



Classificação de vídeos educacionais do YouTube por níveis de ensino utilizando comentários: uma abordagem experimental.

Catrine dos Santos Oliveira, UFSJ - Campus Alto Paraopeba - Ouro Branco - MG - Brasil, catrine.sntsoliveira@gmail.com, <https://orcid.org/0009-0000-4897-8436>
Cristiano Grijó Pitangui, UFSJ - Campus Alto Paraopeba - Ouro Branco - MG - Brasil, pitangui.cristiano@ufsj.edu.br, <https://orcid.org/0000-0002-3961-2042>
Henrique Carlos Fonte Boa Carvalho, UFU - Campus Santa Mônica - Uberlândia - MG - Brasil, henriquefbc@gmail.com, <https://orcid.org/0000-0002-6109-8000>
Alessandro Vivas Andrade, UFVJM - Campus Diamantina - Diamantina - MG - Brasil, alessandrovivas@ufvjm.edu.br, <https://orcid.org/0000-0003-4713-5159>
Luciana Pereira de Assis, UFVJM - Campus Diamantina - Diamantina - MG - Brasil, lpassis@ufvjm.edu.br, <https://orcid.org/0000-0002-7891-7172>
Fabiano Azevedo Dorça, UFU - Campus Santa Mônica - Uberlândia - MG - Brasil, fabianodor@ufu.br, <https://orcid.org/0000-0003-3281-0246>

Resumo: O YouTube é uma plataforma voltada para vídeos que contempla conteúdos de várias áreas, incluindo a Educacional. Embora a ferramenta possua extenso material, seu mecanismo de busca é falho em alguns aspectos, o que é detrimental para alunos e professores que utilizam a plataforma no processo de ensino-aprendizagem. Esse trabalho utiliza Técnicas de Mineração de Texto, e de Aprendizagem de Máquina que usam os comentários dos vídeos como uma alternativa para classifica-los em diversas categorias educacionais. Realizou-se um extenso número de experimentos, e resultados demonstram que os classificadores JRIP, J48, e LMT são eficazes em classificar os vídeos em suas categorias educacionais, com acurácia de até 87%.

Palavras-chave: YouTube, vídeos, educacional, classificação.

Classification of educational YouTube videos by education levels using comments: an experimental analysis.

Abstract: *YouTube is a video platform that posses content from various areas, including the Educational one. Although the tool has an extensive collection of material, its search engine is limited in some aspects, which is detrimental to students and teachers that use the platform in the teaching and learning process. This work uses Text Mining and Machine Learning Techniques that use the video' comments as an alternative to classify them in different educational categories. An extensive number of experiments is performed, and results demonstrate that the JRIP, J48, and LMT classifiers are effective in classifying videos into their educational categories, with an accuracy of up to 87%.*

Keywords: YouTube, video, educational, classification.

1. Introdução

O YouTube é uma plataforma de compartilhamento de vídeos lançada no final de junho de 2005, desenvolvida por Chad Hurley, Steve Chen e Jawed Karim. Nela é possível encontrar conteúdos de entretenimento e também de diversas áreas de conhecimento. A utilização da plataforma é bastante intuitiva, e, desse modo, não exige do seu usuário alto nível de conhecimento técnico.

O YouTube (2019) aponta que mais de 500 horas de conteúdo são enviadas para a plataforma a cada minuto, e destaca também que sem o seu mecanismo de busca e



de classificação de vídeos, seria quase impossível para o usuário encontrar o que deseja assistir em meio ao vasto conteúdo. A diversidade de vídeos educacionais presentes no YouTube permite que a plataforma seja interpretada, no âmbito Educacional, como um repositório de Objetos de Aprendizagem (OA). Segundo Wiley (2000), os OA podem ser entendidos como qualquer recurso digital que possa ser reutilizado como apoio para a Educação. Para Audino (2012), os OA possuem grande importância no processo de ensino-aprendizagem, pois fornecem a capacidade de simular e animar fenômenos, e permitem sua reutilização em vários outros ambientes de aprendizagem.

O YouTube (2019) destaca que durante uma pesquisa na plataforma, o seu mecanismo de busca prioriza: relevância, envolvimento e qualidade. Para garantir a relevância do retorno da busca, são analisados diversos fatores, como o título, *tags* e descrição dos vídeos que correspondam à consulta. Ainda assim, é notável que em muitos casos o retorno da busca apresente um número muito elevado de vídeos, sendo muitos não relacionados a pesquisa realizada, ou mesmo sem relevância com o tema pesquisado.

