Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 20/Jan/2022; 1st round notif.: 14/Mar/2022; New version: 2/Apr/2022; 2nd round notif.: 15/Jun/2022;
Camera ready: 9/Aug/2022; Edition review: 12/Sep/2022; Available online: 30/Sep/2022; Published: 30/Sep/2022;

Classificação automática de vídeos educacionais por meio de comentários apoiada por técnicas de Aprendizado de Máquina: uma análise experimental utilizando o Youtube

Title: Automatic classification of educational videos supported by comment-based machine learning techniques: an experimental analysis using Youtube

Henrique Carlos Fonte Boa Carvalho
Universidade Federal de Uberlândia
ORCID: 0000-0002-6109-8000
henriquefbc@gmail.com

Fabiano Azevedo Dorça
Universidade Federal de Uberlândia
ORCID: 0000-0003-3281-0246
fabianodorca@gmail.com

Cristiano Grijó Pitangui
Universidade Federal de São João Del-Rei
ORCID: 0000-0002-3961-2042
pitangui.cristiano@gmail.com

Luciana Pereira de Assis
Universidade Federal dos Vales do Jequitinhonha e Mucuri
ORCID: 0000-0002-7891-7172
lupassis@gmail.com

Alessandro Vivas Andrade
Universidade Federal dos Vales do Jequitinhonha e Mucuri
ORCID: 0000-0003-4713-5159
alessandro.vivas@gmail.com

Eduardo Augusto Costa Trindade
Universidade Federal dos Vales do Jequitinhonha e Mucuri
ORCID: 0000-0002-5185-6605
eduardoactrindade@hotmail.com

Resumo

O avanço tecnológico permite que novos conteúdos sejam criados e disponibilizados via Web a cada minuto, propiciando grande avanço em diversas áreas. Entretanto, essa disponibilidade também aponta desvantagens no âmbito Educacional. Destaca-se que o excesso de materiais ofertados dificultam o processo de ensino-aprendizagem devido ao demasiado tempo utilizado em busca de conteúdos que atendam às necessidades dos usuários. Nesse sentido, novos métodos para identificar conteúdos educacionais, em vídeos, por exemplo, precisam ser desenvolvidos. Percebe-se, sob essa perspectiva, diferenças significativas nos comentários fornecidos por usuários em vídeos educacionais, indicando assim, potencial para que estes sejam utilizados para a seleção de vídeos educativos. Neste contexto, esse trabalho realiza a análise e coleta de comentários de 500 vídeos da plataforma Youtube, sendo 250 educacionais e 250 não educacionais, e utiliza técnicas de Mineração de Textos e Aprendizado de Máquina para desenvolver um modelo de classificação que, baseado nos vocábulos mais frequentes dos comentários dos vídeos, os categorize em educacionais ou não educacionais. Com isto, provê-se um mecanismo que filtra os vídeos de acordo com a sua classe e retorna ao usuário apenas vídeos de conteúdo educacional. Resultados obtidos demonstram que é possível categorizar vídeos educacionais e não educacionais com acurácia de até 91,30%, ao se utilizar o classificador Random Forest. Ademais, devido aos resultados promissores, desenvolveu-se a API SysVidEduc, que utiliza os comentários dos usuários nos vídeos do Youtube e os classifica automaticamente em educacionais ou não educacionais.
Palavras-chave: *Mineração de Texto; Aprendizado de Máquina; Classificação; Comentários; Vídeos; Youtube.*

Abstract

Technological advances allow new content to be created and be available via Web every minute, providing great progress in several areas. However, this availability also brings drawbacks in the Educational field. It is noteworthy that the excess of materials/content makes teaching-learning process difficult due to the high time spent in searching for content that meets the needs of the users. In this sense, new methods to identify educational content, in videos, for example, need to be developed. From this perspective, it can be seen that significant differences are identified in the comments provided by users on educational videos, thus indicating the potential for using them in the process of

Cite as: Carvalho, H. C. F. B., Dorça, F. A., Pitangui, C. G., Assis, L. P. de, 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, 30, 419-448. DOI: 10.5753/rbie.2022.2455.

selecting these types of videos. In this context, the present work analyzes and collects comments from 500 videos of the Youtube platform, being 250 educational and 250 non-educational, and uses Text Mining and Machine Learning techniques to develop a classification model that, based on the most frequent words of comments on videos, categorize them as educational or non-educational. Thus, we provide a mechanism that filters videos according to their class and returns to the user only videos with educational content. Results demonstrate that it is possible to classify educational and non-educational videos with an accuracy rate of 91.30%, when using the Random Forest classifier. Furthermore, due to the promising results, we developed SysVidEduc, an API that uses the comments from Youtube videos and automatically classifies them as educational or non-educational.
Keywords: *Text Mining; Machine Learning; Classification; Comments; Videos; Youtube.*

1 Introdução

O avanço da tecnologia possibilita que a sociedade seja cada vez mais conectada, com acesso a diversas informações, interação social, entre outros (do Nascimento, Barreto, Primo, Gusmão, & Oliveira, 2017). O aumento de dados nas redes propiciou um enorme avanço para a Educação, permitindo que novos conteúdos sejam criados e compartilhados a cada minuto. No entanto, este vasto acervo de conteúdo apresenta vantagens e desvantagens, 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).

Sob o ponto de vista educacional, o *Youtube* pode ser considerado um repositório de Objetos de Aprendizagem (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, alguns problemas relacionados ao mecanismo de busca do *Youtube* podem ser identificados. Carvalho, Pitangui, Trindade, Assis, and Andrade (2020) apontam, de forma geral, três problemas principais. Tais questões são apresentadas a seguir.

O primeiro problema diz respeito aos resultados retornados pela plataforma a partir de uma busca, pois não é possível filtrar os resultados por conteúdo educacional. Desta forma, são retornados, em muitos casos, um grande número de vídeos, sendo vários destes de baixa qualidade e/ou não muito relacionados à busca realizada sob a ótica educacional. Neste contexto, podem

ser retornados muitos vídeos promocionais, essencialmente comerciais ou vídeos pessoais, o que dificulta a utilização desta plataforma para fins de aprendizagem. Neste caso, o estudante perde muito tempo para averiguar se o vídeo é educacional e se pode ser utilizado em seus estudos.

A segunda dificuldade se relaciona à maneira pela qual a plataforma "interpreta" a expressão de pesquisa. De forma geral, é realizada a comparação entre os termos da busca com os dados dos vídeos, a saber: título, descrição e *tags*. Desta forma, os resultados da pesquisa são vídeos que contenham o(s) termo(s) (da expressão de pesquisa) em seus dados. No caso da expressão busca conter mais de um termo, a plataforma retorna os vídeos que contemplem, em seus dados, todos os termos da expressão. Por exemplo, os vídeos apresentados para a expressão de pesquisa "Futebol", serão aqueles que, em seu título, descrição, ou *tags*, incluam o termo "Futebol". De forma semelhante, caso a busca seja realizada pela expressão "Futebol educação", serão apresentados vídeos que contenham ambos termos em seus dados. No caso de expressões de pesquisa com mais de um termo, o mecanismo de busca da plataforma pode ser considerado bastante restritivo, uma vez que é capaz de desconsiderar vídeos de interesse do usuário.

