Uma abordagem para atribuição automática de metadados sobre enunciados de questões em vídeos educacionais

Gilson R. D. Fonseca[1], **Jairo F. de Souza**[1], **Eduardo Barrére**[1]

[1]LAPIC Research Group – Universidade Federal de Juiz de Fora (UFJF)
Caixa Postal 20.010 – 36.036-900 – Juiz de Fora – MG – Brazil

{gilson, jairo.souza, eduardo.barrere}@ice.ufjf.br

Abstract. *Recommendation systems can be used to organize and retrieve videos used in education based on metadata. Questions in videos can provide useful metadata for this process. However, unlike textual sources, there are no clear markers to identify them, as the content is presented in continuous speech. Additionally, they may be interspersed with explanations given by the teacher during the speech, making identification more complex. This work proposes a model for the automatic identification and classification of questions in english video lessons based on Bloom's Digital Taxonomy. The model was evaluated using classification algorithms, with the BERT model outperforming the others.*

Resumo. *Sistemas de recomendação podem ser utilizados para organizar e recuperar vídeos utilizados na educação, a partir de metadados. Enunciados de questões em videoaulas podem fornecer metadados úteis para esse processo. Porém, ao contrário de fontes textuais, não há marcadores claros que os identifiquem, já que o conteúdo é apresentado em fala contínua. Além disso, eles podem estar mesclados à explicações dadas pelo professor durante a fala, o que torna a identificação mais complexa. Esse trabalho, propõe um modelo de identificação e classificação automática de enunciados de questões em videoaulas em inglês com base na Taxonomia de Bloom Digital. A avaliação do modelo foi realizada utilizando algoritmos classificadores, onde o modelo BERT se destacou como superior aos demais.*

1. Introdução

A vasta quantidade de vídeos disponíveis pode dificultar a identificação dos mais adequados para o aprendizado dos alunos. Sistemas de recomendação podem ajudar, mas a identificação dos temas dos vídeos é complexa devido a descrições textuais geralmente escassas e genéricas [Yang and Meinel 2014] e a dependência dos mecanismos de busca em informações textuais [Coelho and de Souza 2014]. Uma estratégia promissora é o enriquecimento semântico de áudio e vídeo, que melhora a compreensão do conteúdo por meio da inclusão de elementos semânticos [Borges and dos Reis 2019]. Abordagens descritas na literatura atribuem metadados automáticos para, por exemplo, descrever segmentos baseados em tópicos da videoaula [Barrére et al. 2020b] ou o estilo instrucional utilizado no vídeo educacional [Aquino et al. 2023].

Existem na literatura, diversos trabalhos que fazem a identificação e/ou classificação de enunciados de questões [Gani et al. 2023, Razzaghnoori et al. 2018].

DOI: 10.5753/sbie.2024.241876

Eles se utilizam de mídias textuais, como livros e PDFs. Essas mídias possuem marcadores de formatação que permitem identificar partes contendo exercícios e a partir da identificação desses marcadores feita a classificação, como por exemplo, em [Aninditya et al. 2019] que as questões de provas são classificadas segundo a Taxonomia de Bloom e o nível obtido é usado como referência para avaliar o desempenho dos alunos. Abordagens similares podem ser promissoras para vídeos educacionais, fornecendo metadados úteis, como tipo, dificuldade e conceitos abordados na questão, para sistemas de recomendação personalizados ao aprendizado. Entretanto, são escassos os trabalhos para extração desse tipo de metadado em vídeos educacionais.

Ao contrário de mídias textuais, a identificação de enunciados de questões em vídeos enfrenta desafios únicos devido à falta de marcadores visuais claros e à apresentação contínua do conteúdo, onde enunciados de questões estão frequentemente integrados às explicações verbais. Esses desafios podem tornar a abordagem inviável para identificação de enunciados de questões em vídeos e/ou informações insuficientes para uma boa classificação de nível de conhecimento. Com base nisso, são formuladas duas questões de pesquisa neste estudo: **(Q1)**. É possível a identificação dos enunciados de questões em vídeos educacionais, mesmo com a falta de marcadores visuais claros? **(Q2)**. É possível determinar a dificuldade de enunciados de questões em vídeos educacionais?

Para responder as perguntas, foram formuladas as seguintes hipóteses: **(H1)** ferramentas de transcrição automática de áudio podem auxiliar na identificação dos enunciados de questões; **(H2)** modelos de Processamento de Linguagem Natural personalizados, treinados com exemplos específicos de videoaulas, podem melhorar a precisão na identificação de enunciados de questões, mesmo com poucos marcadores visuais; e **(H3)** técnicas de aprendizado de máquina supervisionado podem construir modelos preditivos para determinar a dificuldade dos enunciados de questões, utilizando escalas discretas como a Taxonomia de Bloom.

