nao_edu
ObQs0pwj_A8	Haikaiss e Projota - Pra Te Convencer (VÍDEO OFICIAL)	Nao_edu
Pr7d0zFeC7E	NOSSO MAIOR MICO COMO CASAL - Ep.1242 (Perguntas & Respostas)	Nao_edu
QbxPUFl3cuc	Yoga para Iniciantes: FORTALECENDO e ALONGANDO o CORPO TODO | Fernanda Yoga	Edu
Vb0iORewZDA	A criação dos computadores pessoais | Nerdologia	Edu
WJOB0iHCFhc	Química - Aula 5: Separação de misturas	Edu
WNPcFKlzBfg	De onde vem as notas musicais? | Nerdologia	Edu
XCZyx2IHdxA	INTRODUÇÃO à QUÍMICA: Massa, Volume, Densidade, Estados Físicos, Transformações| Aula 01 (Química I)	Edu
YC7PLXpmsTo	Monstros na Faculdade / 9 Materiais Escolares Para Monstros – Faça Você Mesmo!	Nao_edu
YRZcF0b2CWQ	Entrevista com Caio Castro | The Noite (07/10/20)	Nao_edu
_Ye8eeEH1_E	Aula de italiano #01: Vamos nos apresentar?	Edu
bxJbLb5oFtI	PEGADINHA - ASSALTANTE EDUCADO NA F4VELA DO RIO DE JANEIRO #DESAFIO 128	Nao_edu
cHD2DlUlKYI	VIAJANDO pelo INTERIOR da COREIA DO NORTE (2019)	Nao_edu
dt-EksrDMrs	Hamburguer Gigante Ft Lord e Haru	Nao_edu
eZfFpnvudR0	Conjuntos: Diferença e Complementar (Aula 4 de 4)	Edu
fuOzqtKJN5w	Testei uma Arma de 100 anos Usada na Primeira Guerra Mundial !!	Nao_edu
g_d6VZnaWwA	AS 50 PALAVRAS MAIS USADAS NO INGLÊS (com exemplos!) | Inglês para Iniciantes	Edu
hho15muKq4M	PROGRAMA EU FICO LOKO #04 - JÚLIO COCIELO	Nao_edu
j7v8dMisr78	Primeiras doses da vacina Coronavac estão chegando ao Brasil	Nao_edu
jAl-EaUwKnc	Aula 01 - Corrente Elétrica	Edu
kCKuP2OqfEs	HISTÓRIA GERAL #22 1ª GUERRA MUNDIAL	Edu
oTWF4p9Xc10	Mercado que vendia produtos vencidos passa por fiscalização	Nao_edu
qivkZr2Edpw	REGRA DE TRÊS SIMPLES PARA INICIANTES #regradetres #enem #concurso #encceja	Edu
rLRAZ1Mzqsg	Poríferos - Reino Animalia (Zoologia) - Prof. Paulo Jubilut	Edu
sPfwrl1ugLU	SPOOKY HOUSE: SERÁ QUE FOI TUDO VERDADE? 👻	Nao_edu
sQewkYR4_sg	Geografia - Aula 01 - Orientação e Cartografia	Edu
tHYxjJxtJko	Exercício Python #004 - Dissecando uma Variável	Edu
vmKggNWVvtU	ACENTUAÇÃO: OXÍTONAS, PAROXÍTONAS e PROPAROXÍTONAS || Aula de Português	Edu
wOAkz_o3a-I	Presidente da Sony/Playstation DEFENDE AUMENTO DE PREÇOS na nova geração kkkkkkkkkkkkkkkkkkkkkkkkkkkk	Nao_edu
wVZYyOuz9GI	Questões Comentadas: Conjuntos - Nível Básico	Edu
ycc-_JIpLGQ	CYBERPUNK 2077 - Como Está a Vida em Night City!? | Impressões de Gameplay	Nao_edu
yogFf3xVXRk	TIAGO IORC - Alexandria (Clipe Oficial)	Nao_edu

Figura 45 – Exp 5: Vídeos selecionados para a comparação das variações

5.3 Learning Object Intelligent Search - LOIS

Inicialmente, o sistema foi desenvolvido com o nome de SysVidEduc, foi desenvolvido utilizando-se a linguagem de programação Python, em conjunto com as bibliotecas *string, unidecode, NLTK, e re*, para o pré-processamento de texto, e as bibliotecas *sklearn, pandas, e joblib* para a classificação dos vídeos. Ademais, foram utilizadas as bibliotecas exigidas para a conexão com a API do *Youtube*. Posteriormente o sistema foi nomeado como LOIS (Learning Object Intelligent Search) e passou-se a utilizar o framework Tensorflow como o principal para o desenvolvimento dos códigos de Aprendizagem de Máquina.

Com a adoção do framework Tensorflow e o emprego de uma Rede Neural Simples, o LOIS passou a oferecer recomendações de vídeos educacionais com maior precisão. Baseando-se nas conclusões dos experimentos, o sistema agora adota a abordagem flexível, utilizando um grau de certeza de 40% para realizar suas recomendações. No entanto, esse grau de certeza é configurável, podendo ser ajustado tanto pelos Ambientes Virtuais de Aprendizagem (AVA) quanto pelos próprios usuários, por meio da *url* de acesso ao sistema.

Antes de colocar o sistema LOIS em operação para avaliação, foi dada especial atenção ao tempo de resposta necessário para uma busca, um aspecto crucial para um sistema que operará em tempo real e que será utilizado por docentes e aprendizes. A "Etapa 6", ilustrada na Figura 14, foi projetada para permitir a coleta de até 100 comentários por paginação. No entanto, a quantidade de requisições é limitada por cotas, e o processo envolve uma extensa quantidade de classificações para determinar se um vídeo é educacional, o que pode aumentar o tempo necessário para fornecer uma recomendação aos usuários.

Para otimizar o desempenho e eficiência do sistema, foi essencial determinar a quantidade ideal de comentários a serem coletados. Essa decisão foi baseada na análise do tempo necessário para o sistema entregar uma recomendação e na quantidade de recomendações fornecidas, dependendo do número de comentários processados. A análise conduzida é detalhada na Tabela 5.3, onde a coluna "Termo" identifica os termos de pesquisa utilizados: "Medicina" como Termo #1, "Herança" como Termo #2. A segunda coluna categoriza os resultados em "T.", referindo-se ao tempo entre a solicitação e a entrega da recomendação, e "Rec.", que indica a quantidade de vídeos recomendados. Este estudo foi fundamental para ajustar o sistema de forma a equilibrar a precisão das recomendações com a agilidade necessária para a aplicação prática em ambientes educacionais.

Tabela 27 – LOIS: Comparação do tempo e quantidade de recomendações por quantidade de comentários

Termo		**Quantidade de comentários coletados**										
		10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	**Todos**
#1	**T.**	26s	26.37s	27.06s	29s	30	32s	63s	37s	37s	38s	1197s/20m
	Rec.	9	3	5	4	3	3	2+5	3	3	4	3
#2	**T.**	21s	23.03s	25.2	25.8s	26s	30.7	30.4s	31s	33s	36s	179s
	Rec.	4	3	3	5	4	5	5	5	4	5	6

Um dos desafios significativos na coleta total de todos os comentários de um vídeo é a volumosa quantidade necessária para uma análise completa. Para ilustrar, a busca pelo termo

"Medicina" no YouTube frequentemente retorna o vídeo "Anitta - Medicina [Official Music Video]", que, no momento desta pesquisa, em outubro de 2020, possuía 105.616 comentários. Esta grande quantidade de comentários implica um tempo de processamento consideravelmente extenso para a classificação. Em comparação, se limitarmos a coleta a 100 comentários por vídeo, cada busca no YouTube resultaria na coleta e classificação de 5.000 comentários, equivalendo a capacidade de realizar 21 buscas distintas apenas com os comentários deste único vídeo da cantora Anitta.

Adicionalmente, fixar um número específico de comentários para análise padroniza o tempo de processamento para todos os termos de busca, com variações temporais mínimas de aproximadamente 1 segundo, dependendo do engajamento dos usuários em tópicos específicos.

