josif 2022
14ª JORNADA CIENTÍFICA E TECNOLÓGICA
11º SIMPÓSIO DE PÓS-GRADUAÇÃO
IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

ISSN: 2319-0124

ARQUITETURA PARA EXTRAÇÃO DE LEGENDAS E GERAÇÃO DE CONJUNTO DE DADOS PARA PESQUISAS EM ANÁLISE DE SENTIMENTOS.

Renan M. CICILIO[1]; Diego SAQUI[2]

RESUMO

Em pesquisas com análises de sentimentos normalmente é realizado um processamento de linguagem para avaliar o grau de afetividade e seus efeitos em um texto. Infelizmente não existem muitas pesquisas que fazem uso de dados pertencentes a vídeos, assim como não existem muitos conjuntos de dados estruturados para esta finalidade. A proposta deste artigo é a construção de uma arquitetura para realizar extrações de legendas de vídeos do Youtube de diferentes línguas e construir um *dataset* para futuras pesquisas com análise de sentimentos.

Palavras-chave: Vídeos; Youtube Transcribe;

1. INTRODUÇÃO

Vídeos de plataformas de *streaming on demand* são uma das formas mais comuns de entretenimento, pesquisa, divulgação e educação, sendo utilizados como material complementar ao ensino. Além de serem utilizados como material complementar, esses vídeos são essenciais para o funcionamento de cursos que utilizam Educação a distância (EAD), cuja demanda vem crescendo exponencialmente devido ao fácil acesso à internet proporcionado pelo desenvolvimento das tecnologias de informação nas últimas décadas (NUNES; et al. 2007).

Infelizmente não existem muitas pesquisas que fazem uso de dados pertencentes a vídeos, assim como não existem muitos conjuntos de dados estruturados para esta finalidade. Por exemplo, em pesquisas com análises de sentimentos normalmente é realizado um processamento de linguagem para avaliar o grau de afetividade e seus efeitos em um texto (ZHANG; WANG; LIU, 2018). Para realizar esse tipo de análise em vídeos é necessário ter uma transcrição do vídeo em si, permitindo assim que algoritmos façam suas operações através do texto. Outras estatísticas para avaliar aspectos positivos e negativos do vídeo também podem ser úteis para diferentes análises.

Baseado nas questões descritas, se torna clara a oportunidade que conjuntos de dados com tais informações podem contribuir para a realização de novas pesquisas em análise de sentimentos em vídeos, seja analisando suas estatísticas, comentários, ou a linguagem utilizada. Portanto, este estudo tem o propósito de gerar uma arquitetura para obter informações relevantes e a extração de legendas automaticamente geradas ou manualmente escritas de vídeos da plataforma Youtube para a

[1] **Renan M. Cicilio , Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais - *Campus* Muzambinho. Email:renanarms11@gmail.com .**
[2] **Diego Saqui, Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais - *Campus* Muzambinho. Email: diego.saqui@muz.ifsuldeminas.edu.br.**

construção de um conjunto de dados que contribuirá para futuras pesquisas em análise de sentimentos.

2. MATERIAL E MÉTODOS

Neste estudo, uma arquitetura foi estabelecida para gerar um conjunto de dados para pesquisas com vídeos e é representada na Figura 1. Inicialmente, um conjunto de dados prévios, que contém algumas estatísticas de vídeos, como id do vídeo, *likes*, *deslikes* e *link* disponibilizadas pelo Youtube no Kaggle foi obtido[3]. Posteriormente, utilizando um algoritmo desenvolvido em linguagem Python e bibliotecas auxiliares como Pandas, os dados desse conjunto são percorridos e nossa arquitetura visita o *link de* cada vídeo. Então um outro algoritmo, que possibilita a extração de legendas de vídeos do Youtube é aplicado; O algoritmo consegue obter as legendas inseridas manualmente pelo autor ou geradas automaticamente pelo próprio Youtube e as armazena em um arquivo de texto nomeado conforme o id do vídeo. Para este propósito a biblioteca em Python chamada YoutubeTranscribe que é responsável por extrair as legendas de vídeos e armazená-las em um arquivo de texto foi utilizado. Por fim, nossa arquitetura insere o *caminho* do arquivo da legenda de cada vídeo em uma coluna do conjunto de dados.