Com base nas hipóteses apresentadas, este artigo propõe uma abordagem para identificar e classificar automaticamente enunciados de questões em vídeos educacionais em inglês, fazendo uso de *Large Language Models* (LLM) e com base na Taxonomia de Bloom Digital para representar os níveis de conhecimento necessários para solução de um dado enunciado de questão. Com os resultados encontrados neste estudo, espera-se que sistemas educacionais de recomendação e sistemas tutores inteligentes possam fazer uso de metadados mais detalhados para melhoria de processos de aprendizagem.

2. Trabalhos Relacionados

A geração de metadados a partir de vídeos enfrenta desafios devido à composição contínua de imagens e áudio, mas pesquisas têm avançado nas últimas décadas com diversas abordagens documentadas na literatura, variando conforme a aplicação [Pal et al. 2019]. Estudos como [Balasubramanian et al. 2016] extraíram palavras-chave e segmentos baseados em tópicos de videoaulas por transcrição. O uso de transcrição para segmentação de videoaulas foi também proposto por [Barrére et al. 2020b].

Sistemas de recomendação de vídeos educacionais, como o desenvolvido por [Carvalho et al. 2020], usam as categorias, geradas pelo *Youtube* para organizar os vídeos e guiar o processo de recomendação. O sistema BAVi [Lima Dias et al. 2017] faci-

lita a identificação de recursos didáticos por meio do enriquecimento semântico com transcrição e anotações, podendo ser usado para recomendação de videoaulas no Moodle [Barrére et al. 2018, Barrére et al. 2020a]. [Aquino et al. 2022, Aquino et al. 2023] classificou estilos de videoaulas com base em características visuais para atribuir metadados sobre estilos de mídia instrucional.

O uso de técnicas de classificação automática para enunciados de questões tem sido amplamente estudada [Aninditya et al. 2019, Gani et al. 2023, Mohammed and Omar 2020, Razzaghnoori et al. 2018]. No entanto, há uma lacuna significativa na literatura, pois os estudos atuais apresentam soluções para documentos textuais, os quais possuem uma estruturação de texto e linguagem formal que facilitam a identificação de enunciados. Alguns trabalhos, como o de [Omar et al. 2012], utilizam uma abordagem baseada em regras, classificando enunciados de questões a partir dos verbos na Taxonomia de Bloom e padrões das questões. Este estudo foi expandido por [Haris and Omar 2015], que combinou regras com abordagens estatísticas para melhorar a precisão da classificação. Outros estudos adotam abordagens de aprendizado supervisionado, utilizando características dos enunciados de questões para treinar classificadores, como em [Yahya et al. 2012] e [Setyaningsih and Listiowarni 2021].

Embora os trabalhos apresentem abordagens por vezes eficazes para atribuição de metadados sobre enunciados de questões em fontes textuais devidamente formatadas, estas abordagens não são diretamente aplicáveis para vídeos educacionais. A extração de enunciados em videoaulas é desafiadora devido à ausência de marcadores visuais claros na transcrição contínua das falas, o que torna difícil identificar as perguntas que estão misturadas com as explicações do professor. Ainda, estes não exploram amplamente abordagens recentes como as LLM. O presente estudo apresenta como diferencial o uso de abordagens mais modernas para contribuir na identificação e atribuição de metadados sobre enunciados de questões em vídeos educacionais.

3. Abordagem para identificação e classificação de enunciados de questões

Neste trabalho, considera-se como enunciados de questões os trechos que abordam explicitamente conceitos, princípios ou operações educacionais, apresentando uma pergunta clara ou problema imperativo (questão expressa na forma de uma instrução) para os quais se espera uma resposta ou solução, podendo ou não ser acompanhados de um enunciado explicativo. Assim, trechos que contém questões gerais, opiniões ou discussões do conteúdo sem um comando claro por parte do professor não deverão ser classificados como enunciados de questões. Esta Seção apresenta o processo para identificação e classificação automática de enunciados de questões.

Conforme a Figura 1, o processo se inicia com a recepção da videoaula, onde seu áudio é transcrito para texto. Em seguida, é realizada a identificação de trechos do texto que são possíveis de conter descrição de enunciados. Depois, esses trechos são classificados de acordo com os diferentes níveis da Taxonomia de Bloom Digital. Ao final do processo, tem-se o enunciado de questão e sua classificação correspondente.

3.1. Módulo de Transcrição

O módulo de transcrição tem como objetivo converter o conteúdo de uma videoaula em texto. Foram considerados, para a transcrição, o *wav2vec2-large-xlsr-53-english*

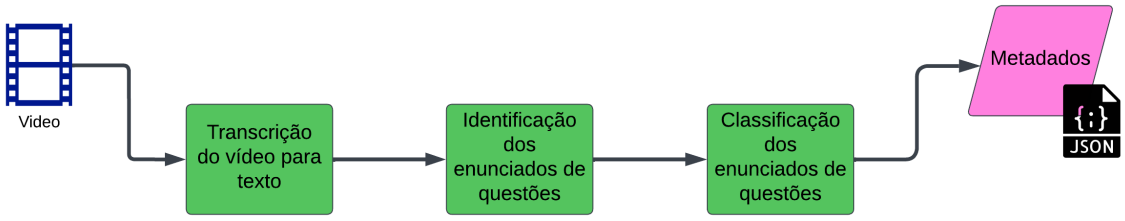

Figura 1. Fluxograma da solução

[Baevski et al. 2020], o *wav2vec2-large-960h-lv60-self* [Baevski et al. 2020] e, finalmente, o *Whisper* [Radford et al. 2023]. Decidiu-se pelo uso do *Whisper*, pois este demonstrou um Word Error Rate (WER) de apenas 9%, inferior a outros modelos [Radford et al. 2023]. O que sugere uma transcrição mais eficaz das videoaulas selecionadas.