A análise dos resultados indicou que, embora a coleta de apenas 10 comentários seja mais rápida, ela tende a produzir algumas classificações incorretas. Esta imprecisão está relacionada à insuficiente quantidade de comentários que o classificador tem para identificar como educacionais, neste caso, apenas 4, levando à recomendação de materiais inapropriados. A Figura 46 exemplifica os resultados para o termo "Herança" com a coleta de apenas 10 comentários.

Figura 46 – LOIS: Resultado para "Herança" com 10 comentários

Para equilibrar a precisão das recomendações com a rapidez necessária para uma aplicação prática, optou-se por coletar 50 comentários por vídeo, o que permite manter o tempo de resposta em aproximadamente 30 segundos. Este ajuste visa otimizar a eficiência do sistema sem comprometer a qualidade das classificações fornecidas.

A tela inicial do LOIS é apresentada na Figura 47, demonstrando a interface através da qual os usuários interagem com o sistema.

A tela inicial é simples, com um campo de busca e um botão. Além disso, o cabeçalho apresenta a logo, o nome e um botão para voltar para a tela inicial. Esse botão é também está presente na tela de recomendação. A Figura 48 apresenta os resultados da busca para o termo "Herança" com a coleta de todos os comentários.

Percebe-se que a recomendação apresenta um maior número de vídeos, 5 vídeos, que o sistema identificou como Educacionais. Essa característica do resultado não se deve exclusivamente à eficácia do sistema ou ao modelo de aprendizado de máquina empregado, uma vez que o ambiente

Figura 47 – LOIS: Página Inicial

Figura 48 – LOIS: Resultado da busca com o termo “Herança” com todos os comentários

do YouTube é extremamente dinâmico, e as recomendações podem variar a qualquer momento. Adicionalmente, independentemente da busca realizada, o sistema foi ajustado para garantir que os usuários recebam pelo menos 3 vídeos educacionais nas recomendações. Quando um vídeo desperta o interesse do usuário, há a possibilidade de assisti-lo diretamente através do sistema. A funcionalidade de visualização de vídeo integrada ao sistema é ilustrada na Figura 49, facilitando o acesso imediato ao conteúdo educacional recomendado.

O LOIS, apresenta uma interface mais limpa e simples para os seus usuários, pensando na facilidade de uso do sistema e a interação com os usuários. Adicionalmente, o sistema apresenta uma nova coluna, “Certeza” na tabela de recomendação dos vídeos. Essa coluna apresenta o Grau de Certeza do vídeo ser educacional. Além disso, o LOIS continua fornecendo as recomendações via *json*.

Figura 49 – LOIS: Execução de vídeo recomendado

Sob o ponto de vista de integração do LOIS à Ambientes Virtuais de Aprendizagem, aponta-se que, uma vez que sistema proposto possui acesso via *Web*, sem a necessidade de instalação da API, tal integração pode ser realizada por meio de retorno dos vídeos via *json*. Tal agregação permite que o LOIS forneça, de forma automática, dinâmica, e transparente, vídeos educacionais pertinentes a um assunto abordado no interior de um Ambiente Virtual de Aprendizagem. Dessa forma, *links* para vídeos complementares ao material fornecido pelo professor, podem ser apresentados ao discente, sem a necessidade deste realizar buscas por vídeos em outras plataformas, o que enriquece, sobremaneira, o processo de ensino-aprendizagem. Ademais, o LOIS é capaz de retornar, também via *json*, metadados importantes para a avaliação da qualidade de um vídeo, tais como número de visualizações, número de "*likes*" e "*dislikes*". Tais metadados podem ser utilizados pelo Ambiente Virtal de Aprendizagem para se ranquear a qualidade dos vídeos a serem recomendados na plataforma.

Para ilustrar, ao enviar um comando, como curl ou wget, o sistema LOIS responderá com informações sobre vídeos educacionais relacionados ao tema buscado. Esta interação é exemplificada na Figura 50, onde é mostrado o retorno do LOIS em formato JSON para a busca com o termo "Herança". Essa funcionalidade destaca a capacidade do sistema de fornecer recomendações acessíveis e práticas por meio de comandos simples, facilitando a integração com outras aplicações e a utilização por usuários que preferem interfaces de linha de comando.

A Tabela 5.3 apresenta os resultados obtidos por meio de um questionário aplicado aos usuários do sistema LOIS. Os resultados incluem a média e o desvio padrão (DP) das respostas. Este questionário foi direcionado a mestres e mestrandos do programa de Mestrado em Educação da Universidade Federal dos Vales do Jequitinhonha e Mucuri. A metodologia do questionário foi adaptada a partir do estudo de (PU; CHEN; HU, 2011), considerando a relevância de suas abordagens para a avaliação de sistemas de recomendação.

A seleção dos participantes se justificou pelo perfil acadêmico e profissional dos envolvidos:

Figura 50 – LOIS: Retorno do LOIS para uma requisição JSON

muitos dos respondentes são professores atuantes, o que os torna usuários diretamente interessados nas funcionalidades e eficácia do LOIS como ferramenta de apoio educacional. O questionário foi aplicado às turmas que ingressaram nos anos de 2019, 2021, 2022, 2023, abrangendo uma variedade de experiências e perspectivas sobre o uso do sistema em contextos educacionais. Notavelmente, em 2020, não houve novas admissões no programa devido às interrupções causadas pela pandemia de COVID-19.

O questionário foi divulgado para aproximadamente 120 pessoas, porém, apenas 5 responderam. Os resultados foram anonimizados, e as questões éticas devidamente consideradas. O envio foi realizado através do Google Formulários, permitindo que apenas os indivíduos interessados participassem da pesquisa.

Este levantamento de feedback visa avaliar a utilidade e usabilidade percebidas do LOIS, além de identificar oportunidades de melhoria para que o sistema possa servir ainda melhor aos educadores em suas práticas diárias.

De acordo com os resultados obtidos e os exemplos anteriormente mencionados, fica evidente que o sistema LOIS é promissor no objetivo de filtrar e recomendar vídeos de conteúdo educacional. Os dados da tabela indicam que os materiais recomendados pelo sistema estão alinhados com os interesses dos usuários; além disso, o LOIS foi avaliado como confiável e os usuários expressaram satisfação com sua funcionalidade. Essa abordagem permitirá que docentes e discentes economizem tempo na busca e seleção de vídeos educacionais, otimizando seus esforços educacionais.

Entretanto, foi identificada a necessidade de tornar o sistema mais atrativo e interativo, enriquecendo a experiência do usuário e aprimorando a personalização. Atualmente, o LOIS

Tabela 28 – LOIS: Avaliação do Sistema de Recomendação

Perguntas	**Média**	**DP**
1 - Os materiais recomendados para mim corresponderam aos meus interesses.	**4.8**	0.45
2 - O sistema de recomendação me ajudou a descobrir novos materiais.	**4.2**	0.84
3 - Os materiais recomendados para mim foram diversificados.	**4,4**	0.55
4 - O layout da interface do sistema de recomendação é atrativo.	**3.4**	1.34
5 - O sistema de recomendação explica porque os materiais foram recomendados para mim.	**3.4**	1.14
6 - As informações fornecidas para os materiais recomendados são suficientes para eu tomar uma decisão de abri-los.	**4**	1.00
7 - Eu achei fácil informar ao sistema se eu não gosto/gosto do item recomendado.	**3**	1.87
8 - Eu me tornei familiar com o sistema de recomendação muito facilmente.	**4.8**	0.45
9 - Me sinto no controle para modificar minhas preferências.	**3.6**	1.95
10 - Eu entendi porque estes conteúdos foram recomendados para mim.	**4.4**	0.89
11 - O recomendador me deu boas sugestões.	**4.4**	0.89
12 - Em geral, eu estou satisfeito com o sistema de recomendação.	**4.8**	0.45
13 - O sistema de recomendação é confiável.	**4.6**	0.54
14 - Eu usarei este recomendador outra vez.	**4.6**	0.54
15 - Qual o tempo aceitável para receber uma recomendação	**10 e 20s**	-

apresenta um excelente potencial para auxiliar estudantes e professores; no entanto, algumas melhorias ainda podem ser realizadas na plataforma, a saber:

- ❏ Aprimorar a interface do usuário, tornando-a mais interativa e atrativa.
- ❏ Integrar uma opção de classificação correta e incorreta.
- ❏ Integrar outras funcionalidades para a recomendação caso os comentários não possam ser utilizados.
- ❏ Recomendação personalizada dos melhores vídeos através dos comentários, ranqueando os vídeos através do feedback dos usuários.