Uma possível maneira de se refinar os resultados de uma busca por vídeos é pela especificação de uma ou mais categoria(s) para os mesmos. Nesse sentido, Carvalho *et al.* (2020) retrata a impossibilidade do mecanismo de busca do YouTube de filtrar a busca por categorias dos vídeos. Esta restrição implica, como apontado, em um grande número de vídeos irrelevantes retornados. Outra limitação apontada em Carvalho *et al.* (2020) se relaciona à restrição presente no mecanismo de busca da plataforma, que, caso a expressão de busca contenha mais de um termo, a pesquisa retornará vídeos que contêm, em seus dados, todos os termos da expressão. Nesse caso, e de maneira oposta a primeira restrição apontada, corre-se o risco de se desconsiderar materiais que seriam úteis ao usuário da plataforma.

Sob a ótica dos apontamentos supracitados, aponta-se que os textos dos comentários dos vídeos presentes na plataforma contêm informações importantes a respeito do conteúdo dos mesmos, como por exemplo, se os usuários que assistiram aos vídeos gostaram dos seus conteúdos, e se, tais conteúdos, condizem ou não com as temáticas apresentadas nos títulos dos vídeos. Ademais, o vocabulário utilizado nos comentários dos vídeos tendem a ter relação aos assuntos neles abordados Carvalho *et al.* (2022). Nesse sentido, Carvalho *et al.* (2022) utiliza-se dessa premissa, e estuda as diferenças entre os vocabulários dos comentários de vídeos educacionais e não educacionais, constatando discrepâncias fundamentais entre eles, e apontando que, em experimentos realizados, o algoritmo *Random Forest* é capaz de categorizar vídeos, por meio dos vocábulos de seus comentários, em educacionais ou não educacionais, com acurácia superior a 90%.

O presente trabalho utiliza Técnicas de Mineração de Texto e de Aprendizagem de Máquina para analisar os comentários de vídeos do YouTube, e apresenta uma metodologia eficiente para classifica-los em diversas “categorias” educacionais. Inicialmente, explora-se a classificação dos vídeos como educacionais ou não educacionais, e, posteriormente, são estudadas classificações considerando o nível de ensino, em: Ensino Fundamental II, Ensino Médio, e Ensino Superior.

Assim, a presente pesquisa apresenta-se como um estudo aprofundado da análise realizada em Carvalho *et al.* (2022), que demonstrou a distinção entre os vocabulários dos comentários dos vídeos educacionais e não educacionais, e a potencial utilização destes para a classificação dos vídeos. Diferentemente de Carvalho *et al.* (2022), este trabalho explora não apenas a classificação entre vídeos educacionais e não educacionais, mas considera diversas classes de ensino, demonstrando que os comentários dos vídeos podem ser utilizados de maneira eficaz para a classificá-los. Sob esta perspectiva, aponta-se



que essa classificação pode ser valiosa para a implementação de uma plataforma de recomendação de vídeos do Youtube, visando auxiliar o processo de ensino-aprendizagem de professores e alunos.

Este trabalho se organiza como segue. A seção 2 apresenta alguns dos principais trabalhos relacionados a esta pesquisa. A seção 3 apresenta a metodologia experimental adotada para a obtenção dos resultados. A seção 4 apresenta e discute os resultados obtidos. Por fim, a seção 5 apresenta as conclusões da pesquisa e perspectivas de trabalhos futuros.

2. Trabalhos Relacionados

Essa seção apresenta uma breve descrição dos trabalhos relacionados ao tema sob investigação, bem como os principais avanços da presente pesquisa em relação aos trabalhos relacionados à mesma.

Abu-El-Haija *et al.* (2016) dispõe-se de uma base de dados de aproximadamente 8 milhões de vídeos, englobando um total de 1,9 bilhões de quadros, e 500 mil horas de vídeos categorizados. Essa base de dados é utilizada para o estudo da classificação de vídeos do YouTube, com intuito de desenvolver um sistema de multi-classificação. A pesquisa é realizada em duas etapas, a saber: 1) obtenção das classes dos vídeos por meio do *Knowledge Graph entities*; 2) processamento *frame a frame* dos vídeos, e categorização dos mesmos por uma Rede Neural Convolutacional pré-treinada no *ImageNet*. O trabalho aponta sua contribuição no sentido de auxiliar o desenvolvimento de pesquisas sobre compreensão de vídeos. Aponta-se que, embora a pesquisa realize a categorização de vídeos em diversas classes, não foi encontrada uma categoria específica para vídeos educacionais.