O *Whisper* foi treinado com 680.000 horas de áudio, compreendendo dados de vários conjuntos, incluindo *Common_Voice* e *VoxPopuli_En*. Dentre os modelos disponíveis para uso, o *small.en* foi o selecionado, pois mantém um desempenho satisfatório em comparação com os modelos de maior tamanho, ao mesmo tempo em que requer um consumo reduzido de recursos computacionais e um tempo de execução menor. O resultado produzido pelo modelo é salvo em um arquivo texto e, para evitar textos muito longos, dividido em partes menores, denominadas *chunks*. Esses pedaços menores são então salvos no mesmo arquivo.

3.2. Módulo de Identificação dos enunciados de questões

Para o módulo de identificação dos enunciados de questões, foi escolhido o GPT-3.5 Turbo, um modelo de linguagem desenvolvido pela OpenAI. Este modelo é reconhecido por sua capacidade abrangente de processamento de linguagem natural, incluindo funções como geração de texto, tradução automática e resposta a perguntas. Além disso, o GPT-3.5 Turbo também conta com uma quantidade significativa de parâmetros, estimada em cerca de 175 bilhões.

Este modelo é responsável por identificar os enunciados de questões contidos no *chunk* fornecido como entrada após o processo de treinamento via *few-shot*. Neste módulo, é feita uma requisição de conversação ao modelo escolhido, no qual são passados os parâmetros necessários para que ele possa identificar e retornar os enunciados. Após conclusão da identificação dos enunciados de questões, as questões identificadas são salvas em um novo arquivo.

3.3. Módulo de Classificação dos enunciados de questões

Por fim, temos o módulo de classificação de enunciados de questões o qual foi construído utilizando métodos distintos para comparação, sendo uma abordagem baseada em regras e abordagens baseadas em aprendizado de máquina. Nestas últimas, foram adotados um modelo Naive Bayes (NB), BERT e o GPT-3.5 Turbo. O Naive Bayes foi escolhido devido à sua velocidade no treinamento e na classificação, enquanto a rede neural BERT foi selecionada por suas diversas vantagens, incluindo a disponibilidade de versões pré-treinadas, o que economiza tempo e recursos computacionais. Por fim, o GPT-3.5 Turbo foi escolhido com a finalidade de testar a utilidade de IA generativa para esse problema.

São recebidos, como entrada, os enunciados identificados pelo módulo anterior e, em seguida, os classifica, em um dos níveis da Taxonomia de Bloom Digital, utilizando os diferentes modelos de classificação apresentados. A classificação resultante é então gravada juntamente com o enunciado.

4. Experimentos

Esta Seção discute a implementação da abordagem para a identificação e classificação automática de enunciados de questões em videoaulas. Sua estrutura é organizada da seguinte forma: as bases de dados utilizadas neste trabalho são expostas em detalhes na Seção 4.1, seguidos pela análise dos resultados na Seção 4.2.

4.1. Base de dados

Aqui, serão detalhadas as bases de dados selecionadas, destacando suas características principais. Todas as bases geradas estão disponíveis no repositório público[1].

Base Bloom: A base de dados utilizada é a *Bloom's Taxonomy Cognitive Levels Data Set*, que consiste em enunciados de questões coletados da Web e classificados segundo a Taxonomia de Bloom. O conjunto de dados, empregado em estudos anteriores [Yahya 2011, Yahya et al. 2012], possui duas colunas: *Question*, com os exercícios, e *Level*, com a classificação manual dos níveis cognitivos feita por um especialista pedagógico. Cada um dos seis níveis da Taxonomia possui 100 enunciados de questões. Essa base foi escolhida para treinamento de classificadores devido a estrutura dos enunciados e à classificação manual realizada por profissionais. Para a classificação automática, a coluna *Level* foi remapeada para números na coluna *Label*, representando os níveis da Taxonomia de Bloom Digital (0=Lembrar, 1=Compreender, 2=Aplicar, 3=Analisar, 4=Avaliar, 5=Criar).

Base de Videoaulas (BVA): Para classificar os enunciados de questões, foram identificadas videoaulas adequadas para a Taxonomia de Bloom Revisada. Selecionaram-se 42 videoaulas licenciadas sob *Creative Commons*. Estas videoaulas, todas em inglês, foram obtidas do *YouTube* e têm uma duração total de 16,28 horas, variando de 4 a 61 minutos, com média de 23 minutos cada. As videoaulas foram categorizadas da seguinte forma: 32 vídeos contém uma aula completa que inclui trechos de explanação do professor e enunciados de questões em um ou mais trechos do vídeo, enquanto os restantes são vídeos que contém apenas de enunciados e resoluções. Essa diversidade representa um desafio adicional na identificação de enunciados de questões, uma vez que, em videoaulas com explanação do conteúdo, é possível que o professor mescle enunciados de questões com informações adicionais.