Apesar das possibilidades de aprimoramento, o sistema LOIS já atende satisfatoriamente à sua finalidade inicial de auxiliar na escolha de materiais educacionais, fornecendo recomendações de Objetos de Aprendizagem com elevada acurácia. As melhorias sugeridas visam enriquecer ainda mais o sistema, aumentando sua utilidade e adaptabilidade às necessidades dos usuários. O desenvolvimento contínuo e a implementação dessas melhorias têm o potencial de transformar o LOIS em uma ferramenta ainda mais eficiente e eficaz, reforçando seu papel no suporte ao ensino e à aprendizagem.

5.4 Limitações e questionamentos

Durante o desenvolvimento do presente trabalho, surgiram dúvidas e questionamentos em suas etapas, desde a coleta dos vídeos até o comportamento do sistema proposto. Acredita-se que é importante apresentar tais questionamentos em conjunto a uma breve discussão sobre os pontos levantados. Tais considerações são apresentadas a seguir.

- **O repositório utilizado é o Youtube, um repositório de vídeos. Essa abordagem funcionaria apenas para esse repositório?**

  O objetivo do trabalho é recomendar Objetos de Aprendizagem em repositórios não estritamente educacionais, através da utilização de técnicas de Aprendizagem de Máquina. Para iniciar e verificar a viabilidade dessa abordagem, foi necessário estabelecer um ponto de partida e identificar métricas que pudessem ser utilizadas.

  Com o intuito de alcançar esses objetivos, planejou-se utilizar as opiniões, ou seja, os comentários das pessoas como métrica para a classificação de OA. Essa abordagem se mostra promissora pois os comentários estão presentes em repositórios de diversos tipos, como vídeos, músicas, filmes, blogs, posts, entre outros. Além disso, percebe-se que essa abordagem, pode ser ampliada e aplicada em diversos outros repositórios como TikTok, Linkedin, entre outros.

  Adicionalmente, acredita-se que a abordagem proposta pode ser aplicada em outros repositórios com ajustes mínimos.

- **Caso os modelos de classificação utilizem um vocabulário estritamente educacional, obterão-se melhores valores de acurácias?**

  Acredita-se que não, pois não se pode garantir que palavras pertencentes ao vocabulário educacional serão apenas utilizadas nos comentários de vídeos educacionais. Dessa forma, e analisando-se apenas o aspecto geral, caso um vídeo não educacional contenha algum vocábulo educacional, ele já poderia ser classificado, erroneamente, como educacional. Não obstante, pretende-se como trabalho futuro, criar e validar um vocabulário educacional no sentido de que o mesmo possa auxiliar no processo de classificação.

- **A qualidade do vídeo, i.e., se os usuários avaliam o mesmo positivamente ou negativamente, poderia afetar sua classificação?**

  Aqui, ressalta-se que buscaram-se por vídeos que atendessem a definição exposta anteriormente que "delimita" um conceito para vídeo educacional. Neste sentido, não se preocupou com a qualidade do vídeo, mas apenas se o mesmo se encaixaria na definição adotada. Destaca-se que a maior parte dos vídeos educacionais utilizada neste trabalho, trata de conteúdos básicos, como geografia, história, português, entre outros, havendo poucos vídeos mais específicos, de nível superior, como cálculo, programação, entre outros.

  Acredita-se que a relevância e a frequência dos comentários, assim como os vocábulos, podem ser determinantes para a classificação de um vídeo, independentemente se o mesmo for avaliado positivamente ou negativamente pelos usuários. Sob este aspecto, aponta-se

que os comentários dos vídeos poderiam ser utilizados por um Sistema de Recomendação que sugere vídeos aos usuários utilizando Análise de Sentimentos para definir a polaridade dos comentários e delimitar se um vídeo é julgado "bom" ou "ruim" pelos usuários.

- **Caso um material não possua comentários, ou seus comentários estejam desativados, ele pode ser classificado?**

  Este ponto aborda uma limitação das abordagens desenvolvidas, que é a dependência dos comentários dos usuários para a classificação de recursos educacionais em plataformas não estritamente educacionais. Embora a abordagem proposta utilize a interação dos usuários para fazer recomendações, é importante reconhecer que outras abordagens podem ser necessárias para recomendar OA quando a abordagem atual não consegue classificar esses materiais devido à falta dos comentários.

  No contexto específico dos experimentos focados em recomendar vídeos para o YouTube, a limitação se refere aos vídeos sem comentários ou com os comentários desativados, que não podem ser classificados pelas abordagens propostas. Nesse caso, a atual versão do LOIS apenas indica que esses vídeos não são classificáveis devido à ausência de comentários. Duas possíveis soluções foram consideradas para contornar essa limitação:

  1. Utilização das legendas automáticas fornecidas pelo próprio YouTube: Essa solução consiste em classificar os vídeos com base nos vocábulos presentes nas legendas geradas automaticamente. No entanto, é importante destacar que essa proposta não é uma solução definitiva, pois o recurso de geração automática de legendas não está disponível em todos os vídeos. Ademais, não sabe-se se é o intuito da plataforma implementar isso em todo o seu material.
  2. Utilização dos frames dos vídeos como forma de classificação: Essa abordagem envolve utilizar os frames (imagens estáticas) dos vídeos para realizar a classificação. No entanto, essa opção apresenta um alto custo computacional para o treinamento e armazenamento do dataset, demandando mais estudos para avaliar sua viabilidade e eficácia.

5.5 Avaliação dos Resultados

Este estudo foi desenvolvido com o objetivo de avaliar e investigar a viabilidade da recomendação de Objetos de Aprendizagem (OAs) em ambientes digitais que não se dedicam exclusivamente à educação. O objetivo principal consistiu em identificar métodos para facilitar o acesso a conteúdos educacionais em plataformas que abrangem uma ampla diversidade de informações, enfrentando os desafios que podem dificultar consideravelmente o ensino e a aprendizagem.

O desenvolvimento deste trabalho foi norteado pelas questões de pesquisa e objetivos específicos previamente apresentados. A primeira questão de pesquisa (QP1) questionou: "É possível recuperar e recomendar materiais educacionais de ambientes não estritamente educacionais, mas que são utilizados com finalidade educacional?". Por sua vez, a QP2 buscou responder: "A opinião dos usuários pode servir como critério para classificar um Objeto de Aprendizagem?".

Para abordar esses questionamentos, iniciou-se a fase de análise de dados. Essa etapa envolveu um estudo detalhado sobre o uso de palavras-chave e a frequência de certos termos em duas classes de vídeos, educacionais e não educacionais. O objetivo dessa análise era identificar palavras marcantes que pudessem ser usadas para reconhecer padrões específicos de cada categoria e, consequentemente, realizar uma classificação acurada. Observou-se que vocábulos especificamente vinculados ao contexto educacional, como "aul" e "profes", bem como expressões que denotam gratidão e aprendizado, tais como "obrig", "aprend" e "ajud", tem uma presença superior em vídeos educativos. Em contraste, percebeu-se que os comentários em vídeos não educacionais tendem a ser mais frequentes e diversificados, refletindo uma interação mais informal dos usuários, incluindo uma maior ocorrência de expressões humorísticas, como "kk" e suas extensões.

Essas descobertas permitiram não apenas responder positivamente à Questão de Pesquisa 1 (QP1), demonstrando a possibilidade de recuperar e recomendar materiais educacionais em ambientes não estritamente educacionais, mas também evidenciaram a utilidade dos comentários dos usuários como uma métrica valiosa para essa finalidade, abordando a Questão de Pesquisa 2 (QP2).

