

# Descoberta de Perfis de *Youtubers* via Aprendizagem de Máquina

Matheus Oliveira Costa, Anderson Alencar, Pedro Henrique Pereira,  
Ticiania L. Coelho da Silva, Georgia Cruz, Andrea Pinheiro

<sup>1</sup>Instituto Universidade Virtual. Fortaleza - Ceará, Brasil

{matheusoliveiracos, anderson.sjav}@gmail.com,

{pedrohenripereira}@alu.ufc.br,

{ticianalc, georgia, andrea}@virtual.ufc.br

**Resumo.** *O uso progressivo de redes sociais, como o Youtube, nos últimos anos tem produzido um grande volume de informações geradas pelos seus usuários, que com frequência compartilham seus sentimentos, opiniões sobre vídeos, bem como novos conteúdos em formato de mídias (vídeos). Compreender o perfil desses usuários produtores de conteúdo, mais conhecidos como youtubers, pode ser um diferencial para sistemas de recomendação que podem sugerir canais na plataforma Youtube para seus usuários com mais precisão, e ainda, auxiliar no entendimento de como a comunicação falada de crianças e adultos está mudando com o passar dos anos. No entanto, analisar esse grande volume de dados de forma não automatizada consiste em um problema não trivial. Este trabalho visa descobrir e identificar os perfis de Youtubers utilizando técnicas de Aprendizagem de Máquina.*

## 1. Introdução e Motivação

As novas aplicações da Web 2.0 permitem aos indivíduos criarem conteúdo e colocarem na rede para outros verem e interagirem de forma praticamente livre e fácil. Um dos maiores representantes dessa nova realidade é o site de compartilhamento de vídeos YouTube.

Criado em 2005, o YouTube permite que qualquer pessoa possa enviar, assistir e interagir com outros usuários na sua rede, de forma gratuita e simples. Esse diferencial possibilitou milhares de pessoas a criarem e disponibilizarem conteúdo original online. O site também conta com um complexo sistema de recomendação, que se baseia no perfil de uso dos usuários para indicar vídeos relacionados.

No entanto, não existe algo de mesma magnitude no caso de recomendação de *youtubers* e também não é de conhecimento dos autores nenhuma ferramenta que execute a ação de relacionar vídeos aos seus criadores de forma a classificar e agrupar tais vídeos de acordo com o perfil do *youtuber*. Isso pode ser relevante para sistemas de recomendação de canais ou de *youtubers* aos usuários da plataforma. E ainda, auxiliar no entendimento de como a comunicação falada de crianças e adultos está mudando com o passar dos anos. Segundo a Oxford, *youtuber* é um usuário frequente do site de compartilhamento de vídeos YouTube, especialmente alguém que produz e aparece em vídeos no site.

Este trabalho visa propor uma solução por meio da análise dos textos utilizando técnicas de aprendizagem de máquina nas legendas dos vídeos do YouTube, de tal sorte que canais ou *youtubers* de alta similaridade (ou mesmo perfil) estejam em agrupados em um mesmo perfil ou grupo. Bem como, canais ou *youtubers* dissimilares estejam em grupos diferentes.

Os resultados desta pesquisa, além de auxiliarem na descoberta dos perfis de *youtubers*, servirão como base para construção de uma base de treinamento para um classificador de vídeos, capaz de categorizar um dado vídeo, se ele se refere a um conteúdo para crianças, para amantes de jogos, maquiagem, entre outros perfis.

## 2. Desenvolvimento da Pesquisa

Clusterização de Dados refere-se a um conjunto de técnicas de aprendizagem de máquina que pode ser definida como o processo de dividir grandes conjuntos de dados em subconjuntos chamados *clusters*. A divisão é feita baseada no algoritmo de análise escolhido, de forma que os objetos dentro de um *cluster* são suficientemente semelhantes entre si e diferentes dos objetos dos demais *clusters*. Nesse contexto, diferentes métodos de análise podem gerar diferentes agrupamentos no mesmo conjunto de dados. Consequentemente, o processo de análise é útil, pois pode levar a descoberta de grupos de dados anteriormente desconhecidos [Han et al. 2011].

Inicialmente, foi realizada uma coleta de diversos canais famosos da plataforma Youtube e extraída a legenda de vídeos de cada um deles, entre eles: Crescendo com Luluca, Mundo da Vivi, Luiza Vinco, Diário da Mafer, Caduzinho Carvalho, Papai Silva, Piero Start, Canal da Lulu, Felipe Neto, Aprendendo a brincar com Sarah Maria, Canal da Lele, Canal do Dudu, Rick Santana, O Mundo Nerd de Tiffany, Pedro Maia, Cadu Top Games, Fabiana Landim - Landimas, Yasmin Galvão, Eduarda Ferrão, Fabinho Ferrão, entre outros. O total de texto atualmente é de 300 legendas.

Algoritmos de clusterização, tais como *K-means*, *DBSCAN* e *Agglomerative Clustering* estão sendo aplicados nestes dados utilizando a ferramenta *scikit-learn* a fim de atingir os objetivos apresentados a seguir.

Objetivo Geral:

- Descoberta de grupos de canais ou *youtubers* que apresentem alta similaridade entre si.

Objetivos Específicos:

- Coletar e disponibilizar a base de textos com as legendas de diferentes vídeos do YouTube.
- Identificar o algoritmo de clusterização que devolve o conjunto de perfis ou *clusters* de maior qualidade de acordo com uma métrica a ser escolhida.

## 3. Perspectivas de Trabalhos

Esta pesquisa pretende ao final do trabalho, disponibilizar uma base de textos das legendas de vídeos do YouTube e ser identificado o algoritmo de clusterização que devolve o conjunto de *clusters* de maior qualidade de acordo com uma métrica a ser escolhida.

Bem como, conforme já dito anteriormente os resultados desta pesquisa serão utilizados para construir uma base de treinamento para um classificador de vídeos. O

papel do classificador será categorizar um dado vídeo, se ele se refere a um conteúdo para crianças, para amantes de jogos, maquiagem, entre outros perfis. Não é do conhecimento dos autores nenhuma abordagem que realize o mesmo processo.

## **Referências**

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.