Base de enunciados extraídos manualmente (BEM): A partir das videoaulas selecionadas, foi elaborada uma base de dados contendo 139 enunciados identificados manualmente. A base é composta por nove colunas: *Question* (enunciados de questões), *Source* (videoaula de origem), *Subject* (tema), *VideoClass Style* (estilo da videoaula), *Bloom's Style* (formatação conforme a Taxonomia de Bloom Digital), *Manual Classification* (classificação manual nos níveis cognitivos da Taxonomia de Bloom Digital) e colunas adicionais para as classificações automáticas dos modelos testados. Além dos

[1] https://drive.google.com/drive/folders/17jP72jU9R6Z1EFmXttuLUiQvBx_EYZas?usp=sharing

139 enunciados válidos, foram incluídos 19 enunciados inválidos, contendo apenas as colunas *Question* e *Source*.

Base de enunciados extraídos automsticamente (BEA): foi gerada a partir do resultado do processo de identificação automática de enunciados de questões. Esta base contém um total de 107 enunciados identificados automaticamente, dos quais 102 foram considerados válidos, pois estão de acordo com a definição de enunciado de questão apresentada no trabalho, enquanto 5 enunciados foram classificados como inválidos, pois ou são perguntas relacionadas a *feedback* ou perguntas feitas durante a resolução de um exercício. Assim como a BEM, esta base também é composta por 9 colunas, contendo as mesmas informações e também utilizada para classificação automática. A distribuição manual dos enunciados entre os diferentes níveis da TBD, foi a seguinte: 42 Lembrar, 2 Compreender, 37 Aplicar, 14 Analisar, 5 Avaliar e 2 Criar. Os demais 5 enunciados inválidos não possuem classificação.

4.2. Análise dos resultados

Nesta Seção, serão abordados os processos de identificação e classificação dos enunciados de questões presentes nas videoaulas. O objetivo é analisar os resultados obtidos, com os métodos e técnicas utilizados para identificar os enunciados nos textos transcritos das videoaulas e, em seguida, classificá-las de maneira automatizada. A Tabela 1 apresenta a relação entre as bases de dados, as ferramentas utilizadas e os processos apresentados.

Tabela 1. Relacional entre o processo, base de dados e ferramenta utilizados

Processo	Base de Dados	Ferramenta
Transcrição	BVA	*Whisper*
Identificação	BEM e BEA	*GPT-3.5 few-shot*
Classificação	Base Bloom, BEM e BEA	BR, NB, BERT e *GPT-3.5 FT*

4.3. Identificação dos enunciados de questões

Foram realizados dois experimentos distintos para identificar os enunciados de questões. No primeiro experimento, o objetivo era determinar enunciados como válidos e inválidos. No segundo experimento, foram identificados e retornados os enunciados válidos a partir da transcrição.

A identificação automática dos enunciados de questões foi realizada utilizando a abordagem de *few-shot learning* no GPT-3.5 Turbo. Uma amostra selecionada da base BEM, composta por cinco exemplos de enunciados válidos e seis exemplos de enunciados inválidos, conforme definido no estudo, foi empregada para treinamento. Para os testes, uma amostra de 29 enunciados de questões, extraída da mesma base, foi utilizada. Destes, 16 eram considerados válidos e 13 inválidos. Os resultados demonstraram a eficácia da abordagem *few-shot*, com valores de precisão, revocação e F1-score atingindo 93,8%, e acurácia de 93,1%.

Um erro identificado foi a classificação errônea da pergunta “Então, vamos analisar o trecho e nos perguntar. Qual é o efeito específico que o escritor está tentando alcançar?” como válida. Enquanto o enunciado “Uma muito semelhante. A única diferença é que agora, em vez de se mover a uma velocidade constante, diz que está

sendo baixado a uma velocidade decrescente. Qual é verdadeiro?", que não seguia a estrutura típica dos demais foi erroneamente classificado como inválido. Este último caso evidencia as dificuldades do modelo em reconhecer enunciados de questões que fogem à norma, apesar de se alinharem à definição estabelecida.

O segundo experimento consistiu em uma reformulação do *prompt*, no qual foi solicitado que, utilizando o conhecimento sobre enunciados válidos e inválidos, o modelo identificasse e retornasse os enunciados válidos contidos no texto transcrito. Foram removidos dessa etapa, 21 enunciados válidos utilizados no primeiro experimento. Foi observado que o modelo GPT, criou enunciados de questões com base nas informações apresentadas nas aulas, extrapolando os resultados fornecidos. Esses enunciados foram classificados como alucinações, sendo consequentemente excluídos dos experimentos. Dos 118 enunciados restantes, o GPT demonstrou uma precisão de 95,3%, uma revocação de 86,4% e um F1-score de 91,0%. A acurácia alcançada foi de 83,0%. 102 enunciados foram identificados corretamente como válidos, enquanto 16 foram classificados como inválidos. Foram encontrados também 5 enunciados na transcrição textual que não correspondiam a enunciados válidos. Esses resultados contribuíram para a criação da BEA.