Além disso, a QP3 procurou verificar: "Algoritmos de Aprendizado de Máquina podem ser empregados, com alta precisão (superior a 90%), para identificar padrões e classificar Objetos de Aprendizagem?". Com base nas análises anteriores, deu-se início ao desenvolvimento do primeiro experimento. Nessa fase, modelou-se o dataset para conseguir representar ambas as classes de forma satisfatória, utilizando os vocábulos mais frequentes de cada classe como features. Este equilíbrio entre as classes e a seleção criteriosa das features permitiram uma distinção eficaz e evitaram que uma classe dominasse o modelo com características exclusivamente suas.

Além da utilização dos vocábulos mais frequentes, modelou-se o dataset para que algoritmos de Aprendizado de Máquina classificassem diretamente um vídeo como educacional ou não, numa abordagem que chamamos de "rígida". Durante o experimento, os algoritmos J48, Random Forest, PART, JRip e GenClust++ foram testados, e apenas o Random Forest alcançou acurácias acima de 90%, atingindo o valor máximo de 91,30%. Adicionalmente, no experimento 2, ao aplicar outros modelos de Aprendizado de Máquina como Redes Neurais, obteve-se uma acurácia de 95,34% ao utilizar uma CNN. Os demais algoritmos também atingiram acurácias superiores a 90%, com o Random Forest alcançando 92,13%, a Rede Neural Simples com 93,86% e a Rede Neural Profunda com 94,46%.

Além disso, o Experimento 5 demonstrou a possibilidade de alcançar uma precisão de até 95,70% utilizando uma Rede Neural Convolucional para classificar vídeos educacionais. Esses resultados confirmam a viabilidade do uso de algoritmos de aprendizado de máquina para a classificação eficaz de OAs, como investigado na Questão de Pesquisa 3 (QP3).

Também foi proposta a metodologia "flexível" para a Recomendação de Objetos de Aprendizagem, na qual classifica-se os comentários individualmente e aplica-se a métrica de "grau de certeza" para avaliar o nível de confiança na classificação do vídeo como educacional. O Experimento 5 responde à Questão de Pesquisa 4 (QP4): "É possível desenvolver uma metodologia ajustável para a recomendação de OAs que possa ser adaptada de acordo com as necessidades específicas de cada temática?". Os resultados demonstraram a viabilidade de desenvolver e aplicar

uma metodologia flexível e precisa. Contudo, é crucial considerar as métricas de recomendação e os casos de classificação cuidadosamente, pois um Grau de Certeza mais alto pode reduzir o número de OAs recomendados. Se bem planejada e estruturada, esta abordagem pode ser eficazmente aplicada em casos específicos.

A realização deste trabalho também foi orientada pelos objetivos específicos. Por meio da execução dos experimentos, exploraram-se os comentários dos usuários, obtendo-se percepções valiosas sobre a interação dos usuários com vídeos educacionais e não educacionais. Adicionalmente, diversos algoritmos de Aprendizado de Máquina foram empregados e testados, como J48, Random Forest, PART, JRIP, GenClust++, Rede Neural Simples, Rede Neural Profunda Densa e Rede Neural Convolucional. Notou-se que as Redes Neurais apresentaram resultados superiores em comparação a outros algoritmos, evidenciando mais uma vez o potencial das redes neurais.

Por fim, desenvolveu-se um modelo de Aprendizado de Máquina que utiliza Redes Neurais e os 500 vocábulos mais frequentes para classificar novos OAs. Esse modelo é empregado junto ao LOIS, um Sistema de Recomendação para vídeos educacionais do YouTube, que auxilia aprendizes e docentes na etapa de seleção de materiais relevantes.

Os resultados obtidos são particularmente expressivos considerando que abordam uma dificuldade anteriormente inexplorada e devido a natureza real dos dados, que são coletados diretamente de interações humanas. Esses dados incluem gírias e as variadas nuances linguísticas típicas do português. Desenvolver métricas que recomendem conteúdo com mais de 90% de acurácia já seria, por si só, um avanço notável. Nesse contexto, as acurácias alcançadas de 95,34% na variação rígida e de 95,70% na variação flexível são particularmente impressionantes.

Os resultados obtidos confirmaram positivamente a hipótese levantada: “A análise de comentários aprimora o processo de recomendação de conteúdos educacionais”. Foi possível observar que os comentários são uma fonte rica de informações e, quando adequadamente utilizados, podem melhorar o processo de recomendação e seleção de materiais educacionais.

Em resumo, este trabalho não apenas enfrentou os desafios da recomendação de OAs em plataformas de conteúdo diversificado, como também utilizou de forma inovadora os comentários dos usuários como dados essenciais para a classificação e recomendação de conteúdo educacional. As metodologias empregadas forneceram informações valiosas e demonstraram o potencial das técnicas de Mineração de Texto e Aprendizado de Máquina para melhorar o acesso a materiais educacionais. Além disso, a implementação dessas descobertas no sistema LOIS constitui uma contribuição expressiva para facilitar a busca por recursos educacionais em ambientes digitais diversificados, beneficiando tanto educadores quanto alunos na descoberta de conteúdo relevante e de qualidade.

Capítulo 6

Conclusão

O compartilhamento massivo de informações tem impulsionado progressos significativos em várias áreas da sociedade, especialmente na educação. No cenário educacional, a ampla disponibilidade de conteúdos oferece acesso privilegiado a uma diversidade de materiais, mas também representa um desafio para educadores e estudantes na fase de identificar e selecionar recursos pertinentes. A vastidão de materiais disponíveis pode tornar este processo exaustivo e pouco estimulante, especialmente em plataformas que não são especificamente destinadas a serem repositórios educacionais.

Diante deste desafio, este estudo introduz uma abordagem inovadora que emprega a filtragem de comentários para a recomendação de Objetos de Aprendizagem (OA) em plataformas de conteúdo diversificado. Optou-se pelo YouTube como plataforma experimental por ser um repositório de vídeos, que é a principal mídia das as novas gerações, e sua extensa utilização como ferramenta de aprendizado. É importante destacar que, embora o YouTube tenha sido a escolha inicial para os experimentos, a metodologia desenvolvida e suas variantes são flexíveis e podem ser adaptadas para qualquer plataforma que faça uso das opiniões dos usuários.

A abordagem proposta utiliza a opinião dos usuários conjuntamente com os termos mais frequentes em vídeos educacionais e não educacionais para recomendar Objetos de Aprendizagem (OAs). Utilizando os termos mais frequentes em cada classe, desenvolveu-se um vocabulário específico para ser utilizado como características para distinguir entre materiais educacionais e não educacionais. O vocabulário foi compilado de forma a manter a equidade entre as palavras mais frequentes encontradas tanto em vídeos educacionais quanto em vídeos não educacionais.

Foram implementadas duas metodologias distintas para a modelagem e classificação de vídeos educacionais: a "rígida" e a "flexível". A metodologia rígida é projetada para classificar os materiais diretamente como educacionais ou não educacionais. Nesta abordagem, um vídeo e seus comentários, são modelados para serem representados unicamente em uma entrada, ou seja, todos os comentários são unificados em uma única entrada representativa (vetor). Cada valor desse vetor corresponde à frequência com que um determinada característica - vocábulo mais frequente - aparece em todos os comentários. Em seguida, um algoritmo de Aprendizado de Máquina avalia esse vetor e fornece uma classificação categórica, indicando se o vídeo é educacional ou não educacional.

Como resultado obtido, nessa variação, conseguimos alcançar os objetivos propostos no início

da pesquisa. Conseguiu-se obter excelente acurácia média de 95,53% ao se utilizar a Rede Neural Convolucional e o dataset composto por 500 vocábulos para classificar OA através dos comentários.

Com a progressão do estudo e o reconhecimento de limitações técnicas, foi desenvolvida a metodologia flexível. Esta abordagem trata a classificação de cada comentário de forma individualizada e, subsequente a isso, emprega a fórmula do "grau de certeza" para determinar o nível de confiança de que o material é educacional. Nesta variação, cada comentário é tratado e modelado como uma entrada representativa (vetor), e cada posição desse vetor corresponde à frequência com que um determinada característica - vocábulo mais frequente - aparece apenas nesse comentário.