Carvalho *et al.* (2020) aponta que, embora o YouTube realize uma classificação de seus vídeos, não é possível que o usuário realize buscas segundo suas categorias. Esse mesmo estudo desenvolve um aplicativo que acessa os dados de vídeos do YouTube e realiza buscas na plataforma pela categoria “Education”. O trabalho, então, analisa se a categorização é assertiva para identificação de vídeos educacionais. Como resultado, constatarem-se três pontos críticos relacionados ao mecanismo de busca do YouTube, a saber: 1) falta de transparência quanto à categorização de um vídeo; 2) informações incorretas quanto à categorização dos vídeos; 3) impossibilidade de se buscar vídeos por suas categorias. Desse modo, o trabalho conclui que a classificação realizada pelo YouTube não pode ser determinante para a busca por vídeos.

Carvalho *et al.* (2022) utiliza técnicas de Mineração de Texto e de Aprendizagem de Máquina com o intuito de analisar os comentários dos vídeos do Youtube e de identificar se eles possuem potencial para serem utilizados na categorização de vídeos educacionais. O trabalho apontou que existem diferenças importantes entre os vocabulários dos comentários de vídeos educacionais e não educacionais, e que o *Random Forest*, em experimentos realizados, foi capaz de categorizar os vídeos, nas classes educacional ou não educacional, com acurácia superior a 90%.

O presente trabalho utiliza-se dos resultados obtidos por Carvalho *et al.* (2022) para realizar uma investigação mais profunda acerca da classificação de vídeos educacionais do YouTube. Nesse sentido, esta pesquisa expande e aprofunda o estudo Carvalho *et al.* (2022) por aplicar técnicas de Mineração de Texto e de Aprendizagem de Máquina para classificar vídeos do YouTube em vários níveis de ensino, não se restringindo apenas às categorias educacional e não educacional. Assim, a presente proposta, além de classificar vídeos do YouTube em educacionais ou não educacionais, também classifica-os por “níveis” de ensino, a saber: Ensino Fundamental II, Médio, e



Superior; sendo que a classe Ensino Médio ainda é subdividida em suas séries (1º ano, 2º ano, e 3º ano).

A presente pesquisa, portanto, amplia e expande o estudo Carvalho *et al.* (2022), no sentido de realizar experimentos mais detalhados com o uso de mais bases de dados, a fim de verificar a potencialidade dos comentários dos vídeos do YouTube para classificá-los em diversos “níveis” de categorias de vídeos educacionais.

3. Metodologia Experimental

A Figura 1 apresenta a metodologia experimental adotada para o desenvolvimento deste trabalho. Em seguida, cada um dos passos metodológicos são explicados.

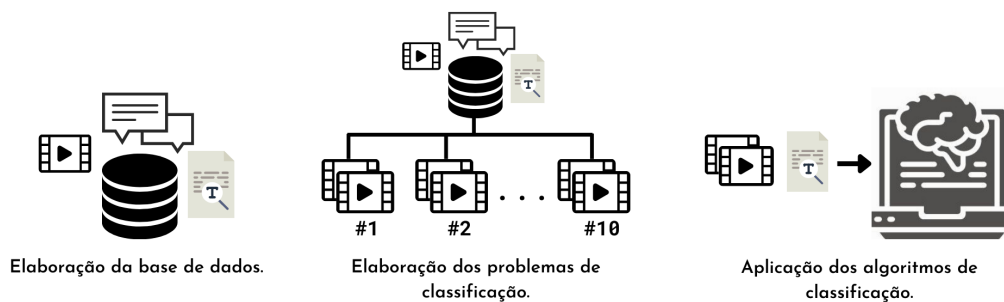


Figura 1. Metodologia experimental adotada - (Fonte: autoria própria).

3.1. Elaboração das Bases de Dados

Coletaram-se, de forma manual, diversos vídeos da plataforma YouTube para a geração inicial das bases de dados, conforme apresentado na Figura 2. No total, têm-se 8 bases de dados, a saber: Ensino Fundamental II, 1º Ano Ensino Médio, 2º Ano Ensino Médio, 3º Ano Ensino Médio, Ensino Médio, Ensino Superior, Educacional, e Não Educacional.



Figura 2. Esquema geral das bases de dados elaboradas para esse estudo - (Fonte: autoria própria).

Após a aquisição manual dos vídeos, seus comentários foram coletados utilizando-se o software *Mozdeh – Big Data Analysis 3*. O software requer uma chave de API do YouTube e os *Ids* dos vídeos que se deseja extrair os comentários. Após sua utilização, o *Mozdeh – Big Data Analysis 3* retorna um arquivo textual contendo os comentários dos vídeos, identificados pelos seus *Ids*.