Por último, outro problema encontrado se relaciona à impossibilidade de se realizar buscas por vídeos segundo suas categorias. De fato, o *Youtube* categoriza seus vídeos, porém, não disponibiliza meios de se realizar pesquisas por elas, e a maioria dos usuários não tem conhecimento sobre essa categorização. Sob este aspecto, Carvalho, Pitangui, Trindade, Assis, and Andrade (2020), apresentam que, apesar de diversos vídeos serem categorizados como *education*, tais classificações são realizadas erroneamente, confirmando, desta maneira, que as categorias fornecidas pela plataforma não devem ser consideradas como determinantes no processo de busca por um vídeo.

Considerando o problema de categorização errônea de vídeos no *Youtube*, Carvalho, Pitangui, Trindade, Assis, Andrade, and de Souza (2020) identificaram diferenças significativas entre os vocabulários (dos usuários) em comentários de vídeos educacionais e não educacionais. O trabalho apontou que os comentários publicados por usuários possuem potencial para serem utilizados como método de categorização dos vídeos da plataforma.

Neste contexto, o presente trabalho propõe uma nova maneira de categorizar os vídeos do *Youtube* em educacionais ou não educacionais, por meio do uso de técnicas de Aprendizado de Máquina que processam os comentários fornecidos pelos usuários. Dessarte, pretende-se fornecer um mecanismo que filtre automaticamente os conteúdos, apresentando ao estudante apenas vídeos com conteúdos realmente educacionais.

A plataforma *Youtube* foi selecionada para a realização dos experimentos por possuir um extenso acervo de vídeos que pode utilizado com um viés educacional, além disso, possibilita que seus usuários expressem suas opiniões a respeito dos vídeos por meio de comentários. No entanto, é importante destacar que a abordagem proposta pode ser expandida e utilizada em outras plataformas de repositórios de vídeos.

Para o desenvolvimento da proposta, inicialmente coletou-se manualmente um total 500 vídeos (250 educacionais e 250 não educacionais) juntamente com seus comentários. Em seguida, os comentários coletados foram pré-processados com o objetivo de remover informações não relevantes ao objetivo proposto. Dessa forma, removeram-se dos comentários os caracteres especiais, números, e *stopwords*, obtendo, por fim, apenas os radicais dos vocábulos, como por exemplo, "profes" (advindo de ambos os vocábulos "professor" e "professora").

Posteriormente, realizou-se a análise da frequência dos vocábulos, e identificou-se elevada diferença nos vocábulos empregados nas duas classes (educacional e não educacional) consideradas. Após essa análise, foram geradas oito bases de dados com diferentes quantidades de atributos, i.e., cada base de dados possui como atributos um determinado número de vocábulos mais frequentes. Os classificadores JRIP, PART, J48, Random Forest e GenClust++ foram utilizados em cada uma das oito bases de dados com objetivo de classificar os vídeos em educacionais ou não educacionais. Por fim, desenvolveu-se a API (*Application Programming Interface*) SysVidEduc, que realiza a classificação automática de vídeos educacionais do *Youtube* por meio de comentários.

Durante a coleta dos comentários, observou-se que alguns vídeos não apresentam ou desabilitam a função de postar comentários. Dessa forma, esses vídeos não foram selecionados para a geração das bases de dados. Aponta-se que, para estes vídeos, outras abordagens de classificação devem ser propostas. Neste sentido, uma possível solução seria utilizar a geração de legendas automáticas da plataforma e utilizá-las para a classificação dos vídeos.

De forma geral, os resultados obtidos neste trabalho demonstram que as técnicas de Aprendizado de Máquina supervisionadas são mais assertivas quando comparadas ao GenClust++, um algoritmo de Aprendizado de Máquina não supervisionado. Optou-se por utilizar o GenClust++ uma vez que este esta técnica, apresenta, de forma geral, resultados superiores quando analisada em relação a outras técnicas de clusterização. Isto se dá, em parte, pois o GenClust++ utiliza um Algoritmo Genético em conjunto com K-means no processo de identificação de clusteres.

De forma breve, a técnica de Aprendizado de Máquina supervisionada que mais se destacou foi o Random Forest, que obteve acurácia mínima de 83,02% (na base de dados com 8 vocábulos) e máxima de 91,30% (na base de dados com 200 vocábulos). Para estas bases de dados, os valores de acurácia obtidos pelo GenClust++ foram de 50,72% e 60,46%, respectivamente.

A API SysVidEduc foi desenvolvida motivada pelos resultados promissores obtidos pelos algoritmos de Aprendizado de Máquina utilizados. A API desenvolvida recebe uma *string*, que pode ser uma expressão de busca ou o *id* de um vídeo, busca os metadados dos vídeos, processa seus comentários e classifica-os utilizando o Random Forest, técnica obteve os melhores resultados dentre os algoritmos avaliados. Por receber os dados via *Web*, o SysVidEduc pode ser facilmente empregado em Ambientes Virtuais de Aprendizagem, propiciando agilidade na escolha dos materiais e transparência para docentes e discentes. Ademais, a API proposta já realiza toda a integração com a API de dados do *Youtube*, processando os comentários e classificando os vídeos sem a necessidade de se instalar ou processar algo diretamente no Ambiente Virtual de Aprendizagem.

O presente trabalho se organiza como segue. A seção 2 aborda os principais conceitos utilizados nesta pesquisa. A seção 3 apresenta os principais trabalhos relacionados ao tema deste trabalho. A seção 4 descreve a metodologia experimental adotada na pesquisa. A seção 5 discute os resultados obtidos. Por fim, a seção 6 apresenta as considerações finais e apontamentos a trabalhos futuros.

2 Referencial Teórico

Essa seção aborda os principais conceitos relacionados a esta pesquisa.

2.1 Aprendizado de Máquina e Mineração de Dados e Texto

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 Inteligencia 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 verdadeira capacidade do classificador em reconhecer as classes apresentadas. Um dos métodos mais utilizados e recomendados para estimar a verdadeira 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). Mineração de Dados pode ser definido como o processo de descoberta de pa-

drões 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 et al. (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, Ilamathi, Nithya, et al., 2015; Sukanya & Biruntha, 2012). Tais etapas são brevemente descritas a seguir.

A etapa de pré-processamento consiste no tratamento do texto antes de se realizar a análise e a aplicação das técnicas propriamente ditas. Essa etapa consiste em padronizar o texto, removendo palavras irrelevantes (como *stop words*, caracteres especiais e numéricos), e aglomerar termos similares, i.e, converter os caracteres para minusculo, corrigir erros ortográficos, etc. (Hickman, Thapa, Tay, Cao, & Srinivasan, 2022). Essa etapa é de grande importância para o processo como um todo, pois melhora a qualidade dos dados a serem analisados. Pode-se pensar que esta etapa "limpa" a base de dados de "ruídos", tais como erros ortográficos, caracteres especiais, entre outros (HaCohen-Kerner, Miller, & Yigal, 2020).

A etapa de aplicação de técnicas de mineração consiste em utilizar algoritmos para processar 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).

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 Algoritmos de Aprendizado de Máquina

Uma das forma de se categorizar os mais diversos algoritmos de Aprendizado de Máquina se faz pela maneira pela qual eles representam o conhecimento. Neste sentido, a representação do conhecimento pode ser realizada por meio de Árvores de Decisão, Regras, ou Clusteres (Witten et al., 2011). A Tabela 1 apresenta os algoritmos utilizados nessa pesquisa categorizados quanto ao tipo de representação do conhecimento e ao tipo de aprendizado.

Tabela 1: Algoritmos utilizados nesta pesquisa.