Ao se utilizar a metodologia flexível, em conjunto com a Fórmula do grau de certeza foi possível obter acurácia média de 95,70% ao classificar um novo vídeo educacional com a utilização de uma CNN. A metodologia flexível, conseguiu obter resultados melhores que na metodologia rígida, que atingiu a melhor média de 95,53%. Além disso, a metodologia flexível, quando combinada com uma rede neural simples, conseguiu superar a acurácia da CNN, oferecendo também um menor custo computacional.

Adicionalmente, ao analisar vídeos com diferentes quantidades de comentários, a rede neural baseada na metodologia flexível apresentou desempenho superior, atingindo uma média de 94,55%, enquanto a abordagem rígida alcançou 91,64%. Esses resultados destacam a maior capacidade de adaptação da metodologia flexível a vídeos com variados níveis de engajamento, tornando-a mais adequada para uso em ambientes reais onde os dados são derivados de interações autênticas.

Essa variação flexível oferece maior adaptabilidade e flexibilidade, ajustando-se eficazmente às exigências práticas do uso, o que permite uma classificação mais precisa de vídeos com diferentes volumes de comentários. Além disso, esta metodologia tem a vantagem significativa de manter a acurácia consistente, independentemente de se usar todos os comentários disponíveis ou apenas uma fração, reduzindo assim o número de requisições necessárias.

Ambas as variações da metodologia alcançaram resultados excelentes, com eficácia superior a 95%. No entanto, a metodologia flexível mostrou-se mais alinhada aos objetivos propostos. Sua capacidade de fazer boas recomendações com diferentes quantidades de comentários é crucial, permitindo uma maior adequação às condições do mundo real, onde os materiais podem apresentar variados níveis de engajamento. Por essa razão, a metodologia flexível destacou-se por sua adaptabilidade e eficácia.

Por fim, foi desenvolvido o LOIS, um Sistema de Recomendação de Objetos de Aprendizagem com o propósito de auxiliar a etapa de seleção de materiais educacionais. O LOIS fornece aos usuários vídeos educacionais classificados de acordo com a metodologia flexível. Além disso, possibilita a integração com ambientes virtuais de aprendizagem, permitindo que os resultados sejam retornados em formato JSON, tornando a integração com outras plataformas mais simples e eficiente. Com essa funcionalidade, aprendizes e professores podem acessar facilmente os recursos educacionais, otimizando a etapa de busca e escolha de materiais para seus estudos e atividades pedagógicas.

Ao avaliar os resultados deste estudo, torna-se claro que a abordagem inovadora e o desen-

volvimento do LOIS como um sistema pioneiro na recomendação de Objetos de Aprendizagem representam um marco significativo na interseção entre tecnologia e educação. Este projeto não só confirma a viabilidade do uso do feedback dos usuários como métrica eficientes para a recomendação e seleção de materiais educacionais mas também demonstra como a inteligência artificial e a análise de comentários podem ser empregados de maneira efetiva para personalizar a educação.

Através do LOIS, foi possível demonstrar como a análise inteligente de comentários e a categorização de conteúdo podem ser utilizadas para fornecer recursos altamente relevantes e adequados às necessidades dos aprendizes. Esta inovação não apenas aprimora a eficiência e a eficácia na localização de materiais adequados, mas também possibilita uma experiência de aprendizado mais engajadora e motivadora.

6.1 Principais Contribuições

O trabalho apresenta uma abordagem e duas variações para a classificação de Objetos de Aprendizagem em repositórios não estritamente educacionais através do uso das opiniões dos usuários e as principais contribuições são apresentadas abaixo:

- Identificação do comportamento dos usuários: Ao analisar os comentários em materiais educacionais e não educacionais, o estudo revelou diferenças significativas nas interações dos usuários. Além disso, a identificação de um vocabulário específico para os materiais educacionais destaca a relevância dos comentários como indicadores de conteúdos educacionais.

- Metodologias para a classificação: O trabalho apresentou duas metodologias para a modelagem e classificação de Objetos de Aprendizagem em repositórios não estritamente educacionais. Ambas utilizam os vocábulos mais frequentes em cada classe, entretanto, a metodologia rígida utiliza técnicas de aprendizado de máquina para classificar o vídeo como educacional ou não educacional. Já a metodologia flexível classifica cada comentário individualmente, fornecendo um grau de certeza para o vídeo ser educacional.

- Potencial dos comentários na classificação: O estudo evidenciou o elevado potencial dos comentários como fonte de informações para determinar se um vídeo é educacional ou não. A utilização de técnicas de aprendizado de máquina alcançou elevada acurácia na classificação de Objetos de Aprendizagem, destacando a eficácia dessa abordagem para recomendações educacionais em plataformas não estritamente educacionais.

- Avanço no estado da arte: O trabalho representa um avanço significativo tanto na área de Educação quanto na de Aprendizado de Máquina ao abordar a classificação de Objetos de Aprendizagem em repositórios não estritamente educacionais. Por meio do desenvolvimento das metodologias e da apresentação de resultados, o estudo contribui para a evolução e progresso das áreas envolvidas.

  No campo da Educação, o trabalho amplia as possibilidades de seleção e recomendação de materiais educacionais em ambientes não estritamente educacionais. Através da análise

dos comentários dos usuários, identifica-se um vocabulário específico e relevante para reconhecer os padrões nos conteúdos educacionais dos não educacionais, contribuindo para tornar otimizar a seleção de recursos e torná-la precisa.

Por outro lado, na área da Aprendizagem de Máquina, o estudo demonstra o poder e a eficácia dessas técnicas na classificação de Objetos de Aprendizagem com base nas opiniões dos usuários. A utilização de algoritmos de aprendizado de máquina, como Random Forest, Part, JRIP, J48, Rede Neural Simples, Rede Neural Profunda Densa e Rede Neural Convolucional, revelou-se altamente precisa na tarefa de classificação e, assim, abre novas possibilidades de aplicação dessas técnicas em contextos educacionais.

- Criação de datasets, modelos e SR: Como resultado do trabalho, foram desenvolvidos dois datasets extensos, que representam recursos valiosos para futuras pesquisas e avaliações em Aprendizagem de Máquina e classificação de Objetos de Aprendizagem. Além disso, desenvolveu-se 2 modelos de aprendizagem de máquina, que podem ser utilizados para classificar novos OA. Por fim, o desenvolvimento de um sistema com o propósito de auxiliar professores, aprendizes e Ambientes Virtuais de Aprendizagem na recomendação de materiais educacionais o que representa um avanço prático e aplicável na área da educação.

- Potencial em outras áreas: A abordagem e as variações aplicadas neste projeto possuem potencial aplicável além da esfera educacional, estendendo-se a outras áreas, incluindo o âmbito das redes sociais. Um exemplo notável dessa versatilidade é o trabalho de Carvalho (2024), que emprega a seleção de vocábulos mais frequentes em combinação com a ciência de redes para identificar notícias falsas. Tal aplicação evidencia a adaptabilidade e o amplo alcance das abordagens desenvolvidas, sugerindo sua viabilidade e eficácia em contextos variados, especialmente em desafios contemporâneos como o combate à disseminação de desinformação.

Esse trabalho é uma contribuição relevante para a área de Educação e de Inteligência Artificial/Aprendizagem de Máquina através da recomendação de Objetos de Aprendizagem em ambientes não estritamente educacionais, apresentando duas excelentes abordagens, boa análise e desenvolvendo recursos capazes de auxiliar professores e aprendizes na etapa de seleção de materiais relevantes. As descobertas e apontamentos apresentados por esse trabalho apresentam potencial para auxiliar e melhorar o aprendizado.

6.2 Trabalhos Futuros

O trabalho demonstrou excelente acurácia na classificação de Objetos de Aprendizagem, porém, percebe-se que estudos ainda podem ser realizados quanto a métricas para a Recomendação de Objetos de Aprendizagem, onde pode-se utilizar, por exemplo, análise de sentimentos para identificar os melhores materiais e os que apresentam melhor qualidade.

Devido a essa observação, pretende-se explorar a análise de sentimentos para auxiliar na Recomendação de Objetos de Aprendizagem, assim como, aprimorar o sistema desenvolvido para

Recomendar os melhores Objetos de Aprendizagem através das opiniões dos usuários utilizando análise de sentimentos.