De posse dos comentários dos vídeos selecionados, com auxílio da biblioteca NLTK, em *Python*, para Mineração de Texto, foi possível pré-processar os comentários, removendo-se caracteres desnecessários ao processo de classificação, tais como sinais de pontuação, *stopwords*, dentre outros. Os vocábulos restantes após o pré-processamento



foram reduzidos aos seus radicais, e, posteriormente, calcularam-se as frequências destes radicais nos comentários dos vídeos.

Cada vídeo nas bases de dados elaboradas é descrito pelo seu *Id*, seguido pelo número de vezes em que cada um dos 100 vocábulos mais frequentes nos comentários dos vídeos de uma determinada categoria figuraram nos comentários do vídeo em questão, seguido pelo número de vezes em que cada um dos 100 vocábulos mais frequentes nos comentários dos vídeos de outra categoria figuraram nos comentários do vídeo em questão, seguido, finalmente, pela categoria do vídeo (por exemplo, educacional, não educacional, 1º ano do ensino médio, etc.). Aponta-se que o número de 100 vocábulos mais frequentes para as diversas classes de vídeos foi escolhido após testes preliminares indicarem que um número maior de vocábulos não melhora a acurácia dos classificadores, bem como leva a modelos de classificação mais complexos.

A seguir, descrevem-se brevemente cada uma das 8 bases de dados geradas.

- **Ensino Fundamental II (200 vídeos):** utilizou-se como fonte de dados um canal do YouTube, o YouTube Edu. Os 200 vídeos pertencentes a base de dados Ensino Fundamental II referem-se aos vídeos das *playlists* do 6º, 7º, 8º e 9º ano do ensino fundamental, das disciplinas de Português, Matemática, Ciências, Geografia, e História. Os vídeos foram selecionados preocupando-se em manter quantidades semelhantes de vídeos de cada série e de cada matéria, para não se criar uma base de dados tendenciosa e/ou desbalanceada.
- **1º Ano Ensino Médio (200 vídeos):** utilizou-se também o YouTube Edu como fonte de dados. Os 200 vídeos pertencentes a base de dados referem-se aos vídeos da *playlist* do 1º ano do ensino médio, das disciplinas de Português, Matemática, Biologia, Química, Física, Geografia, História, e Literatura. Os vídeos foram selecionados preocupando-se em manter quantidades semelhantes de vídeos de cada série e de cada matéria, para não se criar uma base de dados tendenciosa e/ou desbalanceada. Não foram utilizados os vídeos da disciplina de Inglês, pois os comentários destes vídeos apresentavam muitos termos em inglês, o que traria dificuldades para a etapa de identificação dos vocábulos mais frequentes, visto que esta etapa foi realizada de maneira direcionada a vocábulos em língua portuguesa.
- **2º Ano Ensino Médio (200 vídeos):** elaborada de maneira semelhante à base de dados 1º Ano Ensino Médio, com a distinção de que os vídeos selecionados referem-se a *playlist* 2º Ano do Ensino Médio do canal YouTube Edu. Esta base também é composta por 200 vídeos.
- **3º Ano Ensino Médio (200 vídeos):** elaborada de maneira semelhante à base de dados 1º Ano Ensino Médio, com a distinção de que os vídeos selecionados referem-se a *playlist* 3º Ano do Ensino Médio do canal YouTube Edu. Esta base também é composta por 200 vídeos.
- **Ensino Médio (600 vídeos):** os 600 vídeos que compõem essa base de dados são os vídeos das bases 1º Ano Ensino Médio, 2º Ano Ensino Médio, e 3º Ano Ensino Médio.
- **Ensino Superior (200 vídeos):** utilizou-se como fonte de dados *playlists* de vídeos de disciplinas do Ensino Superior, bem como vídeos presentes em canais de Instituições de Ensino Superior. Os 200 vídeos que compõem essa base de dados foram coletados das mais diversas áreas, com objetivo de se evitar a geração de uma base de dados tendenciosa a cursos ou a assuntos específicos.
- **Educacional (1000 vídeos):** os 1000 vídeos que compõem essa base de dados são os vídeos das bases Ensino Fundamental II, Ensino Médio, e Ensino Superior.



- **Não Educacional (200 vídeos):** os 200 vídeos que compõem essa base de dados referem-se a vídeos que não abordam nenhum conteúdo de intuito estritamente educativo, como por exemplo, *vlogs*, clipes musicais, etc. Os vídeos dessa base de dados foram selecionados sobre temas diversos por meio do mecanismo de busca do YouTube.