Algoritmo	Representação do Conhecimento	Tipo de Aprendizado
JRIP	Regras	Supervisionado
PART	Regras	Supervisionado
J48	Árvore de Decisão	Supervisionado
Random Forest	Árvore de Decisão	Supervisionado
GenClust++	Clusteres	Não supervisionado

As Árvores de Decisão são umas das formas mais simples e mais bem-sucedidas de se classificar dados. 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). Cada nó interno da árvore corresponde a um teste do valor de um dos atributos de entrada. Cada nó folha representa uma classe para uma determinada entrada (Kesavaraj & Sukumaran, 2013; Allahyari et al., 2017).

Os algoritmos baseados em Regras podem ser definidos basicamente como uma simples condição SE-ENTÃO, onde tem-se uma condição, também chamada de antecedente e uma predição ou conclusão. Basicamente as decisões baseadas em Regras funcionam da seguinte forma: SE

a condição X é atendida ENTÃO classifica-se essa sequência de características em determinada classe (Witten et al., 2011).

Por sua vez, os algoritmos baseados em Clusteres 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).

A seguir, apresenta-se uma breve descrição dos algoritmos utilizados na presente proposta.

- O JRip (RIPPER - *Repeated Incremental Pruning to Produce Error Reduction)* gera regras de classificação tratando inicialmete dos exemplos de uma classe específica. Após a geração de regras para uma classe, ele passa para incrementalmente para cada classe até que todas as classes tenham sido tratadas (Rajput, Aharwal, Dubey, Saxena, & Raghuvanshi, 2011) (Cohen, 1995).
- O J48 É um algoritmo que utiliza a estratégia de dividir para conquistar para gerar uma árvore de decisão. Cada nó da árvore está associado a um conjunto de exemplos. No início, apenas o nó raiz está presente e associado ao conjunto de treinamento. Para a geração de um novo nó, o algoritmo de dividir para conquistar é executado, objetivando identificar a melhor escolha local (Ruggieri, 2002).
- O Part baseia-se na árvore de decisão. A cada iteração, ele constrói uma árvore parcial e transforma o caminho até o melhor nó em uma regra de classificação (Frank & Witten, 1998).
- O Random Forest é um modelo de classificação em conjunto baseado em árvores de decisão, ou seja, é um algoritmo que constrói diversas árvores e a classificação de um exemplo é dada pela votação deste conjunto de classificadores (Breiman, 2001).
- O algoritmo GenClust++ é uma combinação entre o K-Means e um Algoritmo Genético com um novo arranjo de operadores genéticos para realizar a clusterização. Esta técnica possui a capacidade de identificar como soluções iniciais um conjunto de cromossomos de alta qualidade (Islam, Estivill-Castro, Rahman, & Bossomaier, 2018).

3 Trabalhos Relacionados

Esta seção aborda os principais trabalhos relacionados a esta pesquisa. Destaca-se que poucos trabalhos foram identificados referentes a categorização de vídeos educacionais do *Youtube*. Neste sentido, esta seção relaciona os trabalhos que abordam os temas de Recomendação de OA utilizando a *Wikipédia* e o *Youtube*, e trabalhos que utilizam técnicas de Aprendizado de Máquina com foco no *Youtube*. 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 2 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.

Tabela 2: Comparação entre os trabalhos relacionados e a abordagem proposta.

Autores	Plataforma?	OA?	Class.?	Característica?	Com. Edu.?
Menolli, Malucelli, and Reinehr (2011)	Wikipedia	Sim	Sim	Textos	Não
Abu-El-Haija et al. (2016)	Youtube	Não	Sim	Frames / imagens	Não
Júnior and 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 and Duque (2019)	Youtube	Não	Não	Comentários	Não
Carvalho, Pitangui, Assis, and Andrade (2020)	Youtube	Sim	Não	Categoria do vídeo	Não
Carvalho, Pitangui, Trindade, Assis, and Andrade (2020)	Youtube	Sim	Não	Categoria do vídeo	Não
Amanda and Negara (2020)	Youtube	Não	Sim	Títulos e descrição	Não
Trindade et al. (2020)	Youtube	Sim	Não	Títulos, descrição e tags	Não
Carvalho, Pitangui, Trindade, Assis, Andrade, and de Souza (2020)	Youtube	Sim	Não	Comentários	Sim
Zheng, Xue, Sun, and Zhu (2021)	Youtube	Não	Não	Comentários	Não
Abordagem proposta	Youtube	Sim	Sim	Comentários	Sim

De forma geral, observa-se que apenas a presente proposta e o trabalho desenvolvido por Carvalho, Pitangui, Trindade, Assis, Andrade, and de Souza (2020) utilizam os comentários e realizam a análise de comentários educacionais, mas que apenas o trabalho atual realiza a classificação de vídeos do *Youtube* por meio de técnicas de Aprendizado de Máquina baseado-se em comentários.

Menolli et al. (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 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 and 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 uti-

lizados 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 and 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, Pitangui, Assis, and Andrade (2020) 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, Pitangui, Trindade, Assis, and Andrade (2020) 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 and 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 meta-

dados 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, Pitangui, Trindade, Assis, Andrade, and de Souza (2020) 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 em 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 contribui para estabelecer um feedback em tempo real entre a sociedade e a autoridade de saúde publica, 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 cujo o propósito era identificar e auxiliar o processo de escolha de vídeos educacionais da plataforma *Youtube*. Inicialmente, o trabalho, (Abu-El-Haija et al., 2016), foi identificado quanto a 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 se identificar outros meios de classificação dos vídeos educacionais do *Youtube*. Posteriormente, identificou-se o trabalho (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 (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*.

Após a leitura de (Thelwall, 2018), iniciou-se a busca por maneiras de se referenciar e utilizar os comentários do *Youtube*, e identificou-se que a API de dados fornecida pela plataforma poderia auxiliar neste processo. Inicialmente, constatou-se que os metadados dos vídeos da plataforma possuem uma categoria e, que, dentre as categorias disponíveis, existe uma denominada "*Education*". Assim, acreditou-se que tal categoria poderia ser utilizada para auxiliar a busca/seleção de vídeos educacionais da plataforma. Nesse sentido, foram realizados os trabalhos Carvalho, Pitangui, Assis, and Andrade (2020) e Carvalho, Pitangui, Trindade, Assis, and Andrade (2020), onde, no primeiro, desenvolve-se um sistema que realiza a buscas por vídeos no *Youtube* pela categoria "*Education*", e no segundo, avalia-se a qualidade das categorizações dos vídeos na pla-

taforma. Percebeu-se, por meio de ambos os trabalhos, que, de forma geral, as categorias dos vídeos do *Youtube* não são confiáveis e que, portanto, não devem ser consideradas determinantes no processo de busca por vídeos. Após tal conclusão, procuraram-se maneiras de coletar e analisar os comentários dos vídeos da plataforma. Nesse sentido, desenvolveu-se o trabalho Carvalho, Pitangui, Trindade, Assis, Andrade, and de Souza (2020) que identificou diferenças significativas entre os vocábulos empregados em vídeos educacionais e não educacionais do *Youtube*. Tal ponto levantou a possibilidade explorada neste trabalho, que diz respeito ao uso de técnicas de Aprendizado de Máquina para a categorização de vídeos educacionais do *Youtube* por meio de comentários.