Apesar das abordagens apresentadas atingirem os objetivos propostos, elas apresentam uma limitação quanto a dependência dos comentários para realizar a classificação, demonstrando que caso os comentários não estejam presentes ou serem impossíveis de serem obtidos, outras abordagens devem ser utilizadas para ultrapassar a falta de classificação que as abordagens propostas. Devido a isso, e utilizando o Youtube como repositório experimental, podemos apresentar duas abordagens, a saber:

- ❏ Utilização das legendas automáticas: Nessa abordagem, deve-se analisar a possibilidade de utilizar as legendas automáticas dos vídeos e verificar se as abordagens propostas podem ser utilizadas. Caso as abordagens propostas não possam ser utilizadas devemos desenvolver um novo dataset para realizar a classificação dos objetos de aprendizagem.

- ❏ Utilização dos frames para a classificação: Essa abordagem consiste em obter os frames dos vídeos para realizar a classificação. Essa abordagem também apresenta elevada complexidade devido a necessidade de obter frames de um vídeo e classificá-lo.

Além desses apontamentos, existe o interesse em aplicar as metodologias propostas em outras plataformas, verificando se o comportamento continua o mesmo, assim como, verificar se o modelo desenvolvido funcionaria diretamente em outras plataformas.

Adicionalmente, existe o interesse em verificar se o comportamento em outras línguas, principalmente Inglês e Espanhol, são os mesmos e quais os ajustes seriam necessários.

6.3 Contribuições em Produção Bibliográfica

As pesquisas com relação ao trabalho proposto teve início em agosto de 2019, com o início do mestrado em Educação e teve sua continuidade no Doutorado. As publicações (CARVALHO et al., 2020c), (CARVALHO et al., 2020a) e (CARVALHO et al., 2020b) foram desenvolvidas no Mestrado e serviram de base para a continuidade da pesquisa. De 2021, entrada no Doutorado, até a defesa foram produzidos os seguintes trabalhos ligados ao tema:

- ❏ (CARVALHO et al., 2022) - **CARVALHO, H. C. F. B.**; DORÇA, F. A.; PITANGUI, C. G.; ASSIS, L. P.; ANDRADE, A. V.; TRINDADE, E. A. C.; Classificação automática de vídeos educacionais por meio de comentários apoiada por técnicas de aprendizado de máquina: uma análise experimental utilizando o youtube. Revista Brasileira de Informática na Educação, v. 30, p. 419–448, set. 2022. Disponível em: <https://sol.sbc.org.br/journals/index.php/rbie/article/view/2455>.

- ❏ (CARVALHO et al., 2024) - **CARVALHO, H. C. F. B.**; DORÇA, F. A.; PITANGUI, C. G.; ASSIS, L. P.; ANDRADE, A. V.; TRINDADE, E. A. C.; Improving the educational experience on Youtube: a machine learning approach to classifying and recommending educational videos. Revista Gestão e Secretariado, v. 15, n. 4, p. e3587, abr. 2024 -

Disponível em:
<https://ojs.revistagesec.org.br/secretariado/article/view/3587>

- (CARVALHO et al., 2023) - **CARVALHO, H. C. F. B.**; PITANGUI, C. G.; DORÇA, F. A.; OLIVEIRA, C. S.; ASSIS, L. P.; ANDRADE, A. V.; TRINDADE, E. A. C.; Probabilistic classification of educational videos considering comments: an experimental analysis on Youtube. In: Anais do XXXIV Simpósio Brasileiro de Informática na Educação. Porto Alegre, RS, Brasil: SBC, 2023. p. 1408–1418. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/26765>.

- (OLIVEIRA et al., 2024) - OLIVEIRA, C. S.; PITANGUI, C. G.; **CARVALHO, H. C. F. B.**; ANDRADE, A. V.; ASSIS, L. P.; DORÇA, F. A.; Classificação de vídeos educacionais do youtube por níveis de ensino utilizando comentários: uma abordagem experimental. Revista Novas Tecnologias na Educação, v. 22, n. 1, p. 589–599, 2024.

Além das publicações, foi realizado o Registro de Programa de Computador no INPI.

- **CARVALHO, H. C. F. B.**; DORCA, F. A. ; PITANGUI, C. G. ; TRINDADE, E. A. C. ; ANDRADE, A. V. ; ASSIS, L. P. . LOIS (Learning Object Intelligent Search). 2024. Patente: Programa de Computador. Número do registro: 512024002083-7, data de registro: 01/02/2024, título: "LOIS (Learning Object Intelligent Search)" , Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial.

Além desses trabalhos, as abordagens propostas auxiliaram na seguinte dissertação de mestrado:

- CARVALHO, J. A.; PEREIRA, F. S. F; TRAVENÇOLO, B. A. N; Detecção e identificação de notícias falsas em redes sociais utilizando abordagem de ciência de redes. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de Uberlândia, Uberlândia, Jan 2024. Available at Universidade Federal de Uberlândia Repository:
<https://repositorio.ufu.br/handle/123456789/41258>.

Referências

ABU-EL-HAIJA, S. et al. Youtube-8m: A large-scale video classification benchmark. **arXiv preprint arXiv:1609.08675**, 2016.

AFONSO, A. R.; DUQUE, C. G. Análise de sentimentos em comentários de vídeos do youtube utilizando aprendizagem de máquinas supervisionada. **Ciência da Informação**, v. 48, n. 3, 2019.

ALLAHYARI, M. et al. A brief survey of text mining: Classification, clustering and extraction techniques. **arXiv preprint arXiv:1707.02919**, 2017.

AMANDA, R.; NEGARA, E. S. Analysis and implementation machine learning for youtube data classification by comparing the performance of classification algorithms. **Jurnal Online Informatika**, v. 5, n. 1, p. 61–72, 2020.

ASSEMBLYAI. **Neural Networks explained in 60 seconds!** 2022. Disponível em: <https://www.youtube.com/watch?v=kQl45ophSVQ#t=0m43s>.

BARRÉRE, E. et al. Utilização de enriquecimento semântico para a recomendação automática de videoaulas no moodle. **Revista Brasileira de Informática na Educação**, v. 28, n. 1, 2020. Disponível em: <https://doi.org/10.5753/rbie.2020.28.0.319>.

BERRAR, D. Cross-validation. In: RANGANATHAN, S. et al. (Ed.). **Encyclopedia of Bioinformatics and Computational Biology**. Oxford, UK: Academic Press, 2019. p. 542–545. ISBN 978-0-12-811432-2. Disponível em: <https://10.1016/B978-0-12-809633-8.20349-X>.

BERRY, M. J.; LINOFF, G. S. **Data mining techniques: for marketing, sales, and customer relationship management**. [S.l.]: John Wiley & Sons, 2004.

BIANCHINI, Â. R. Arquitetura de redes neurais para o reconhecimento facial baseado no neocognitron. Universidade Federal de São Carlos, 2001.

BRAGA, J.; MENEZES, L. **Objetos de aprendizagem, volume 1: introdução e fundamentos**. UFABC, 2014. v. 1. 153 p. Disponível em: <https://pesquisa.ufabc.edu.br/intera/wp-content/uploads/2015/12/objetos-de-aprendizagem-v1.pdf>.

BRAME, C. J. Effective educational videos: Principles and guidelines for maximizing student learning from video content. **CBE—Life Sciences Education**, Am Soc Cell Biol, v. 15, n. 4, p. es6, 2016. Disponível em: <https://doi.org/10.1187/cbe.16-03-0125>.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, n. 2, p. 123–140, 1996. Disponível em: <https://doi.org/10.1007/BF00058655>.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>.

BREIMAN, L. et al. **Classification and Regression Trees**. Routledge, 1984. Disponível em: <https://doi.org/10.1201/9781315139470>.

BURGESS, J.; GREEN, J. Youtube e a revolução digital. **São Paulo: Aleph**, 2009. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/2205278/mod_resource/content/1/Burgess\%20et\%20al.\%20-\%202009\%20-\%20YouTube\%20e\%20a\%20Revolu\%C3\%A7\%C3\%A3o\%20Digital\%20Como\%20o\%20maior\%20fen\%C3\%B4meno\%20da\%20cultura\%20participativa\%20transformou\%20a\%20m\%C3\%ADdia\%20e\%20a\%20socieda.pdf>.