É importante ressaltar que a escolha em se utilizar os vídeos retornados via plataforma YouTube Edu baseia-se na dificuldade em se determinar em qual ano de Ensino o conteúdo de um vídeo educacional se enquadra, visto que, normalmente, os títulos dos vídeos só apresentam o tema apresentado, e, alguns destes temas, são abordados em diferentes profundidades em mais de um ano de Ensino. Sob este aspecto, para a construção das bases de dados sem a utilização do YouTube Edu, seria necessário o auxílio de um grupo de professores da Educação Básica, para analisar e classificar os vídeos em suas diversas categorias de ensino.

3.2. Elaboração dos Problemas de Classificação

De posse das bases de dados previamente descritas, elaboram-se, no total, 10 problemas de classificação. Tais problemas foram gerados visando-se uma maior diversificação, a fim de se verificar a capacidade dos classificadores em categorizar vídeos em em diversos e distintos problemas no âmbito educacional. A Tabela 1 apresenta os 10 problemas de classificação abordados. Cada um dos 10 problemas, na coluna “Problema”, utiliza parte ou a totalidade das bases de dados geradas, conforme apresentado numericamente em cada uma das outras colunas da tabela.

Ressalta-se que cada problema de classificação trabalha com 200 vídeos de cada classe (por exemplo, educacional, não educacional, etc.) e, desse modo, para as bases de dados que ultrapassam esse número de vídeos, apenas uma fração delas é utilizada para a geração do problema. Neste sentido, o problema #E-NE, a exemplo, utiliza os 200 vídeos da base de dados Não Educacional, para a representação da classe não educacional, e 200 vídeos (40 do Ensino Fundamental II, 40 do 1º Ano Ensino Médio, 40 do 2º Ano Ensino, 40 do 3º Ano Ensino Médio, e 40 do Ensino Superior) selecionados dos 1000 vídeos que compõem a base de dados Educacional, para a representação da classe educacional. A geração dos demais problemas de classificação deve ser entendida de maneira equivalente ao exemplo anterior.

Ao se selecionar os vídeos das bases para comporem os problemas, preocupou-se em equilibrar o número de vídeos de cada disciplina, a fim de manter o problema de classificação não tendencioso, uma vez que um problema com vídeos majoritariamente de uma disciplina poderia implicar em um vocabulário nos comentários mais propenso a ela. A seguir, apresenta-se cada um dos problemas sob investigação.

- **Problema #E-NE:** as classes deste problema são, **Educacional (E)**, e **Não Educacional (NE)**.
- **Problema #M-F-S:** as classes desse problema são, **Ensino Médio (M)**, **Ensino Fundamental II (F)**, e **Ensino Superior (S)**.
- **Problema #M-F:** as classes desse problema são, **Ensino Médio (M)**, e **Ensino Fundamental II (F)**.
- **Problema #M-S:** as classes desse problema são, **Ensino Médio (M)**, e **Ensino Superior (S)**.
- **Problema #M1-F:** as classes desse problema são, **1º Ano do Ensino Médio (M1)**, e **Ensino Fundamental II (F)**.
- **Problema #M2-F:** as classes desse problema são, **2º Ano do Ensino Médio (M2)**, e **Ensino Fundamental II (F)**.



Tabela 1. Relação entre os problemas gerados e bases de dados utilizadas.

Problema	Educacional					
	Não Educacional	Ens. Fund. II	Ens. Médio			Ens. Superior
			1º Ano Ens. Médio	2º Ano Ens. Médio	3º Ano Ens. Médio	
#E-NE	200	40	40	40	40	40
#M-F-S		200	66	67	67	200
#M-F		200	66	67	67	
#M-S			66	67	67	200
#M1-F		200	200			
#M2-F		200		200		
#M3-F		200			200	
#M1-S			200			200
#M2-S				200		200
#M3-S					200	200

- **Problema #M3-F:** as classes desse problema são, **3º Ano do Ensino Médio (M3)**, e **Ensino Fundamental II (F)**.
- **Problema #M1-S:** as classes desse problema são, **1º Ano do Ensino Médio (M1)**, e **Ensino Superior (S)**.
- **Problema #M2-S:** as classes desse problema são, **2º Ano do Ensino Médio (M2)**, e **Ensino Superior (S)**.
- **Problema #M3-S:** as classes desse problema são, **3º Ano do Ensino Médio (M3)**, e **Ensino Superior (S)**.