4.4. Classificação dos enunciados de questões

Para todos os classificadores propostos utilizou-se a Base Bloom contendo 600 enunciados de questões.

A abordagem baseada em regras (BR) utiliza um conjunto suplementar de dados contendo verbos associados aos diversos níveis da Taxonomia de Bloom. Resumidamente, os enunciados são comparados com essa lista de verbos até que uma correspondência seja encontrada, o que permite determinar o nível cognitivo associado a cada exercício. Para melhorar a precisão dessa técnica, os enunciados foram convertidos em vetores de palavras, com remoção de palavras com menos de 3 caracteres e sinais de pontuação. Isso simplificou o processo de correspondência, já que o segundo conjunto de dados não incluía verbos com menos de 3 caracteres, reduzindo a complexidade da busca nos vetores de palavras. Posteriormente, as palavras presentes em cada enunciado foram comparadas com os verbos contidos na base de dados. A presença de uma correspondência indicava o nível específico da Taxonomia de Bloom associado ao enunciado.

Na classificação com NB e BERT, a base foi dividida em duas partes: 498 enunciados de questões para validação e 102 para teste. A base de validação foi subdividida em conjuntos de treinamento e validação utilizando a técnica de validação cruzada *10-fold*. Nesse método, os dados foram divididos em 10 partes iguais, onde em cada iteração, 9 partes foram utilizadas para treinar o modelo e 1 parte foi reservada para validar seu desempenho. Esse processo foi repetido 10 vezes para assegurar a utilização de todas as partes dos dados para treinamento e validação, possibilitando a comparação dos resultados.

No treinamento do NB, após a etapa de separação das bases de treinamento e validação, as amostras passam por uma etapa de vetorização para prepará-las para a execução do treinamento. Devido à natureza do problema, que envolve múltiplas respostas, foi utilizada a versão Multinomial do NB. Essa escolha permite identificar não apenas se um enunciado pertence ou não a um determinado nível cognitivo em uma classificação binária, mas sim determinar o nível cognitivo específico ao qual o enunciado pertence.

No treinamento do modelo BERT, utiliza-se um modelo pré-treinado *bert-base-uncased*[2] [Devlin et al. 2019], introduzido em 2018. Este modelo foi treinado com uma grande quantidade de dados, incluindo textos da *Wikipedia* em inglês e do *Toronto Book Corpus*, que contém mais de 11 mil livros não publicados. Os textos são tokenizados usando a técnica *WordPiece*, com um vocabulário de tamanho 30 mil, e o modelo emprega uma variedade de objetivos de modelagem de linguagem, como a aplicação de máscaras e a predição da próxima sentença do texto.

No modelo BERT, as *features* das instâncias foram *tokenizadas*. O tamanho máximo das sentenças foi limitado a 256 *tokens* devido a restrições de hardware. Máscaras de atenção foram criadas para indicar quais *tokens* são relevantes para a classificação e quais são *paddings*, adicionados para padronizar o tamanho das sentenças. Durante o treinamento, foi utilizado o otimizador *AdamW*, com uma taxa de aprendizado de 2E-5 e um valor de ε igual a 1E-8. O treinamento foi conduzido em oito épocas. A classificação com o GPT utilizou a técnica de *fine-tuning*. A base foi preparada em formato *jsonl* para este propósito e dividida em conjuntos de validação e teste. Após a preparação, os arquivos foram carregados na plataforma da *OpenAI* para realizar o *fine-tuning* e criar o novo modelo.

Após a configuração dos classificadores, os resultados foram avaliados. O método NB alcançou uma média de 67% de acurácia, com baixo desvio padrão de 6%, indicando consistência nas classificações. O método BERT obteve uma média de 74% de acurácia, com desvio padrão de 12%, mostrando maior variabilidade. Já o GPT 3.5 com *fine-tuning* destacou-se com uma média de 91% de acurácia, porém com desvio padrão de 14%, sugerindo uma variação mais ampla nos resultados.

Após o treinamento, a base de teste avaliou a classificação usando NB, BERT e GPT. Esses métodos mostraram desempenho mais satisfatório que a abordagem baseada em regras. Os modelos treinados foram salvos para uso posterior na etapa de classificação da base de enunciados de questões, a qual foi construída a partir das videoaulas. A Tabela 2 compila os resultados, incluindo dados do modelo descrito em [Yahya 2011].