CARVALHO, H. et al. Probabilistic classification of educational videos considering comments: an experimental analysis on youtube. In: **Anais do XXXIV Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2023. p. 1408–1418. ISSN 0000-0000. Disponível em: <https://doi.org/10.5753/sbie.2023.235155>.

CARVALHO, H. C. F. B. et al. Classificação automática de vídeos educacionais por meio de comentários apoiada por técnicas de aprendizado de máquina: uma análise experimental utilizando o youtube. **Revista Brasileira de Informática na Educação**, v. 30, p. 419–448, set. 2022. Disponível em: <https://doi.org/10.5753/rbie.2022.2455>.

CARVALHO, H. C. F. B. et al. Improving the educational experience on youtube: a machine learning approach to classifying and recommending educational videos. **Revista de Gestão e Secretariado**, v. 15, n. 4, p. e3587, abr. 2024. Disponível em: <https://doi.org/10.7769/gesec.v15i4.3587>.

CARVALHO, H. C. F. B. et al. Educavídeos: Um sistema de recomendação de objetos de aprendizagem de vídeos educacionais do youtube. In: **ESUD 2020 - XVII Congresso Brasileiro de Ensino Superior a Distância**. [s.n.], 2020. Disponível em: <https://esud2020.ciar.ufg.br/wp-content/anais-esud/210418.pdf>.

CARVALHO, H. C. F. B. et al. Categorização de vídeos educacionais do youtube por meio de comentários. **RENOTE**, v. 18, n. 2, p. 621–629, 2020. Disponível em: <https://doi.org/10.22456/1679-1916.110305>.

CARVALHO, H. C. F. B. et al. Learning objects and youtube: an analysis of videos and their categories. In: **LACLO 2020 - XV Latin American Conference on Learning Technologies**. [S.l.: s.n.], 2020.

CARVALHO, J. A. d. **Detecção e identificação de notícias falsas em redes sociais utilizando abordagem de ciência de redes**. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de Uberlândia, Uberlândia, Jan 2024.

CHAVES, A. d. C. F. **Extração de Regras Fuzzy para Máquinas de Vetores Suporte (SVM) para Classificação em Múltiplas Classes**. Tese (Doutorado) — PUC-Rio, 2006.

COHEN, W. W. Fast effective rule induction. In: PRIEDITIS, A.; RUSSELL, S. (Ed.). **Machine Learning Proceedings 1995**. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 115–123. ISBN 978-1-55860-377-6. Disponível em: <https://doi.org/10.1016/B978-1-55860-377-6.50023-2>.

COPPIN, B. **Inteligência artificial**. [S.l.]: Grupo Gen-LTC, 2015.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995. Disponível em: <https://doi.org/10.1007/BF00994018>.

COX, C. **Listen to this**. [S.l.]: C. Cox Publishing Company, 1922.

DANG, S.; AHMAD, P. H. Text mining: Techniques and its application. **International Journal of Engineering & Technology Innovations**, v. 1, n. 4, p. 22–25, 2014.

FRANK, E.; HALL, M. A.; WITTEN, I. H. **The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"**. Morgan Kaufmann Publishers, 2016. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf>.

FRANK, E.; WITTEN, I. H. Generating accurate rule sets without global optimization. In: SHAVLIK, J. (Ed.). **Fifteenth International Conference on Machine Learning**. [S.l.]: Morgan Kaufmann, 1998. p. 144–151.

FRIEDMAN, R. Tokenization in the theory of knowledge. **Encyclopedia**, MDPI, v. 3, n. 1, p. 380–386, 2023. Disponível em: <https://doi.org/10.3390/encyclopedia3010024>.

GAIKWAD, S. V.; CHAUGULE, A.; PATIL, P. Text mining methods and techniques. **International Journal of Computer Applications**, Foundation of Computer Science, v. 85, n. 17, 2014. Disponível em: <https://doi.org/10.5120/14937-3507>.

GOMES, L. Vídeos didáticos: uma proposta de critérios para análise. **Revista Brasileira de Estudos Pedagógicos**, v. 89, n. 223, 2008. Disponível em: <https://doi.org/10.24109/2176-6681.rbep.89i223.688>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>.

HACOHEN-KERNER, Y.; MILLER, D.; YIGAL, Y. The influence of preprocessing on text classification using a bag-of-words representation. **PloS one**, Public Library of Science, v. 15, n. 5, p. 1–22, 2020. Disponível em: <https://doi.org/10.1371/journal.pone.0232525>.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2001.

HEILBRON, F. D. C. **Understanding Human Activities at Large Scale**. Tese (Doutorado), 2019.

HICKMAN, L. et al. Text preprocessing for text mining in organizational research: Review and recommendations. **Organizational Research Methods**, SAGE Publications Sage CA: Los Angeles, CA, v. 25, n. 1, p. 114–146, 2022. Disponível em: <https://doi.org/10.1177/1094428120971683>.

HOBBS, J. R.; RILOFF, E. Information extraction. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of natural language processing**. Boca Raton, FL, USA: Chapman and Hall/CRC, 2010. p. 511–532.

HORTA, E. G. Previsores para a eficiência da quimioterapia neoadjuvante no câncer de mama. **M. Sc., Universidade Federal de Minas Gerais (UFMG)**, 2008.

HORTA, E. G. et al. Extração de características e casamento de padrões aplicados à estimação de posição de um VANT. **UFMG**, 2011.

IEEE. Ieee standard for learning object metadata. ieee standard 1484.12.1. In: . New York, NY, USA: Institute of Electrical and Electronics Engineers, 2002. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1032843>.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: **Proceedings of the 32nd International Conference on Machine Learning (ICML)**. [s.n.], 2015. p. 448–456. Disponível em: <https://doi.org/10.48550/arXiv.1502.03167>.

ISLAM, M. Z. et al. Combining k-means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering. **Expert Systems with Applications**, Elsevier, v. 91, p. 402–417, 2018. Disponível em: <https://doi.org/10.1016/j.eswa.2017.09.005>.

JÚNIOR, C. B.; DORÇA, F. Uma abordagem para a criação e recomendação de objetos de aprendizagem usando um algoritmo genético, tecnologias da web semântica e uma ontologia. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [s.n.], 2018. p. 1533–1542. Disponível em: <https://doi.org/10.5753/cbie.sbie.2018.1533>.

JUSOH, S.; ALFAWAREH, H. M. Techniques, applications and challenging issue in text mining. **International Journal of Computer Science Issues (IJCSI)**, International Journal of Computer Science Issues (IJCSI), v. 9, n. 6, p. 431, 2012.

KANNAN, S. et al. Preprocessing techniques for text mining. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2014.

KESAVARAJ, G.; SUKUMARAN, S. A study on classification techniques in data mining. In: IEEE. **2013 fourth international conference on computing, communications and networking technologies (ICCCNT)**. 2013. p. 1–7. Disponível em: <https://doi.org/10.1109/ICCCNT.2013.6726842>.

KOVÁCS, Z. L. **Redes neurais artificiais**. [S.l.]: Editora Livraria da Fisica, 2002.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, p. 1097–1105, 2012. Disponível em: <https://doi.org/10.1145/3065386>.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Springer, v. 521, n. 7553, p. 436–444, 2015. Disponível em: <https://doi.org/10.1038/nature14539>.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, IEEE, v. 86, n. 11, p. 2278–2324, 1998. Disponível em: <https://doi.org/10.1109/5.726791>.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007. Disponível em: <https://doi.org/10.22456/2175-2745.5690>.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, p. 115–133, 1943. Disponível em: <https://doi.org/10.1007/BF02478259>.

MENOLLI, A.; MALUCELLI, A.; REINEHR, S. Criaçao semi-automatica de objetos de aprendizagem a partir de conteúdos da wiki. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2011.

MIRANDA, R. M. d. **GROA: um gerenciador de repositórios de objetos de aprendizagem**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, BR, 7 2004.