3.3. Algoritmos de Classificação e Métrica de Avaliação

Utilizou-se o Weka (Frank E. 2016) para a resolução dos 10 problemas de classificação propostos. Especificamente, foram utilizados os classificadores J48 (*Decision Tree Classifier*), JRIP (*Repeated Incremental Pruning to Produce Error Reduction*), e LMT (*Logistic Model Tree*) implementados no referido *framework*. A escolha destes classificadores se justifica uma vez que os mesmos geram modelos de classificação inteligíveis, baseados em regras, que podem ser facilmente interpretados por seres humanos.

Adotou-se a configuração padrão (*default*) de cada classificador, conforme parametrizado no *framework* Weka. Uma vez que as bases de dados são balanceadas, i.e., possuem o mesmo número de instâncias (exemplos) de cada classe, utilizou-se a acurácia como métrica de avaliação dos classificadores. Ademais, adotou-se a técnica de validação cruzada de 10 *folds* (*10-fold cross-validation*) para a obtenção das acurácias e dos modelos de classificação, uma vez que esta metodologia é uma das mais recomendadas e utilizadas para esta finalidade (Berrar 2019; Mitchell *et al.* 1997).

4. Discussões e Resultados

A Tabela 2 apresenta as acurácias obtidas pelos classificadores em cada um dos dez problemas estudados. Adicionalmente, a tabela fornece a acurácia média de cada classificador considerando os dez problemas, bem como a acurácia média por problema, levando em consideração os três classificadores empregados. Os valores mais elevados para cada problema e para as médias calculadas estão destacados em negrito.

De acordo com a Tabela 2, observa-se que o LMT apresentou desempenho superior em 9 dos 10 problemas investigados. Tal resultado está associado a



Tabela 2. Resultados para os problemas de classificação investigados.

Problema	LMT	JRIP	J48	Média
#E-NE	87,2%	85,2%	81,5%	84,6%
#M-F-S	59,8%	47%	46,8%	51,2%
#M-F	63,5%	56,5%	58,3%	59,4%
#M-S	73,5%	71,5%	71,5%	72,2%
#M1-F	71%	59%	60%	63,3%
#M2-F	56%	54,5%	54%	54,8%
#M3-F	63,3%	64,8%	63,8%	63,9%
#M1-S	74,3%	65,4%	65,2%	68,3%
#M2-S	71,1%	68,2%	68,7%	69,3%
#M3-S	73,1%	70,1%	63,8%	69%
Média	69,3%	64,2%	63,4%	65,6%

características específicas deste classificador, pois, embora todos os modelos investigados utilizem regras como forma de classificação, o modelo gerado pelo LMT é baseado em Regressão Logística em conjunto com Árvores de Decisão. Nesse sentido, os modelos gerados pelo LMT, embora mais acurados, são também mais complexos.

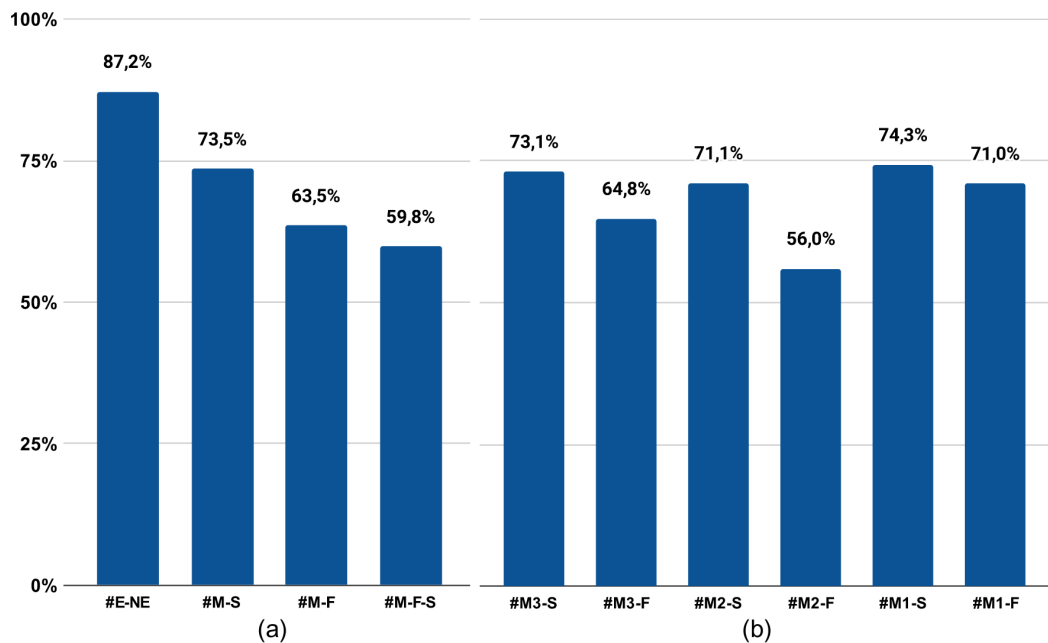


Figura 3. Acurácia do melhor classificador para cada problema investigado - (Fonte: autoria própria).