A abordagem proposta considera a opinião dos usuários, ou seja, seus comentários, para a classificação dos vídeos do *Youtube* em educacionais ou não educacionais. Destaca-se, uma vez mais, que esta abordagem não foi identificada em estudos previamente realizados. Para alcançar o objetivo proposto, são utilizadas técnicas de Mineração de Texto e Aprendizado de Máquina. Nesse sentido, ressalta-se que optou-se por utilizar algoritmos de Aprendizado de Máquina cujos modelos de classificação são inteligíveis, i.e., são interpretáveis pelo ser humano. Tal ponto se justifica, uma vez que possui-se o intuito de mapear e compreender o motivo das classificações realizadas, auferindo, assim, os vocábulos nos comentários que influenciam um vídeo a ser considerado educacional ou não educacional. Adotou-se o *Youtube* como plataforma experimental por esta apresentar um enorme acervo de vídeos dos vais variados tipos, conteúdos, e de diversas qualidades. No entanto, observa-se que, apesar da presente proposta ter sido realizada em uma plataforma de vídeos, é possível ampliá-la e aplica-lá a outros repositórios que possuam comentários/opiniões dos seus usuários.

4 Metodologia

A metodologia adotada para o desenvolvimento do presente trabalho é apresentada na Figura 1 e detalhada a seguir.

1. **Análise dos vídeos.** Esta etapa consistiu na identificação manual (por meio de visualização) de vídeos educacionais e não educacionais da plataforma. Nessa etapa foram analisados 500 vídeos, sendo 250 educacionais, e 250 não educacionais.

   Para a coleta dos vídeos, foi utilizada a seguinte 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. 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 a 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 por-

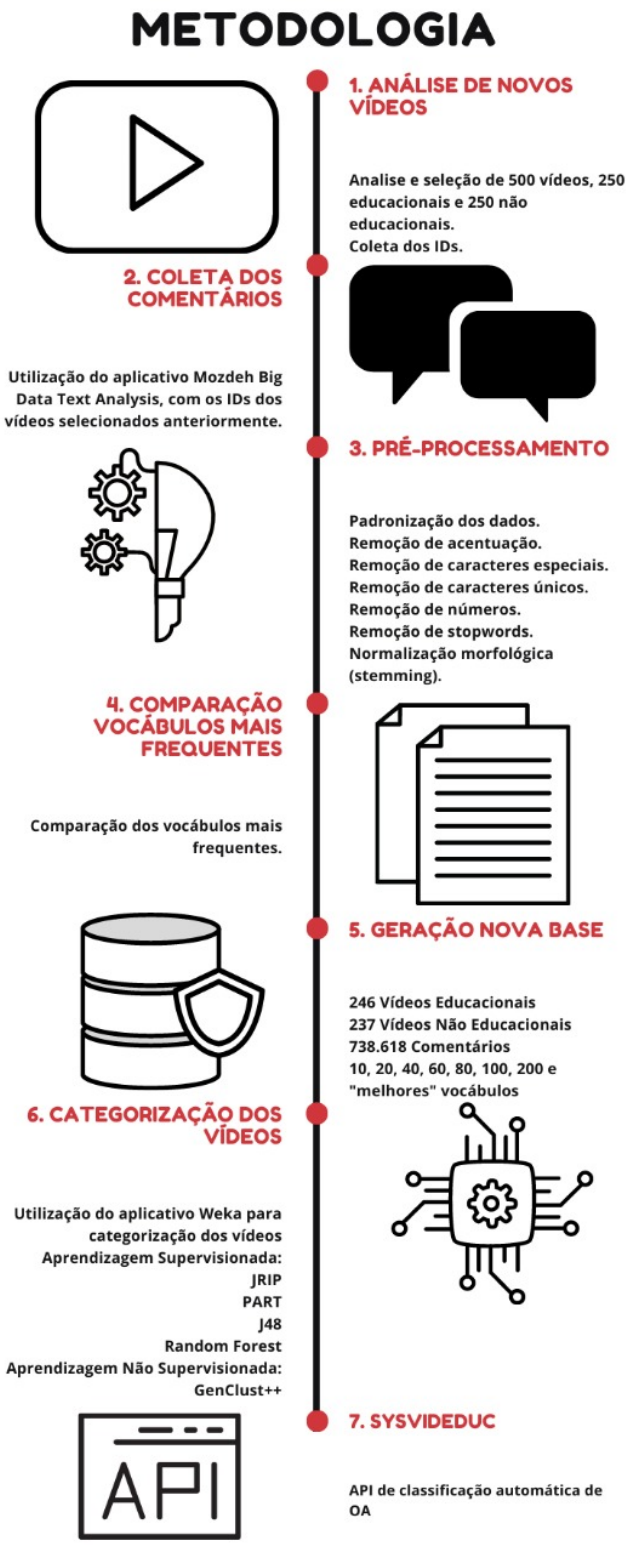

Figura 1: Metodologia utilizada para a pesquisa..

tuguês. 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.

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, como previamente apontado, devem-se buscar outras alternativas para proceder a classificação dos mesmos.

2. **Coleta dos comentários**. Esta etapa consistiu em armazenar os identificadores (*IDs*) dos vídeos selecionados na etapa anterior, e coletar os comentários utilizando o aplicativo *Mozdeh Big Data Text Analysis*, também utilizado em (Thelwall, 2018) e (Carvalho, Pitangui, Trindade, Assis, Andrade, & de Souza, 2020). O aplicativo foi utilizado apenas para a coleta dos comentários, pois apesar de possuir outras funcionalidades, estas são voltadas especificamente para vocábulos em língua inglesa.
3. **Pré-processamento**[1] dos comentários. Esta etapa consistiu nos seguintes passos, a saber.

    - **Padronização dos dados:** transformação de todas as letras em maiúsculo para minúsculo.
    - **Remoção de acentuação:** remoção da acentuação das palavras.
    - **Remoção de caracteres especiais:** 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”, sejam considerados equivalentes.
    - **Remoção de caracteres isolados:** eliminação de caracteres isolados, como “e”, “a”, dentre outros.
    - **Remoção de mais de dois espaços:** remoção de espaçamento extra.
    - **Remoção de números:** eliminação de numerais.

[1] O pré-processamento dos comentários foi realizado por meio de um aplicativo desenvolvido pelos autores utilizando Python e as bibliotecas unidecode, re, e NLTK.

- **Remoção de *stopwords*:** as *stopwords* podem ser consideradas palavras não relevantes para o texto. 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.
- **Normalização morfológica (*stemming*):** consiste em remover os prefixos e os sufixos dos vocábulos, mantendo apenas o radical da palavra, ou seja, fazendo com que, por exemplo, palavras como "professor" e "professora", sejam reduzidas a "profes".

4. **Análise dos vocábulos nos comentários dos vídeos selecionados.** Esta etapa consistiu na identificação dos vocábulos nos comentários dos vídeos selecionados. Após a etapa 3, obteve-se um total de 500 vocábulos nos comentários dos vídeos selecionados, mas nem todos eles são utilizados para a geração das bases de dados, conforme descrito na etapa a seguir.

5. **Geração das bases de dados.** As bases de dados geradas contêm, para cada vídeo selecionado, o *Id*, os vocábulos em seus comentários, e classe do mesmo (educacional ou não educacional). A geração de cada base de dados foi realizada por meio de um aplicativo desenvolvido em Python.

