UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Leandro Massetti Ribeiro Oliveira

Composição de Objetos de Aprendizagem Multimídia Através de Sumarizadores Automáticos de Texto Baseados em Modelos Deep Learning

São Luís - MA

2022

Leandro Massetti Ribeiro Oliveira

Composição de Objetos de Aprendizagem Multimídia Através de Sumarizadores Automáticos de Texto Baseados em Modelos Deep Learning

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Programa de Pós-Graduação em Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Carlos de Salles Soares Neto

São Luís - MA

2022

Leandro Massetti Ribeiro Oliveira

Composição de Objetos de Aprendizagem Multimídia Através de Sumarizadores Automáticos de Texto Baseados em Modelos Deep Learning

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Trabalho Aprovado. São Luís - MA, 16 de Setembro de 2022:

Prof. Dr. Carlos de Salles Soares Neto
Orientador
Universidade Federal do Maranhão

Prof. Dr. Alexandre César Muniz de Oliveira
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Windson Viana de Carvalho
Examinador Externo
Universidade Federal do Ceará

São Luís - MA
2022

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Ribeiro Oliveira, Leandro Massetti.
Composição de Objetos de Aprendizagem Multimídia Através de Sumarizadores Automáticos de Texto Baseados em Modelos Deep Learning / Leandro Massetti Ribeiro Oliveira. - 2022.
50 f.

Orientador(a): Carlos de Salles Soares Neto.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, São Luís - MA, 2022.

1. Deep Learning. 2. Objeto de Aprendizagem. 3. Sumarização de Texto. 4. Transformers. I. de Salles Soares Neto, Carlos. II. Título.

"Se a educação sozinha não transforma a sociedade,
sem ela tampouco a sociedade muda."

(Paulo Freire)

Resumo

Um Objeto de Aprendizagem (OA) é um recurso digital, que pode ser utilizado e reutilizado ou referenciado durante um processo de suporte tecnológico ao ensino e aprendizagem. Apesar de serem principalmente multimídia, com áudio, vídeo, texto e imagens sincronizados entre si, alguns recursos digitais de educação possuem textos como um de seus elementos principais no processo de ensino, como sites, textos, vídeo-aulas, seminários, e a sumarização desses textos podem ser uma forma de composição de OAs multimídia. No entanto, a sumarização de textos é um processo oneroso em tempo e esforço, gerando a necessidade de buscar novas formas de gerar esse conteúdo. Este trabalho apresenta uma solução para a composição de OAs multimídia através de sumarizadores automáticos de texto baseados em modelos *Deep Learning Transformers* a partir de dois experimentos: O primeiro fazendo a composição de OAs a partir de textos educacionais na língua portuguesa utilizando tradutores e sumarizadores de texto, neste experimento os resultados apresentados foram positivos e permitem comparar o desempenho dos resumos como geradores de OA em formato de texto; O segundo experimento apresenta uma solução de sumarização de vídeos educacionais utilizando os mesmos modelos de *Deep Learning* para a sumarização da legenda, os testes foram realizados utilizando o *dataset* EDUVSUM no qual foi possível melhorar os resultados do artigo original alcançando 26,53% de acurácia em um problema multi-classe e erro absoluto médio de 1,49 por *frame* do vídeo e 1,45 por segmento de vídeo.

Palavras-chave: Sumarização de Textos, Objeto de Aprendizagem, Deep Learning, Transformers.

Abstract

A Learning Object (LO) is a digital resource that can be used and reused or referenced during a process of technological support for teaching and learning. Despite being mostly multimedia, with audio, video, text and images synchronized with each other, some digital education resources have texts as one of their main elements in the teaching process, such as websites, texts, video classes, seminars, and the summarization of these texts can be a way of composing multimedia LOs. However, text summarization is a costly process in time and effort, creating the need to seek new ways to generate this content. The present work show a solution for the composition of multimedia LOs through automatic text summarizers based on Deep Learning Transformers models from two experiments: The first one composing LOs from educational texts in Portuguese using translators and text summarizers, in this experiment the results presented were positive and allow comparing the performance of summaries as generators of LO in text format; The second experiment presents an educational video summarization solution using the same Deep Learning models for subtitle summarization, the tests were performed using the EDUVSUM dataset in which it was possible to improve the results of the original article reaching 26.53% accuracy in a multi-class problem and average absolute error of 1.49 per video frame and 1.45 per video segment.

Keywords: Text Summarization, Learning Object, Deep Learning, Transformers.