O Coeficiente de Correlação de Matthews (MCC) foi utilizado para avaliar as previsões, dada a desigualdade entre as classes. Este coeficiente varia de -1 a +1, onde +1 indica uma previsão perfeita, 0 uma previsão média aleatória e -1 uma previsão inversa. Os valores do MCC foram 23% para o BR, 36% para o GPT, 65% para o NB e 84% para o BERT.Os valores de F1 médio foram 27%, 49%, 73% e 88% para BR, GPT, NB e BERT, respectivamente. As acurácias micro mostraram pouca variação para GPT, NB e BERT (47%, 71% e 86%), mas para o BR a diferença foi significativa, obtendo 34%. O desempenho inferior da abordagem baseada em regras, comparado aos demais classificadores, pode ser atribuído à presença de termos implícitos ou com múltiplos significados, dificultando a classificação precisa com essa abordagem.

A abordagem GPT-3.5 Turbo *Fine-Tuning* teve um desempenho inferior ao NB e ao BERT. Isso pode ser atribuído ao processo de *fine-tuning* realizado na plataforma web da *OpenAI*, que não permite a configuração de certos hiperparâmetros, como o número de épocas, limitado a 3. Além disso, o GPT-3.5 possui uma vasta lista de 175 bilhões de parâmetros, o que pode resultar em um desempenho mais baixo no *fine-tuning* em

[2] https://github.com/google-research/bert

Tabela 2. Resultado do treinamento com a Base Bloom

	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
		BR			**Naive Bayes**			**BERT**	
Lembrar	0,43	0,83	0,57	0,83	0,59	0,69	1,00	1,00	**1,00**
Compreender	0,37	0,48	0,42	0,67	0,71	0,69	0,89	0,94	**0,91**
Aplicar	0,57	0,67	0,61	0,60	0,71	0,65	0,83	0,88	**0,86**
Analisar	0,09	0,02	0,04	0,93	0,76	0,84	0,93	0,82	**0,87**
Avaliar	0,07	0,03	0,04	0,72	0,76	0,74	0,68	0,88	**0,77**
Criar	0,07	0,03	0,04	0,60	0,71	0,65	0,92	0,65	**0,76**
Média	0,27	0,36	0,29	0,73	0,71	0,71	**0,88**	**0,86**	**0,86**
micro-ACC			0,34			0,71			0.86
MCC			0,23			0,65			**0.84**
	GPT 3.5 Turbo			**[Yahya 2011]**					
Lembrar	0,62	0,94	0,74	1,00	0,09	0,17			
Compreender	0,83	0,59	0,69	0,50	0,25	0,33			
Aplicar	0,53	0,47	0,50	1,00	0,08	0,14			
Analisar	0,79	0,65	0,71	0,75	0,19	0,30			
Avaliar	0,09	0,12	0,10	1,00	0,50	0,67			
Criar	0,08	0,06	0,07	0,90	0,64	0,75			
Média	0,49	0,47	0,47	0,86	0,29	0,39			
micro-ACC			0.47			**0.87**			
MCC			0.36			N/D			

uma base de dados pública e conhecida, como a Base Bloom. O método BERT mostrou um desempenho superior na tarefa de classificação, com um valor mais elevado do MCC em comparação com o NB, sugerindo uma correlação mais forte entre as previsões e as observações reais.

Comparando o desempenho do BERT com o modelo SVM apresentado por [Yahya 2011], observamos uma maior precisão em um conjunto de classes para cada abordagem, resultando em uma acurácia micro ligeiramente similar (87% para o SVM e 86% para o BERT). Embora a abordagem de [Yahya 2011] seja mais confiável para algumas classes específicas, ela não consegue equilibrar precisão e revocação de forma aceitável. Os baixos valores de revocação e F1 relatados por [Yahya 2011] são atribuídos ao baixo volume de instâncias na base (apenas 600 registros). Por outro lado, a proposta atual supera esse problema, alcançando um F1 significativamente maior para todas as classes, resultando em uma média dos valores de F1 mais alta (86% vs. 39%).

Após treinamento e teste dos modelos, foram classificadas as bases BEM e BEA, utilizando os classificadores que obtiveram um desempenho satisfatório na etapa anterior, as Tabelas 3 e 4 apresentam os resultados obtidos. Os resultados para a base de enunciados de questões manualmente identificados foram superiores aos obtidos automaticamente, sugerindo que a forma como os enunciados de questões foram selecionados pode não ser a mais adequada para a etapa de classificação subsequente. O método Naive Bayes (NB) apresentou valores semelhantes em ambas as classificações, com um MCC de 23% e 20% para as bases BEM e BEA, respectivamente. Apesar de não alcançar o desempenho dos outros dois métodos, a consistência dos resultados é evidente, com apenas

uma divergência de 3pp na acurácia entre elas.