   Conforme apontado na etapa 4, nem todos os 500 vocábulos retornados na etapa 3 são utilizados para representarem os vídeos nas bases de dados geradas. Nesse sentido, para cada uma das oito bases de dados construídas, utilizaram-se quantidades distintas de vocábulos mais frequentes pertencentes aos comentários dos vídeos. Assim, foram foram geradas bases de dados considerando, isoladamente, os 10, 20, 40, 60, 80, 100, e 200 vocábulos mais frequentes nos comentários dos vídeos. O valor máximo, de 200 vocábulos, foi alcançado após testes preliminares indicarem que valores acima de 200 vocábulos não melhoram a acurácia dos modelos de classificação como, adicionalmente, aumentam a complexidade dos mesmos, torando-os, em alguns casos, ininteligíveis. Por sua vez, os valores de 10, 20, 40, 60, 80, e 100 vocábulos foram utilizados no sentido de apontar como os modelos de classificação se comportam com um diverso espectro de quantidades de vocábulos mais frequentes nos comentários dos vídeos selecionados.

   Como já pontuado, foram geradas, no total, 8 bases de dados, sendo que a diferença entre elas se faz pelo número de vocábulos mais frequentes nos comentários dos vídeos. Nesse sentido, a base de dados #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.

   A título de esclarecimento, aponta-se, por exemplo, que a base de dados #1, que possui os

Tabela 3: 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

10 vocábulos mais frequentes nos comentários dos vídeos, contém os 5 primeiros vocábulos mais frequentes nos comentários dos vídeos educacionais, e os 5 primeiros vocábulos mais frequentes nos comentários dos vídeos não educacionais. 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. Esse modelo de estruturação foi utilizado para a formação das sete bases de dados contendo, como apontado, os 10, 20, 40, 60, 80, 100, e 200 vocábulos mais frequentes nos comentários dos vídeos. Por sua vez, a base de dados #8 foi construída conforme previamente apresentado.

Em relação a estrutura 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 3 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.

Observa-se que um mesmo vocábulo pode figurar como o mais frequente em vídeos de ambas as classes (educacional e não educacional). De forma geral, tal ponto não se mostrou um problema para a classificação dos vídeos.

6. **Categorização dos vídeos.** Para a categorização das bases de dados, foram utilizados os algoritmos implementados no *framework* Weka, que fornece uma vasta coleção de ferra-

mentas para classificação, regressão, seleção de atributos, dentre outros (Frank et al., 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 (Cohen, 1995), PART (Frank & Witten, 1998), J48 (Quinlan, 1993) e Random Forest (Breiman, 2001). Para os experimentos “não supervisionados", foi utilizado o GenClust++ (Islam et al., 2018). Todas as técnicas utilizadas foram configuradas com os valores *default* dos seus parâmetros.

Ressalta-se, que optou-se pela utilização de algoritmos baseados em regras e em árvores devido inteligibilidade dos modelos gerados, i.e, devido à facilidade em se abordar e se interpretar as classificações realizadas por esta “categoria"de algoritmos.

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 este 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*).

7. **Desenvolvimento do SysVidEduc.** Desenvolveu-se a primeira versão do SysVidEduc, uma API que utiliza os comentários dos vídeos do *Youtube* e classifica-os em educacionais ou não educacionais. De forma simplificada, o SysVidEduc recebe uma *string* de entrada que representa a expressão de busca ou o *Id* de um vídeo, e exibe os vídeos resultantes da busca, classificados como educacionais ou não educacionais. O SysVidEduc utiliza o Random Forest como classificador. Todo o sistema foi desenvolvido utilizando a linguagem Python, e as bibliotecas: NLTK, re, unidecode, pandas, joblib, sklearn e string. A API é executada via *Web*, possibilitando, desta forma, que Ambientes Virtuais de Aprendizagem realizem uma requisição utilizando uma *string* de busca e obtenham o retorno em um formato *json* com diversos metadados que podem ser utilizados com diversas finalidades.

5 Resultados e Discussões

5.1 Classificação de vídeos do Youtube

A Tabela 4 exibe os vocábulos mais frequentes que figuraram nos vídeos educacionais e não educacionais, após toda a etapa de pré-processamento.

Observa-se que apesar do aumento do número de vídeos e comentários presentes neste trabalho em relação ao Carvalho, Pitangui, Trindade, Assis, Andrade, and de Souza (2020), os prin-

Tabela 4: Vocábulos mais frequentes nos comentários dos 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.

cipais 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 significativa (15.549 vezes em vídeos educacionais e 27.556 vezes em vídeos não educacionais).

A Tabela 5 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, e a coluna "Média” apresenta a média das acurácias obtidas para as técnicas considerando isoladamente cada base de dados. Por fim, a Tabela 5 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.

Tabela 5: Resultados dos experimentos.

Exp.	Vocábulos	JRIP	PART	J48	Random Forest	GenClust++	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

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 expe-

rimentos #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 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 2 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 2: Regras de classificação do JRIP para o experimento #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.