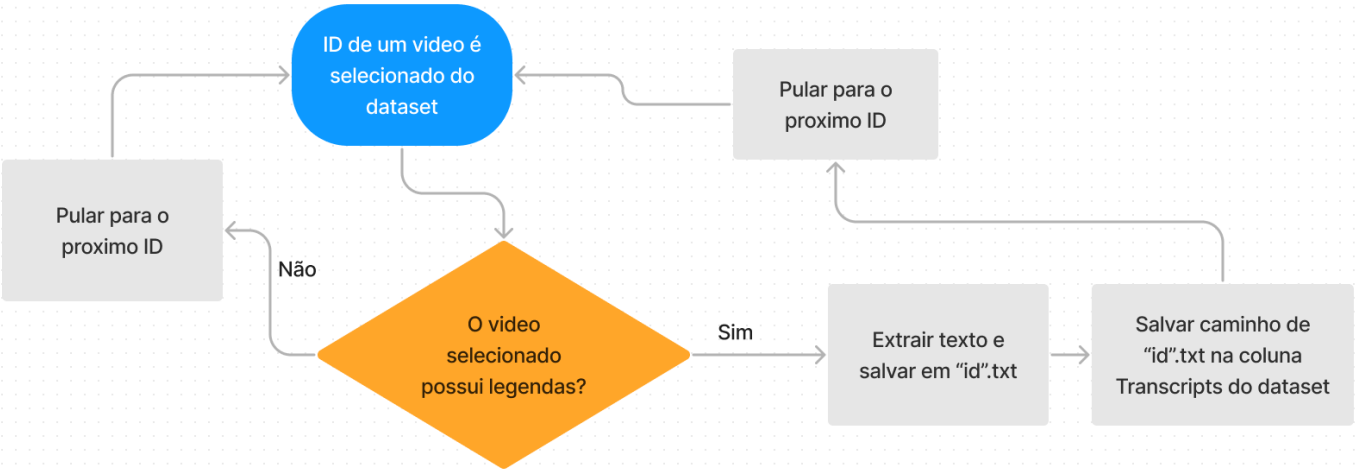

Figura 1. Arquitetura para geração do conjunto de dados para pesquisas com vídeos.

Fonte: Elaboração do autor (2022)

3. RESULTADOS E DISCUSSÕES

Fazendo uso da metodologia explicada anteriormente o algoritmo primeiramente selecionou os *ids* dos vídeos listados e o usou para extrair as legendas dos mesmos e transcrevê-las em um arquivo de texto nomeado com o *id* do vídeo para facilitar buscas futuras. Após isso o algoritmo escreve o caminho, ou *path*, do arquivo de texto em uma nova coluna nomeada "Transcript" *dataset*

[3] https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset

inicial e o salva como um novo documento. Uma amostra do *dataset* gerado com a coluna *Transcript* armazenando o caminho da legenda é mostrado na Figura 2 e exemplos de arquivos contendo as legendas são mostrados na Figura 3.

C	D	E	F	G	H
channelTitle	view_count	likes	dislikes	comment_count	Transcript
Brawadis	1514614	156908	5855	35313	./content/3C66w5Z0ixs.txt
Apex Legends	2381688	146739	2794	16549	./content/M9Pmf9AB4Mo.txt
jacksepticeye	2038853	353787	2628	40221	./content/J78aPJ3VyNs.txt
XXL	496771	23251	1856	7647	./content/kXLn3HkpjaA.txt
Mr. Kate	1123889	45802	964	2196	./content/VIUo6yapDbc.txt
Professor Live	949491	77487	746	7506	./content/w-aidBdvZo8.txt
Les Do Makeup	470446	47990	440	4558	./content/uet14uf9NsE.txt
CGP Grey	1050143	89190	854	6455	./content/ua4QMFQATco.txt
Louie's Life	1402687	95694	2158	6613	./content/SnsPZj91R7E.txt
Rancho Humilde	741028	113983	4373	5618	./content/SsWHMAhshPQ.txt
CaseyNeistat	940036	87111	1860	7052	./content/49Z6Mv4_WCA.txt
Smosh Pit	591837	44168	409	2652	./content/nt3VVyv5pxQ.txt
Ubisoft North America	320872	14288	774	2085	./content/I6hswz4rlrU.txt
LilYachtyVEVO	413372	26440	293	1495	./content/W7VK4DUHvKU.txt
Kyle Exum	921261	124183	1678	16460	./content/W9Aen8hG20Y.txt
Tyler Cameron	105955	4511	69	673	./content/BNeDH6UTmXw.txt
HollywoodLife	1007540	10102	7932	2763	./content/6TIsR_7nrNc.txt
Cole The Cornstar	277338	37533	197	3666	./content/gPdUsIndvVI.txt
Chloe Ting	1648441	130147	1425	15773	./content/GTp-0S82guE.txt
JYP Entertainment	5999732	714287	15174	31039	./content/jbGRowa5tIk.txt
Mark Rober	14684474	544038	15818	33507	./content/vePc5V4h_kg.txt
Screen Junkies	833369	50181	1120	4634	./content/5WjcDji3xYc.txt
FaZe Rug	3061467	206840	2646	14934	./content/FoplxceEr8g.txt
James Charles	3662673	394675	5757	27346	./content/p7HGUZWq_8s.txt