Tabela 3. Tabela de resultado da classificação BEM

	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
	Naive Bayes			**BERT**			**GPT-3.5 Turbo FT**		
Lembrar	0,74	0,31	0,44	0,71	0,94	**0,81**	0,68	0,83	0,75
Compreender	0,06	0,33	0,10	0,60	1.00	**0,75**	0,50	0,33	0,40
Aplicar	0,52	0,59	0,55	0,82	0,70	**0,76**	0,73	0,59	0,65
Analisar	0,47	0,37	0,41	1,00	0,50	**0,67**	0,64	0,47	0,55
Avaliar	0,11	0,17	0,13	0,60	0,50	**0,55**	0,18	0,33	0,24
Criar	0,08	0,33	0,12	0,67	0,67	**0,67**	0,50	0,33	0,40
Média	0,33	0,35	0,29	**0,73**	**0,71**	**0,69**	0,54	0,48	0,50
micro-ACC			0,42			**0,76**			0,65
MCC			0,23			**0,64**			0,48

Tabela 4. Tabela de resultado da classificação BEA

	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
	Naive Bayes			**BERT**			**GPT-3.5 Turbo FT**		
Lembrar	0,68	0,36	0,47	0,68	0,81	**0,74**	0,65	0,36	0,46
Compreender	0,11	0,50	0,18	0,67	1,00	**0,80**	0,00	0,00	0,00
Aplicar	0,44	0,43	0,44	0,63	0,51	**0,57**	0,57	0,70	0,63
Analisar	0,33	0,36	0,34	0,67	0,43	**0,52**	0,40	0,71	0,51
Avaliar	0,10	0,20	0,13	0,60	0,50	**0,55**	0,50	0,40	0,44
Criar	0,20	1,00	0,33	0,50	1,00	**0,67**	0,50	0,50	0,50
Média	0,31	0,47	0,32	**0,61**	**0,73**	**0,64**	0,44	0,45	0,42
micro-ACC			0,39			**0,65**			0,53
MCC			0,20			**0,48**			0,35

Os resultados obtidos com os demais métodos foram superiores aos do NB, porém há uma falta de consistência semelhante entre os resultados da BEM e da BEA em ambos. No caso do GPT-3.5 Turbo *Fine-Tunned*, o MCC reduziu de 48% para 35%, e a acurácia de 65% para 53%. Para o BERT, o MCC foi de 64% para 48%, e a acurácia saiu de 76% para 65%. Embora os resultados dos métodos GPT e BERT tenham sido superiores aos do NB, a consistência entre as classificações da BEM e da BEA não foi tão evidente.

Pode-se observar que o método BERT manteve um desempenho superior em comparação aos demais. Além disso, ao analisar ambas as bases, observa-se que os valores de precisão, revocação e F1-score do BERT são consistentes entre os diferentes níveis da TBD, demonstrando um certo equilíbrio. Este padrão é evidenciado pelo fato de que o método obteve uma acurácia superior a 65% em ambas as bases.

Em contraste, tanto o NB quanto o GPT enfrentaram dificuldades na classificação dos níveis com poucos enunciados identificados, como Compreender, Avaliar e Criar. Por exemplo, o método NB teve uma precisão máxima de 20% para o nível Criar da base BEA, enquanto o GPT obteve 0% para o nível Compreender da mesma base. O que destaca desafios específicos enfrentados por esses métodos ao lidar com dados escassos ou desbalanceados, evidenciando a necessidade de aprimoramento e ajustes para melhorar sua eficácia nessas situações.

5. Considerações finais

O estudo apresentou um modelo de classificação automatizada de enunciados de questões com base na Taxonomia Digital de Bloom. Os resultados usando o GPT foram satisfatórios, apesar dos dados limitados, confirmando as hipóteses **H1** e **H2**. O BERT se destacou entre os classificadores, confirmando a importância da semântica na classificação **H3**. Isso ressalta a necessidade de explorar a semântica dos exercícios como um recurso para aprimorar a classificação automática e o desempenho dos sistemas de recomendação educacional.

Além disso, essa classificação possibilita a incorporação de informações semânticas adicionais em videoaulas, permitindo que sistemas de recomendação educacionais ofereçam conteúdos mais personalizados e precisos de acordo com as necessidades de aprendizado do aluno. Isso abre caminho para uma maior eficácia no fornecimento de materiais educacionais alinhados aos objetivos de aprendizado do aluno.

As limitações desse trabalho podem ser abordadas em quatro áreas distintas. Primeiramente, a escolha da ferramenta para transcrever os vídeos pode afetar os resultados, especialmente considerando variações linguísticas. Além disso, a identificação dos enunciados de questões depende de uma técnica de engenharia de *prompt*, cujas escolhas podem não ser ideais para obter os melhores resultados, e está sujeita às capacidades do modelo GPT-3.5 Turbo, que possui um número limitado de parâmetros em comparação com versões mais recentes. A falta de avaliação por especialistas educacionais pode resultar em classificações imprecisas dos enunciados de questões, sugerindo a necessidade de revisões por especialistas ou validação manual.