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 3 apresenta as regras de classificação geradas pelo PART para o experimento #8.

```
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 3: Regras de classificação do 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 4 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 4: Árvore de decisão do J48 para o experimento #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 inteligível.

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

Finalmente, 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.

5.2 O sistema SysVidEduc

O sistema 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*. De forma geral, o SysVidEduc funciona de acordo com os seis passos descritos a seguir.

1. O SysVidEduc recebe, via *Web*, uma expressão de busca, que pode ser o *Id* do vídeo ou termos referentes a uma pesquisa.
2. O sistema se conecta à API do *Youtube* e retorna até 50 vídeos, caso a busca tenha sido realizada por uma expressão de busca, ou retorna um único vídeo, caso a busca tenha sido realizada pelo *Id* do vídeo.
3. O sistema busca pelos demais metadados do(s) vídeo(s), incluindo os comentários.
4. Os comentários dos vídeos são processados, tal como descrito na sessão de Metodologia. O resultado desta etapa é enviado ao módulo de Aprendizado de Máquina do SysVidEduc.
5. O módulo de Aprendizado de Máquina do sistema, por meio do Random Forest, classifica o(s) vídeo(s) em educacional ou não educacional, utilizando os comentários processados.
6. O sistema retorna os vídeos classificados como educacionais por meio de uma página *Web*, ou um *json*, que pode ser utilizado conforme desejado.

À título de ilustração, a Figura 5 apresenta o resultado da busca (passo 2) utilizando o termo “herança” (passo 1). Observam-se que são retornados 50 vídeos.

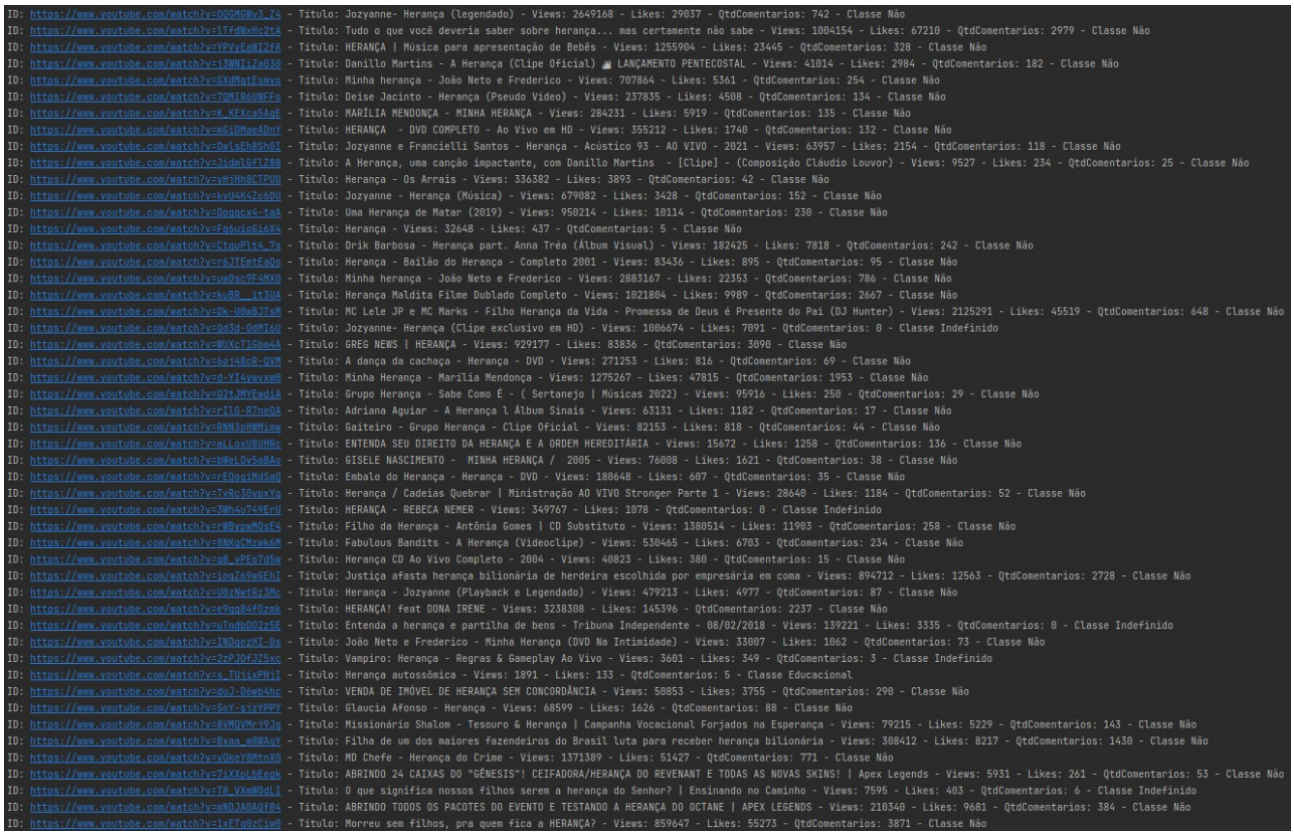

Figura 5: Vídeos retornados pelo Youtube, por meio do SysVidEduc, utilizando-se a expressão de busca “herança”.

A Figura 6 apresenta o único vídeo, dos 50, que foi classificado como educacional pelo SysVidEduc. Neste caso, tal resultado é exibido por meio de uma página *Web*. Por sua vez, a Figura 7 apresenta este mesmo vídeo em formato *json*.

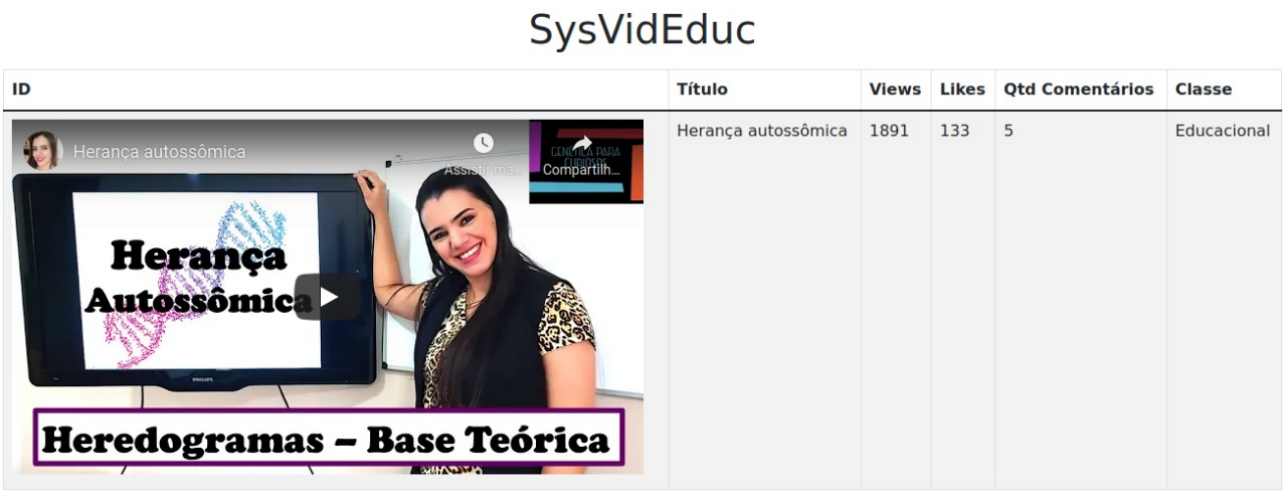

Figura 6: Vídeo classificado como educacional pelo SysVidEduc, apresentado via página *Web*.

["{\n \"classe_random\": \"Educacional\",\n \"comentarios\": [\n \"boa aul ajud bast\",\n \"otim aul\",\n \"aul boa\",\n \"bom parab cust ach vide trat dess assunt facil dinam entend obrig aul\",\n \"obrig aul\"\n],\n \"id\": \"s_TUjixPHjI\",\n \"likes\": \"133\",\n \"qtdcomentarios\": \"5\",\n \"título\": \"Heran\\u00e7a autoss\\u00f4mica\",\n \"views\": \"1891\"\n}"]

Figura 7: Vídeo classificado como educacional pelo SysVidEduc, apresentado via *json*.

A Figura 8 apresenta os 7 primeiros vídeos retornados pelo *Youtube* utilizando a expressão de busca "herança". A busca foi realizada utilizando-se o *Google Chrome* em modo anônimo, sem o uso do sistema SysVidEduc.

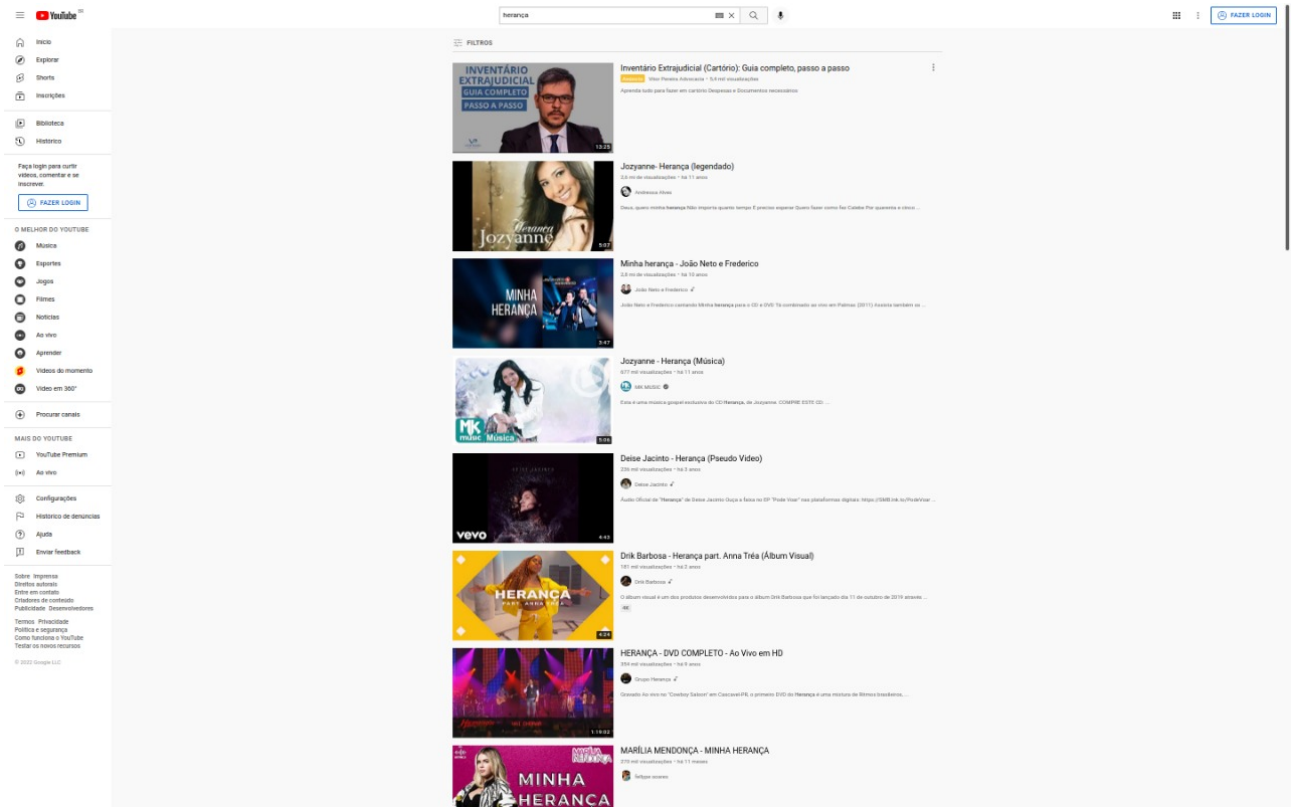

Figura 8: Vídeos retornados pelo Youtube, utilizando-se a expressão de busca "herança".

Observa-se que a busca utilizando-se o temo "herança" no *Youtube* retorna como principais resultados vídeos relacionados a conteúdo musical. Desse modo, caso se opte por vídeos de conteúdo educacional, esta busca deve ser refinada no sentido de se restringir a gama de vídeos retornados. Como previamente apontado, tal refinamento da pesquisa tende a consumir um tempo importante dos discentes e docentes ao pesquisarem por um conteúdo com viés educacional.

Ainda, considerando-se a busca realizada utilizando-se o sistema SysVidEduc e o termo "herança" (Figura 5), observam-se que os vídeos retornados são praticamente os mesmos apresentados na busca exposta anteriormente (com exceção da ordem de alguns vídeos). Assim, aponta-se que o SysVidEdu é capaz de refinar a busca automaticamente, no sentido de que o sistema classificou apenas um vídeo, dentre os 50 retornados, como educacional. O vídeo em questão, de *Id* "s_TUjixPHjI", e título "Herança autossômica", apresenta uma aula sobre as bases gerais e

Herança Autossômica da disciplina de Genética de um curso superior da Universidade Federal do Triângulo Mineiro (UFTM).

Por meio dos exemplos apresentados anteriormente, aponta-se que o sistema proposto se mostra promissor no objetivo de filtrar vídeos de conteúdo educacional. Nesse sentido, indica-se que o SysVidEduc possibilita que os discentes e docentes usem menos tempo na busca e seleção de vídeos de conteúdo educacional.

Sob o ponto de vista de integração do SysVidEduc 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 SysVidEduc 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 SysVidEduc 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.

Durante o desenvolvimento do presente trabalho, levantaram-se questionamentos em relação a como a coleta dos vídeos, qualidade dos mesmos, e seus comentários, poderiam influenciar 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.

- **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 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 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.

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