A Figura 3 apresenta a acurácia do melhor classificador para cada problema investigado. O gráfico (a) considera os problemas envolvendo o Ensino Médio como um todo. O gráfico (b), por sua vez, refina a análise do Ensino Médio, em níveis específicos (1º, 2º e 3º anos).

Observando o gráfico (a) na Figura 3, é notável a acurácia para o problema #E-NE, de 87,2%. Este resultado pode estar ligado à clara distinção entre os vocabulários utilizados nos comentários de vídeos Educacionais e não Educacionais.



Adentrando nas classificações por níveis de ensino, e considerando-se as classes Ensino Médio e Ensino Superior, problema #M-S, observa-se uma queda no desempenho do melhor classificador, que obtém 73,5% de acurácia. Sob essa mesma lógica, ao se substituir a classe Ensino Superior pela classe Ensino Fundamental II, problema #M-F, o desempenho do melhor classificador diminui para 63,5% de acurácia, evidenciando a maior complexidade na classificação da classe Ensino Fundamental II, em relação à classe Ensino Superior. Ademais, no problema #M-F-S, que envolve três classes, ocorre novamente uma piora no desempenho do melhor classificador, que obtém 59,8% de acurácia. Essa queda adicional na acurácia reflete as dificuldades encontradas pelos classificadores ao lidarem com um problema contendo três classes.

No gráfico (b) da Figura 3, os resultados de acurácia para problemas #M3-S, #M2-S, e #M1-S foram de 73,1%, 71,1%, e 74,3%, respectivamente. Esses resultados estão próximos da acurácia obtida na instância #M-S (gráfico (a)), que é de 73,5%; problema que considera, inclusive, todos os anos do Ensino Médio. Dentre os três problemas analisados, nota-se que a menor acurácia, de 71,1%, foi obtida na instância #M2-S. Tal fato indica que os comentários dos vídeos do 2º Ano do Ensino Médio apresentam um vocabulário mais desafiador para a classificação, em comparação aos vocabulários dos comentários dos vídeos do 1º e do 3º ano. É relevante destacar que as acurácias para estes três problemas mantiveram-se acima de 70%.

Analisando os problemas #M3-F, #M2-F, e #M1-F, no gráfico (b), nos quais há variações nos anos do Ensino Médio, mas mantém-se a classe Ensino Fundamental II, observa-se que o #M2-F apresenta a menor acurácia, de apenas 56%. Essa combinação específica, de Ensino Médio II e Ensino Fundamental II, parece ser a mais desafiadora. Nesse sentido, nota-se que o problema #M2-F compartilha as mesmas bases de dados com os problemas #M-F e #M-F-S, que envolvem o 2º Ano do Ensino Médio e o Ensino Fundamental II, sugerindo uma dificuldade de classificação da combinação entre a classe Ensino Fundamental e 2º ano do Ensino Médio (vide as menores acurácias do gráfico (a)).

Ao se comparar o resultado obtido no problema #M-F ao #M2-F, observa-se uma queda significativa na acurácia, de 63,5% para 56%. Por outro lado, ao se analisar o resultado do problema #M-F ao do #M3-F, a acurácia se mantém praticamente a mesma, com pequena elevação de 63,5% para 64,8%. Além disso, o problema #M1-F apresenta acurácia 71%; o melhor desempenho dentre as bases de dados que utilizam anos isolados do Ensino Médio (#M1-F (71%), #M2-F (56%), #M3-F (64,8%)).

Ao se considerar a classe Ensino Superior, o desempenho dos classificadores nesses problemas apresentam uma acurácia mais elevada em relação aos problemas relacionados à classe Ensino Fundamental II. Nesse sentido, a acurácia obtida no problema #M1-S foi de 74,3%, e no #M1-F, de 71%. Para o caso do problema #M2-S, obteve-se uma acurácia de 71,1%, e no #M2-F de 56%. Esse mesmo padrão é observado no problema #M3-S, no qual obteve-se uma acurácia de 73,1%, e no #M3-F de 64,8%. Assim, aponta-se que a classe Ensino Fundamental II é mais difícil de ser diferenciada pelos classificadores utilizados do que classe Ensino Superior, para todas as combinações de classes que envolvem o Ensino Médio (1º, 2º e 3º anos).