Este estudo inaugura um campo de investigação, sendo possivelmente a primeira abordagem a classificar enunciados de questões em videoaulas, abrindo caminho para a criação de metadados específicos para esse tipo de conteúdo. Foram apresentadas duas estratégias distintas de classificação de enunciados de questões: uma baseada em regras e outras em aprendizado de máquina, usando apenas o áudio. No entanto, considerando que os vídeos possuem múltiplos canais de comunicação, abordagens híbridas como as propostas em [Haris and Omar 2015] podem oferecer resultados mais robustos. O uso de segmentação de vídeos [Barrére et al. 2020b] e reconhecimento óptico de caracteres (OCR) também pode ser explorado para agregar informações adicionais à classificação. Além disso, a análise da complexidade da aula como um todo através da classificação dos enunciados de questões pode proporcionar novos insights, considerando a interdependência dos diferentes níveis da TBD. Experimentos futuros com conjuntos de dados mais diversificados podem revelar modelos de classificação ainda mais precisos do que o BERT.

Referências

Aninditya, A., Hasibuan, M. A., and Sutoyo, E. (2019). Text mining approach using tf-idf and naive bayes for classification of exam questions based on cognitive level of bloom's taxonomy. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pages 112–117, Bali, Índia. IEEE, IEEE.

Aquino, B., Barrére, E., and de Souza, J. F. (2022). Classificação automática de estilos de videoaulas. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 956–967, Porto Alegre, RS, Brasil. SBC.

Aquino, B., de Souza, J. F., and Barrére, E. (2023). Automatic classification of learning material styles. *Revista Brasileira de Informática na Educação*, 31:906–924.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Balasubramanian, V., Doraisamy, S. G., and Kanakarajan, N. K. (2016). A multimodal approach for extracting content descriptive metadata from lecture videos. *Journal of Intelligent Information Systems*, 46:121–145.

Barrére, E., Alvim de Almeida, M., Aparecida Vitor, M., and Francisco de Souza, J. (2020a). Utilização de enriquecimento semântico para a recomendação automática de videoaulas no moodle. *Revista Brasileira de Informática na Educação*, 28(1).

Barrére, E., Souza, J., and Soares, E. R. (2020b). Framework para segmentação temporal de vídeos educacionais. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 972–981. SBC.

Barrére, E., Souza, J., Vitor, M. A., and de Almeida, M. A. (2018). Recomendação automática de videoaulas no moodle. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1613.

Borges, M. V. M. and dos Reis, J. C. (2019). Semantic-enhanced recommendation of video lectures. In *2019 ieee 19th international conference on advanced learning technologies (icalt)*, volume 2161, pages 42–46. IEEE.

Carvalho, H., Pitangui, C., ASSIS, L., and VIVAS, A. (2020). Educavıdeos: Um sistema de recomendaçao de objetos de aprendizagem de vıdeos educacionais do youtube. In *XVII Congresso Brasileiro de Ensino Superior a disttância*.

Coelho, S. A. and de Souza, J. F. (2014). Anotação semântica de transcritos para indexação e busca de vídeos. In *Conferência Ibero Americana (WWW/INTERNET)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Gani, M. O., Ayyasamy, R. K., Sangodiah, A., and Fui, Y. T. (2023). Bloom's taxonomy-based exam question classification: The outcome of cnn and optimal pre-trained word embedding technique. *Education and Information Technologies*, pages 1–22.

Haris, S. S. and Omar, N. (2015). Bloom's taxonomy question categorization using rules and n-gram approach. *Journal of Theoretical & Applied Information Technology*, 76(3).

Lima Dias, L., Barrére, E., Siqueira Barbosa, J., and de Souza, J. F. (2017). Uma abordagem para identificação de similaridade entre recursos educacionais utilizando bases de conhecimento externas. *Revista Brasileira de Informática na Educação*, 25(2).

Mohammed, M. and Omar, N. (2020). Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec. *PloS one*, 15(3):e0230442.

Omar, N., Haris, S. S., Hassan, R., Arshad, H., Rahmat, M., Zainal, N. F. A., and Zulkifli, R. (2012). Automated analysis of exam questions according to bloom's taxonomy. *Procedia-Social and Behavioral Sciences*, 59:297–303.

Pal, S., Pramanik, P. K. D., Majumdar, T., and Choudhury, P. (2019). A semi-automatic metadata extraction model and method for video-based e-learning contents. *Education and Information Technologies*, 24:3243–3268.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Razzaghnoori, M., Sajedi, H., and Jazani, I. K. (2018). Question classification in persian using word vectors and frequencies. *Cognitive Systems Research*, 47:16–27.

Setyaningsih, E. R. and Listiowarni, I. (2021). Categorization of exam questions based on bloom taxonomy using naïve bayes and laplace smoothing. In *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, pages 330–333. IEEE.

Yahya, A. (2011). Automatic classification of questions into bloom's cognitive levels using support vector machines. In *The International Arab Conference on Information Technology*.

Yahya, A. A., Toukal, Z., and Osman, A. (2012). Bloom's taxonomy–based classification for item bank questions using support vector machines. In Ding, W., Jiang, H., Ali, M., and Li, M., editors, *Modern Advances in Intelligent Systems and Tools*, pages 135–140, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yang, H. and Meinel, C. (2014). Content based lecture video retrieval using speech and video text information. *IEEE transactions on learning technologies*, 7(2):142–154.