  Este ponto é uma limitação da presente proposta, uma vez que vídeos sem comentários ou com comentários desativados não poderão ser classificados. A atual versão do SysVidEduc apenas aponta que estes vídeos não são classificáveis, devido à falta de comentários. Uma possível solução para esta limitação seria a utilização das legendas automáticas fornecidas pelo próprio *Youtube*. Assim, os vídeos poderiam ser classificados por meio dos próprios vocábulos das legendas. De qualquer forma, tal proposta não se apresenta como uma solução definitiva, visto que o próprio recurso de geração automática de legendas ainda se mostra em estado inicial de desenvolvimento.

6 Considerações Finais e Trabalhos Futuros

Técnicas de Aprendizado de Máquina têm sido utilizadas com elevada eficácia para solucionar problemas em diversas áreas. Neste trabalho é apresentada e testada uma proposta de aplicação de Técnicas de Aprendizado de Máquina supervisionadas no intuito de classificar vídeos educacionais do *Youtube* por meio de seus comentários. Aponta-se que tal proposta foi capaz de diferenciar, com elevada acurácia (mínima de 83,02% e máxima de 91,30%), vídeos educacionais de vídeos não educacionais. Apontou-se, também, que a técnica de Aprendizado de Máquina não supervisionada utilizada não performou bem nos problemas abordados.

Em um sentido mais prático, o presente trabalho contribui para a para a área Educacional demonstrando a capacidade de utilização de técnicas de Aprendizado de Máquina supervisionadas para auxiliar docentes e discentes a agilizarem o processo de busca e escolha de vídeos educacionais.

Nesse sentido, apresenta-se a primeira versão da API SysVidEduc. A API demonstra elevado potencial para auxiliar docentes e discentes durante a seleção de vídeos educacionais, conferindo agilidade ao processo de escolha de materiais, pois retorna apenas os vídeos voltados para a área da Educação. Além disso, por ser executado via *Web* e poder retornar um *json* como resultado de sua pesquisa, o SysVidEduc pode ser implementado em Ambientes Virtuais de Aprendizagem sem a necessidade de desenvolvimento de novos módulos específicos para cada um desses ambientes.

A continuidade dessa pesquisa se faz em cinco frentes, a saber: 1) desenvolvimento de um sistema de Recomendação de vídeos educacionais do *Youtube* que utiliza os metadados e os comentários dos mesmos a fim de categorizá-los e recomendá-los; 2) uso de técnicas de Aprendizado de Máquina para identificação de vídeos relacionados ao nível de escolaridade e ao conteúdo ministrado; 3) criação e validação de um vocabulário educacional por meio dos comentários coletados; 4) validação do SysVidEduc por docentes e discentes.

Referências

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*. [GS Search]

Afonso, A. R., & Duque, C. G. (2019). Análise de sentimentos em comentários de vídeos do youtube utilizando aprendizagem de máquinas supervisionada. *Ciência da Informação*, *48*(3). [GS Search]

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*. doi: 10.48550/arXiv.1707.02919 [GS Search]

Amanda, R., & Negara, E. S. (2020). Analysis and implementation machine learning for youtube data classification by comparing the performance of classification algorithms. *Jurnal Online Informatika*, *5*(1), 61–72. [GS Search]

Berrar, D. (2019). Cross-validation. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (p. 542-545). Oxford, UK: Academic Press. doi: 10.1016/B978-0-12-809633-8.20349-X [GS Search]

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons. [GS Search]

Braga, J., & Menezes, L. (2014). *Objetos de aprendizagem, volume 1: introdução e fundamentos* (Vol. 1). UFABC. Retrieved from https://pesquisa.ufabc.edu.br/intera/wp-content/uploads/2015/12/objetos-de-aprendizagem-v1.pdf [GS Search]

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1). doi: 10.1023/A:1010933404324 [GS Search]

Carvalho, H. C. F. B., Pitangui, C. G., Assis, L. P., & Andrade, A. V. (2020). 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.* Retrieved from https://esud2020.ciar.ufg.br/wp-content/anais-esud/210418.pdf [GS Search]

Carvalho, H. C. F. B., Pitangui, C. G., Trindade, E. A. C., Assis, L. P., & Andrade, A. V. (2020). Learning objects and youtube: an analysis of videos and their categories. In *Laclo 2020 - xv latin american conference on learning technologies.* doi: 10.1109/LACLO50806.2020.9381145 [GS Search]

Carvalho, H. C. F. B., Pitangui, C. G., Trindade, E. A. C., Assis, L. P. d., Andrade, A. V., & de Souza, D. P. B. (2020). Categorização de vídeos educacionais do youtube por meio de comentários. *RENOTE*, *18*(2), 621-629. doi: 10.22456/1679-1916.110305 [GS Search]

Cohen, W. W. (1995). Fast effective rule induction. In A. Prieditis & S. Russell (Eds.), *Machine learning proceedings 1995* (p. 115-123). San Francisco, CA, USA: Morgan Kaufmann. doi: 10.1016/B978-1-55860-377-6.50023-2 [GS Search]

Dang, S., & Ahmad, P. H. (2014). Text mining: Techniques and its application. *International Journal of Engineering & Technology Innovations*, *1*(4), 22–25. [GS Search]

do Nascimento, P., Barreto, R., Primo, T., Gusmão, T., & Oliveira, E. (2017). Recomendaçao de objetos de aprendizagem baseada em modelos de estilos de aprendizagem: Uma revisao sistemática da literatura. , *28*(1), 213. doi: 10.5753/cbie.sbie.2017.213 [GS Search]