Figura 2. Exemplo de algumas colunas e amostras do *dataset* criado

Fonte: Elaboração do autor (2022)

3C66w5Z0ixs.txt

```
1 i wanna mess up your covers
2 [Music]
3 i'm the best thing that has ever
4 happened to
5 you here's your time to shine and show
6 me what you can do
7 [Music]
8 the way my lips feel on your body
9 [Music]
10 but baby i don't want to lead you on
11 looking for something steady
12 what is going on you guys brandon here
13 aka broads and today
14 i am bringing you guys the most special
15 video on this channel today
16 august 10 2020 i am going to step up to
17 the plate
18 and ask my best friend the girl i care
19 about the most
20 in my life the girl i spend the most
```

1mZ1f_8YOEc.txt

```
1 THIS IS CHICAGO'S VERY OWN
2 WGN MORNING NEWS AT 5.
3 CHAOS IN DOWNTOWN CHICAGO
4 THIS MORNING LOOTING AND
5 RIOTING LEADS TO ARREST
6 STRETCHING FROM MICHIGAN AVENUE
7 TO THE SOUTH LOOP.
8 AND BROUGHT A LITTLE TEA HAS
9 BEEN WATCHING IT ALL FROM SKY
10 CAM 9 THIS MORNING, SHE'S LIVE
11 WITH WHAT WE LEARNED RONA.
12 IT IS QUITE A MORNING.
13 WE'RE LOOKING AT THE OHIO
14 THEATER RIGHT NOW IT IS BLOCKED
15 OFF ACCESS TO GAIN 10 LOT OF
16 CLOSURES OVER THE CITY THIS
17 MORNING A LOT OF CHAOS EARLIER
18 IN THE NIGHT THOUGH THINGS DO
19 SEEM TO BE A LITTLE BIT MORE
20 UNDER CONTROL THOUGH STILL
```

Figura 3. Exemplo de legendas extraídas

Fonte: Elaboração do autor (2022)

Os resultados demonstram que é possível não apenas realizar extrações de legendas de forma ordenada como visto nos exemplos acima, mas também é possível evitar erros que ocorrem ao tentar extrair dados de um vídeo cujo legendas estão desabilitadas, evitando assim poluição no dataset e melhorando os resultados de uma busca.

No total 100 vídeos foram usados para extração, dos quais apenas 3 não possuem legendas acessíveis; dos videos cujas legendas foram extraídas corretamente é possível ver variações entre textos onde as legendas foram geradas automaticamente em comparação com legendas inseridas manualmente; legendas manuais possuem mais coerção e paragrafação bem definida, tornando-os o ideal para ser usado em análises, já legendas automáticas são mais curtas e possuem menos coerção e quase não possui paragrafação, porém ainda é possível usá-lo para realização de análises.

4. CONCLUSÕES

Concluir-se que é possível fazer a utilização da extração de legendas em larga escala para realização de estudos que requerem grandes quantias de dados para serem analisados e estudados; já que mesmo possuindo algumas falhas, possuem coerção o bastante para serem lidos e compreendidos sem problemas significativos por humanos ou algoritmos de análise de dados.

REFERÊNCIAS

NUNES, Thiago Soares; TECCHIO, Edivandro Luiz. A utilização de vídeo-aulas e videoconferências no aprendizado do estudante na educação a distância. Repositório Institucional, [*S. l.*], p. 9 -11, 29 nov. 2007. Disponível em: <https://repositorio.ufsc.br/handle/123456789/89366>. Acesso em: 9 mar. 2022.

ZHANG, Lei; WANG, Shuai; LIU, Bing. Deep learning for sentiment analysis: A survey. WIREs, [*S. l.*], ano 2018, 30 mar. 2018. Encyclopedia of Computer Science, p. 1 - 5. DOI <https://doi.org/10.1002/widm.1253>. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1253>. Acesso em: 18 mar. 2022.