MITCHELL, T. M. **Machine learning**. New York, NY, USA: McGraw-hill New York, 1997.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

NASCIMENTO, P. do et al. Recomendaçao de objetos de aprendizagem baseada em modelos de estilos de aprendizagem: Uma revisao sistemática da literatura. v. 28, n. 1, p. 213, 2017. Disponível em: <https://10.5753/cbie.sbie.2017.213>.

NASCIMENTO, R. F. F. et al. O algoritmo support vector machines (svm): avaliação da separação ótima de classes em imagens ccd-cbers-2. **Simpósio Brasileiro de Sensoriamento Remoto**, v. 14, p. 2079–2086, 2009.

OLIVEIRA, C. dos S. et al. Classificação de vídeos educacionais do youtube por níveis de ensino utilizando comentários: uma abordagem experimental. **Revista Novas Tecnologias na Educação**, v. 22, n. 1, p. 589–599, 2024.

PINHEIRO, R. R. A. et al. **Sistema de recomendação de vídeos educacionais: um estudo de caso no Youtube**. Dissertação (Mestrado) — Universidade Federal de Alagoas, Maceió, AL, BR, 2018.

PIRHONEN, J.; RASI, P. Student-generated instructional videos facilitate learning through positive emotions. **Journal of Biological Education**, Taylor & Francis, v. 51, n. 3, p. 215–227, 2017. Disponível em: <https://doi.org/10.1080/00219266.2016.1200647>.

PU, P.; CHEN, L.; HU, R. A user-centric evaluation framework for recommender systems. In: **Proceedings of the fifth ACM conference on Recommender systems**. [s.n.], 2011. p. 157–164. Disponível em: <https://doi.org/10.1145/2043932.2043962>.

QUINLAN, R. **C4.5: Programs for Machine Learning**. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1993.

RAMOS, J. L. C. et al. A comparative study between clustering methods in educational data mining. **IEEE Latin America Transactions**, IEEE, v. 14, n. 8, p. 3755–3761, 2016. Disponível em: <https://doi.org/10.1109/TLA.2016.7786360>.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, American Psychological Association, v. 65, n. 6, p. 386–408, 1958. Disponível em: <https://doi.org/10.1037/h0042519>.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. **Learning representations by back-propagating errors**. Nature Publishing Group UK London, 1986. v. 323. 533–536 p. Disponível em: <https://doi.org/10.1038/323533a0>.

RUSSELL, S.; NORVIG, P. **Artificial intelligence: a modern approach**. Upper Saddle River, NJ, USA: Pearson Education, 2010.

SCIKIT-LEARN DEVELOPERS. **1.10. Decision Trees — scikit-learn 1.5.2 documentation**. [S.l.], 2023. Disponível em: <https://scikit-learn.org/stable/modules/tree.html>.

SEMOLINI, R. **Support vector machines, inferência transdutiva e o problema de classificação**. Tese (Doutorado) — Universidade Estadual de Campinas, 2002.

SHEPHERD, G. M.; KOCH, C. Introduction to synaptic cir cuits. **The Synaptic Organization fo the Brain.**, Oxford Unviersity, p. 3–31, 1990.

SMOLA, A. J.; SCHÖLKOPF, B. **Learning with kernels**. [S.l.]: Citeseer, 1998.

SRIVASTAVA, N. et al. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, v. 15, n. 1, p. 1929–1958, 2014.

SUKANYA, M.; BIRUNTHA, S. Techniques on text mining. In: IEEE. **2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)**. 2012. p. 269–271. Disponível em: <https://doi.org/10.1109/ICACCCT.2012.6320784>.

SUMATHI, S.; SIVANANDAM, S. **Introduction to data mining and its applications**. Springer, 2006. v. 29. Disponível em: <https://doi.org/10.1007/978-3-540-34351-6>.

THELWALL, M. Social media analytics for youtube comments: Potential and limitations. **International Journal of Social Research Methodology**, Routledge, v. 21, n. 3, p. 303–316, 2018. Disponível em: <https://doi.org/10.1080/13645579.2017.1381821>.

TRINDADE, E. A. C. et al. Modelagem do problema de cobertura de conjunto para recomendação de objetos de aprendizagem aplicado ao repositório do youtube. **RENOTE**, v. 18, n. 2, p. 358–367, 2020. Disponível em: <https://doi.org/10.22456/1679-1916.110254>.

VAPNIK, V. N.; VAPNIK, V. **Statistical learning theory**. [S.l.]: Wiley New York, 1998. v. 1.

VIEIRA, F. J. R.; NUNES, M. A. S. N. Dica: Sistema de recomendação de objetos de aprendizagem baseado em conteúdo. **Scientia Plena**, v. 8, n. 5, 2012.

VIJAYARANI, S. et al. Preprocessing techniques for text mining-an overview. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2015.

VIJAYARANI, S.; JANANI, R. et al. Text mining: open source tokenization tools-an analysis. **Advanced Computational Intelligence: An International Journal (ACII)**, v. 3, n. 1, p. 37–47, 2016. Disponível em: <https://doi.org/10.5121/acii.2016.3104>.

VM, N.; R, D. A. K. Implementation on text classification using bag of words model. In: **Proceedings of the second international conference on emerging trends in science & technologies for engineering systems (ICETSE-2019)**. [S.l.: s.n.], 2019.

WIEDERHOLD, G.; MCCARTHY, J. Arthur samuel: Pioneer in machine learning. **IBM Journal of Research and Development**, IBM, v. 36, n. 3, p. 329–331, 1992. Disponível em: <https://doi.org/10.1147/rd.363.0329>.

WILEY, D. A. **Learning object design and sequencing theory**. Tese (Doutorado) — Brigham Young University, 2000.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2011. v. 3. 629 p. Disponível em: <https://doi.org/10.1016/C2009-0-19715-5>.

YOUTUBE. **Youtube Insights 2017**. 2017. Acesso em: 16 de Abril de 2019. Disponível em: <https://www.thinkwithgoogle.com/intl/pt-br/youtubeinsights/2017/introducao/>.

YOUTUBE. **YouTube Insights**. 2019. Acesso em: 17 de Abril de 2019. Disponível em: <https://youtubeinsights.withgoogle.com>.

YOUTUBE. **Youtube para imprensa**. 2019. Acesso em: 16 de Abril de 2019. Disponível em: <https://www.youtube.com/intl/pt-BR/yt/about/press/>.

ZHAO, R.; MAO, K. Fuzzy bag-of-words model for document representation. **IEEE transactions on fuzzy systems**, IEEE, v. 26, n. 2, p. 794–804, 2017. Disponível em: <https://doi.org/10.1109/TFUZZ.2017.2690222>.

ZHENG, C. et al. Public opinions and concerns regarding the canadian prime minister's daily covid-19 briefing: Longitudinal study of youtube comments using machine learning techniques. **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 23, n. 2, p. e23957, 2021. Disponível em: <https://doi.org/10.2196/23957>.

Apêndices

APÊNDICE A

Questionário

O questionário a seguir (ResQue) é baseado no trabalho de Pu, Chen e Hu (2011). Para as questões, você responderá de 1 a 5 de acordo com a seguinte escala:

1 - Discordo totalmente
2 - Discordo
3 - Nem discordo e nem concordo
4 - Concordo
5 - Concordo totalmente

1. Os materiais recomendados para mim corresponderam aos meus interesses.
2. O sistema de recomendação me ajudou a descobrir novos materiais.
3. Os materiais recomendados para mim foram diversificados.
4. O layout da interface do sistema de recomendação é atrativo.
5. O sistema de recomendação explica porque os materiais foram recomendados para mim.
6. As informações fornecidas para os materiais recomendados são suficientes para eu tomar uma decisão de abri-los.
7. Eu achei fácil informar ao sistema se eu não gosto/gosto do item recomendado.
8. Eu me tornei familiar com o sistema de recomendação muito facilmente.
9. Me sinto no controle para modificar minhas preferências.
10. Eu entendi porque estes conteúdos foram recomendados para mim.
11. O recomendador me deu boas sugestões.
12. Em geral, eu estou satisfeito com o sistema de recomendação.
13. O sistema de recomendação é confiável.
14. Eu usarei este recomendador outra vez.

Caso queira, deixe aqui alguma observação que você julga relevante acerca das recomendações.