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The weka workbench. online appendix for "data mining: Practical machine learning tools and techniques"*. Morgan Kaufmann Publishers. Re-

trieved from https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf [GS Search]

Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization. In J. Shavlik (Ed.), *Fifteenth international conference on machine learning* (p. 144-151). Morgan Kaufmann. [GS Search]

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17). doi: 10.5120/14937-3507 [GS Search]

Gomes, L. (2008). Vídeos didáticos: uma proposta de critérios para análise. *Revista Brasileira de Estudos Pedagógicos*, *89*(223). doi: 10.24109/2176-6681.rbep.89i223.688 [GS Search]

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, *15*(5), 1-22. doi: 10.1371/journal.pone.0232525 [GS Search]

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114-146. doi: 10.1177/1094428120971683 [GS Search]

Hobbs, J. R., & Riloff, E. (2010). Information extraction. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (p. 511-532). Boca Raton, FL, USA: Chapman and Hall/CRC. doi: 10.1201/9781420085938 [GS Search]

IEEE (2002). Ieee standard for learning object metadata. ieee standard 1484.12.1. New York, NY, USA: Institute of Electrical and Electronics Engineers. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1032843 [GS Search]

Islam, M. Z., Estivill-Castro, V., Rahman, M. A., & Bossomaier, T. (2018). Combining k-means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering. *Expert Systems with Applications*, *91*, 402-417. doi: 10.1016/j.eswa.2017.09.005 [GS Search]

Júnior, C. B., & Dorça, F. (2018). 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)* (p. 1533-1542). doi: 10.5753/cbie.sbie.2018.1533 [GS Search]

Jusoh, S., & Alfawareh, H. M. (2012). Techniques, applications and challenging issue in text mining. *International Journal of Computer Science Issues (IJCSI)*, *9*(6), 431. [GS Search]

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16. [GS Search]

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (icccnt)* (pp. 1–7). doi: 10.1109/ICCCNT.2013.6726842 [GS Search]

Menolli, A., Malucelli, A., & Reinehr, S. (2011). 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).* [GS Search]

Miranda, R. M. d. (2004). *Groa: um gerenciador de repositórios de objetos de aprendizagem*. Unpublished master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, BR. [GS Search]

Mitchell, T. M. (1997). *Machine learning*. New York, NY, USA: McGraw-hill New York. [GS Search]

Morais, E. A. M., & Ambrósio, A. P. L. (2007). Mineração de textos. *Relatório Técnico–Instituto de Informática (UFG)*. [GS Search]

Pinheiro, R. R. A., et al. (2018). *Sistema de recomendação de vídeos educacionais: um estudo de caso no youtube*. Unpublished master's thesis, Universidade Federal de Alagoas, Maceió, AL, BR. [GS Search]

Quinlan, R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA, USA: Morgan Kaufmann Publishers. [GS Search]

Rajput, A., Aharwal, R. P., Dubey, M., Saxena, S., & Raghuvanshi, M. (2011). J48 and jrip rules for e-governance data. *International Journal of Computer Science and Security (IJCSS)*, *5*(2), 201. [GS Search]

Ruggieri, S. (2002). Efficient c4. 5 [classification algorithm]. *IEEE transactions on knowledge and data engineering*, *14*(2), 438-444. doi: 10.1109/69.991727 [GS Search]

Russell, S., & Norvig, P. (2010). *Artificial intelligence: a modern approach*. Upper Saddle River, NJ, USA: Pearson Education. [GS Search]

Sukanya, M., & Biruntha, S. (2012). Techniques on text mining. In *2012 ieee international conference on advanced communication control and computing technologies (icaccct)* (p. 269-271). doi: 10.1109/ICACCCT.2012.6320784 [GS Search]

Sumathi, S., & Sivanandam, S. (2006). *Introduction to data mining and its applications* (Vol. 29). Springer. doi: 10.1007/978-3-540-34351-6 [GS Search]

Thelwall, M. (2018). Social media analytics for youtube comments: Potential and limitations. *International Journal of Social Research Methodology*, *21*(3), 303–316. doi: 10.1080/13645579.2017.1381821 [GS Search]

Trindade, E. A. C., de Assis, L. P., Andrade, A. V., Carvalho, H. C. F. B., Pitangui, C. G., & Dorça, F. A. (2020). Modelagem do problema de cobertura de conjunto para recomendação de objetos de aprendizagem aplicado ao repositório do youtube. *RENOTE*, *18*(2), 358–367. doi: 10.22456/1679-1916.110254 [GS Search]

Vieira, F. J. R., & Nunes, M. A. S. N. (2012). Dica: Sistema de recomendação de objetos de aprendizagem baseado em conteúdo. *Scientia Plena*, *8*(5). [GS Search]

Vijayarani, S., Ilamathi, M. J., Nithya, M., et al. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16. [GS Search]

Vijayarani, S., Janani, R., et al. (2016). Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, *3*(1), 37-47. [GS Search]

Wiederhold, G., & McCarthy, J. (1992). Arthur samuel: Pioneer in machine learning. *IBM Journal of Research and Development*, *36*(3), 329-331. doi: 10.1147/rd.363.0329 [GS Search]

Wiley, D. A. (2000). *Learning object design and sequencing theory*. Unpublished doctoral dissertation, Brigham Young University. [GS Search]

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (Vol. 3). Morgan Kaufmann. doi: 10.1016/C2009-0-19715-5 [GS Search]

Zheng, C., Xue, J., Sun, Y., & Zhu, T. (2021). 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*, *23*(2), e23957. doi: 10.2196/23957 [GS Search]