Lista de ilustrações

Lista de abreviaturas e siglas

ACC	Acurácia
API	*Application Programming Interface*
B_ACC	Acurácia Balanceada
BART	*Bidirectional and Auto-Regressive Transformers*
BERT	*Bidirectional Encoder Representations from Transformers*
CCN	*CommonCrawl News*
DistilBART	*Distiled Bidirectional and Auto-Regressive Transformers*
DistilBERT	*Distiled Bidirectional Encoder Representations from Transformers*
EAM	Erro Absoluto Médio
EDUVSUM	*Educational Video Summarization*
GPT	*Generative Pretrained Transformer*
GPU	Unidade de Processamento Gráfico
LDA	*(Latent Dirichlet Allocation*
LSA	*Latent Semantic Analysis*
MLP	*Multilayer Perceptron*
MS MARCO	*Microsoft Machine Reading Comprehension*
PCG	*Procedural content generation*
PEGASUS	*Pre-training with Extracted Gap-sentences for Abstractive Summarization*
PLN	Processamento de Linguagem Natural
RNN	*Recurrent Neural Network*
RoBERTa	*A Robustly Optimized BERT Pretraining Approach*
ROGUE	*Recall-Oriented Understudy for Gisting Evaluation*
SBIE	Simpósio Brasileiro de Informática na Educação

STS	*Semantic Text Similarity*
SVM	*Support Vector Machines*
TF-IDF	*Term Frequency – Inverse Document Frequency*
TvSum	*Title-based Video Summarization*
OA	Objeto de Aprendizagem

Sumário

1 Introdução

A popularização da internet facilitou o acesso do público geral a diversos conteúdos digitais, conteúdos que são documentados nos mais variados domínios. Considerando isso, os meios digitais se tornaram uma opção como forma de educação. Desta forma é possível adaptar recursos educacionais para ferramentas digitais como texto, áudio, vídeo, animações entre outros. Estes recursos são denominados Objetos de Aprendizagem (OA), a flexibilidade e possibilidade de reutilização são algumas de suas características (TAROUCO et al., 2014).

No entanto, o processo de autoria desses OAs multimídias se trata de uma tarefa de grande complexidade, desde o seu processo de *desing*, seus elementos de interação como até a inserção de objetos pedagógicos definidos. Todo esse processo gera mais tempo, dinheiro e recursos humanos necessários para produzir esse tipo de conteúdo, além também da aquisição de *softwares* e *hardwares* (ARAÚJO et al., 2014).

Considerando por exemplo os OAs do tipo vídeo, a edição de vídeos por si só já é uma tarefa que consome bastante tempo de um editor devido a complexidade de manter a continuidade e outros fatores de *design*. Quando se trata de vídeos educacionais exige uma equipe multidisciplinar para a tarefa, como por exemplo a necessidade de um conteudista (aquele que tem conhecimento no domínio abordado pelo OA); Um *designer* instrucional (aquele que indicará a melhor forma de apresentar o conteúdo); Desenvolvedor/Programador; Além de outras tarefas como a revisão ortográfica e metodologias ativas.

Muitos recursos digitais contém texto como um de seus elementos, como por exemplo resumos textuais, perguntas e respostas, vídeos (com legendas ou textos nas imagens), *podcasts* com transcrição e animações.

Esses recursos podem ser utilizados como objeto de aprendizagem devido a sua capacidade de ensino, reutilização e flexibilidade. Levando esse fator em comum, é possível compor OAs através de tratamento de texto com uso de ferramentas de Processamento de Linguagem Natural. Uma dessas ferramentas são os modelos de *Deep Learning Transformers*.

O Processamento de Linguagem Natural (PLN) surgiu devido à necessidade de compreensão e geração automática da linguagem natural. Trata-se de forma extrair as informações de textos, facilitar a entrada de dados nos sistemas e a estruturação desses dados, além da geração de textos semelhantes aos de uma comunicação pessoal (SANTOS et al., 2015). De acordo com Chowdhary (2020), PLN é a coleção de técnicas computacionais para a análise automática e representação das linguagens humanas, motivadas pela teoria.