De maneira geral, os problemas #M-F-S, #M-F, #M1-F, #M2-F e #M3-F apresentam os valores de acurácias mais baixos, sendo que apenas no problema #M1-F obteve-se acurácia superior a 70%. Todos esses problemas têm em comum o fato de incluírem a classe Ensino Fundamental II em suas bases de dados. Tal ponto sugere que a presença dessa classe traz dificuldades ao processo de classificação. É importante destacar que esses problemas utilizam comentários de estudantes de faixa etária mais baixa, em comparação as demais fases de ensino abordadas nesse estudo. Por outro lado,



os problemas, #E-NE, #M-S, #M1-S, #M2-S e #M3-S apresentam, todos, uma acurácia superior a 70%. Com exceção do problema #E-NE, esses problemas não utilizam a base de dados Ensino Fundamental II e, portanto, trabalham apenas com os comentários de alunos das faixas etárias mais altas desse estudo.

A baixa faixa etária dos estudantes pode contribuir para um maior número de erros ortográficos, o que dificulta a construção do vocabulário de palavras mais frequentes. Além disso, os comentários dessa faixa etária podem conter termos desconexos aos assuntos abordados nos vídeos, o que também afeta a qualidade da base de dados. Por outro lado, obteve-se um valor de acurácia de destaque no problema #E-NE, 87,2%, mesmo tal instância considerando todas as faixas etárias desse estudo. Tal ponto pode ser justificado pelo fato de que esse problema trabalha com todas as bases de dados, e realiza uma classificação mais “geral”, dividindo os comentários nas categorias Educacional e não Educacional. Dessa forma, o impacto dos vídeos com comentários de alunos de baixa faixa etária é reduzido.

Destaca-se que em 6 dos 10 problemas abordados, alcançou-se acurácia superior a 70%, e em somente 2 problemas, a acurácia obtida foi abaixo de 60%, sendo que nesses, a classe Ensino Fundamental II era sempre uma das classes trabalhadas.

5. Considerações finais e Trabalhos Futuros

Esse trabalho utilizou Técnicas de Mineração de Texto e de Aprendizagem de Máquina para analisar os comentários de vídeos do YouTube, e apresentou uma metodologia eficiente para classifica-los em diversas “categorias” educacionais. Realizaram-se experimentos computacionais com o objetivo de verificar a potencialidade dos comentários dos vídeos do YouTube com objetivo de classifica-los. Aponta-se a potencialidade da metodologia adotada para a classificação de vídeos do YouTube em diversas “categorias” educacionais.

Como perspectivas de trabalhos futuros, pretende-se explorar a classificação de vídeos de conteúdo do Ensino Médio, em disciplinas e matrizes de ensino, de acordo com o ENEM, visto que as classificações dos vídeos de Ensino Médio apresentaram resultados promissores. Outro tópico a ser investigado é a utilização de outros metadados dos vídeos com objetivo de auxiliar a classificação dos mesmos, bem como a parametrização específica dos algoritmos de classificação para cada problema investigado.

Referências

Abu-El-Haija, S. *et al.* Youtube-8m: A large-scale video classification benchmark. **CoRR**, abs/1609.08675, 2016. Disponível em: <http://arxiv.org/abs/1609.08675>.

Audino, D. F. **Objetos de aprendizagem hipermídia aplicado à cartografia escolar no sexto ano do ensino fundamental em geografia**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Santa Catarina (UFSC), 2012.

Berrar, D. Cross-validation. In: Ranganathan, S.; Gribskov, M.; Nakai, K.; Schönbach, C. (Ed.). **Encyclopedia of Bioinformatics and Computational Biology**. Oxford, UK: Academic Press, 2019. p. 542–545. ISBN 978-0-12-811432-2.

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://sol.sbc.org.br/journals/index.php/rbie/article/view/2455>.



Carvalho, H. C. F. B.; Pitangui, C. G.; Assis, L. P.; Andrade, A. V. 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. GS Search. Disponível em: <https://esud2020.ciar.ufrgs.br/wp-content/anais-esud/210418.pdf>.

Frank E., H. M. A. . W. 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.

Mitchell, T. M. *et al.* **Machine learning**. New York: McGraw-Hill New York, 1997.

Wiley, D. A. Learning object design and sequencing theory. **Tese (Doutorado) — Brigham Young University**, 2000.

YouTube. **YouTube: Características do Produto**. 2019. Acesso em: 31 de Outubro de 2022. Disponível em: <https://www.youtube.com/howyoutubeworks/product-features/search/>.