Uma das aplicações de PLN é a sumarização automática de textos, que trata-se da tarefa de produzir resumos concisos e naturais sem perder informações chave de conteúdo e o seu significado geral. É uma tarefa complexa e desafiadora, nós humanos quando resumimos um texto, geralmente o lemos inteiramente para desenvolver nossa compreensão e, em seguida, escrever um resumo destacando seus pontos principais. É complexo fazer um computador assimilar uma linguagem natural, tornando a tarefa de resumir não-trivial (ALLAHYARI et al., 2017). A tarefa de sumarização de textos entra no domínio da PLN denominado *sequence-to-sequence* no qual um texto resumido (sequência de palavras) é gerado através de um texto maior.

Quando se trata da geração de texto (*sequence-to-sequence*) as redes Transformers (VASWANI et al., 2017) foram uma grande melhora de performance em tarefas de processamento de linguagem natural. Através dela surgiram vários outros modelos como o BERT, DistilBERT, RoBERTa, para tarefas de geração de linguagem natural, inferência de palavras, *question-answering*, tradução, sumarização de textos entre outras tarefas. Devido a sua robustez essas ferramentas podem ser utilizadas em diversos segmentos, incluindo na geração de objetos de aprendizagem.

A sumarização automática de textos é uma das tarefas de geração de linguagem natural que os modelos baseados em Transformers podem ser aplicados. No contexto educacional existe a necessidade do seu uso para a geração de OAs devido ao processo de criação ser oneroso e complexo. Como por exemplo na tarefa de resumir conteúdos educacionais para disponibilização de forma digital para alunos, a realização desta tarefa de maneira automática diminuiria o tempo necessário para a realização. Os modelos de Deep Learning baseados em Transformers podem ser uma solução pois conseguem atingir resultados semelhantes ao de resumo feito por especialistas (VASWANI et al., 2017).

A sumarização de texto utilizando modelos Transformers, no entanto, possui algumas limitações para a composição de OA. A principal é a desses modelos não serem treinados para todos os idiomas, devido a necessidade de *datasets* gigantescos e grande quantidade de tempo de treinamento, a maioria dos modelos são para a língua inglesa e os que existem para a língua portuguesa por enquanto conseguem apenas fazer uma representação vetorial das palavras.

Neste sentido, este trabalho contribui para a solução de dois problemas: O primeiro trata-se da criação de OAs de texto baseados no resumo dos mesmos. Já no segundo problema, almeja-se que vídeos educacionais possam ser sumarizados, através da classificação de importância dos segmentos, com base no resumo de suas legendas.

Para a primeira etapa, que envolve a sumarização automática de textos educacionais, pretende-se empregar modelos de *Deep Learning* baseados em redes *Transformers*, onde será verificado se os mesmos conseguem manter sua estrutura gramática além de sua estrutura educacional, considerando que esses modelos são para a língua inglesa, deseja-se verificar

a viabilidade do uso de tradutores automáticos para complementar este experimento.

Já para a segunda etapa, deseja-se expandir para a geração de OAs do tipo vídeo, por meio da sumarização dos mesmos utilizando sua transcrição em texto. Nesta etapa será verificado se é possível classificar os segmentos mais importantes de uma vídeo-aula, seminário ou palestra com foco na legenda e por meio da sumarização automática da transcrição da fala do orador.

1.1 Objetivos

Este trabalho tem como objetivo verificar a viabilidade do uso de modelos *Deep Learning* pré-treinados baseados em *Transformers* para a geração automática de Objetos de Aprendizagem, tanto textuais como de vídeos educacionais.

1.1.1 Objetivos Específicos

Para alcançar tal objetivo, este trabalho é dividido em dois experimentos:

- A geração de objetos de aprendizagem textuais através da sumarização automática de textos provenientes da Wikipédia, além disso, avaliar a sua capacidade educacional por pesquisa com especialistas da área
- A sumarização de vídeo educacionais através da classificação dos segmentos mais importantes de uma apresentação em vídeo. Para isso é utilizado a transcrição da fala do orador do vídeo para sumarização de texto e em seguida correlação com os segmentos em que o texto remanescente pertence. Por fim a utilização de um *dataset* de vídeos educacionais para avaliação da sumarização dos vídeos.

1.1.2 Organização do Trabalho

Este trabalho está estruturado da seguinte forma:

- O Capítulo 2 apresenta os trabalhos relacionados a geração de Objetos de Aprendizagem e a sumarização de vídeos educacionais.
- O Capítulo 3 trata da fundamentação teórica dos modelos utilizados nos experimentos seguintes, nele teremos uma visão geral sobre as redes neurais e os modelos de *Deep Learning* utilizados no trabalho.
- O Capítulo 4 apresenta o primeiro experimento relacionado a geração de OAs através da sumarização de textos educacionais a partir de modelos *Deep Learning*.

- O Capítulo 5 apresenta o segundo experimento relacionado a sumarização de vídeos educacionais utilizando sumarizadores de textos baseados em modelos *Deep Learning*.
- O Capítulo 6 apresenta as considerações finais do trabalho realizado nos dois experimentos além do levantamento de possíveis trabalhos futuros.

2 Trabalhos Relacionados

Neste Capítulo são descritos os trabalhos relacionados aos dois experimentos abordados nesta dissertação. Desta forma ele está dividido em duas seções: A Seção 2.1 apresenta os trabalhos relacionados a geração automática de OAs; A Seção 2.2 apresenta os trabalhos relacionados a sumarização de vídeos educacionais.

2.1 Geração Automática de Objetos de Aprendizagem

De acordo com Kurdi et al. (2020), estudos recentes com foco na geração de conteúdo educacional utilizam a avaliação de profissionais especialistas como forma de validar o conteúdo gerado. Essa abordagem parece ser um procedimento padrão e geralmente é um fator de boa qualidade. No entanto, apontam que a avaliação dos profissionais é apenas um fator. É fundamental medir a avaliação dos alunos e demonstrar a usabilidade prática, legibilidade e discernimento do conteúdo gerado.

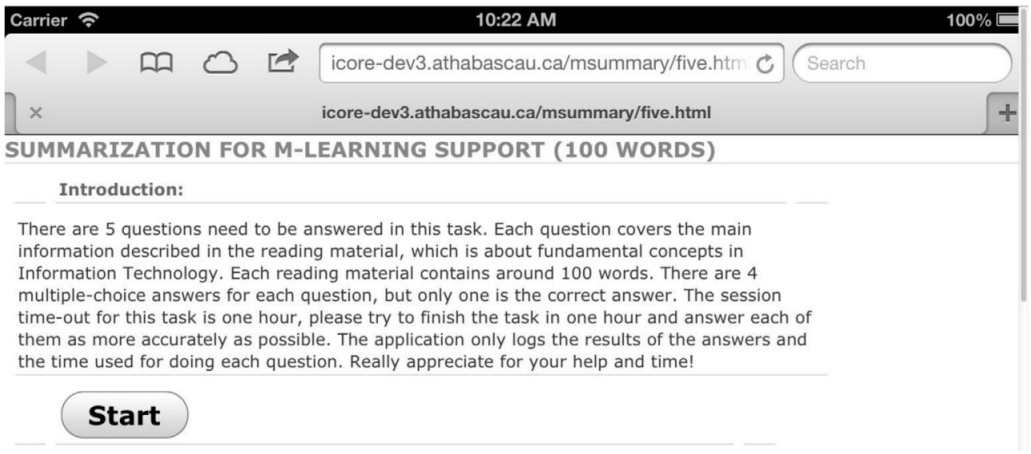
Rocha et al. (2020), por exemplo, propõem um método que explora bases educacionais estruturadas para gerar perguntas de forma automática. Na experimentação, professores voluntários validaram a qualidade de 100 questões geradas por seu algoritmo, obtendo uma classificação média de 3,5 em uma escala entre 1 e 5.

Yang et al. (2013) defende que os conteúdos educacionais são mais difíceis de serem absorvidos quando são extensos. Assim, o autor apresenta uma solução para resumir esses conteúdos educacionais para melhor se adaptar a um ambiente de dispositivos móveis, A Figura 1 apresenta um exemplo da interface da aplicação com os textos resumidos por sumarizadores automáticos. No estudo, o autor utilizou questões para avaliar o aprendizado dos alunos por meio de resumos gerados por um sumarizador automático. Como resultado, o artigo mostra que a metodologia é eficiente para que o aluno adquira conhecimento e se adapte melhor em um ambiente de dispositivo móvel.

Rüdian, Heuts e Pinkwart (2020) apresentam em seu artigo um problema referente à geração automática de perguntas, já que a maioria deles trabalha apenas em sentenças de entrada única. Isso limita a geração de boas perguntas em larga escala. Propõem, então, utilizar modelos de sumarização de textos em alemão para essa tarefa, comparando alguns utilizados na literatura por meio da avaliação de 30 professores da área. Desta forma, o algoritmo *LexRank* apresentou os melhores resultados de desempenho quanto à legibilidade do conteúdo, podendo assim ser utilizado como parâmetro na geração de questões.

Li e Xing (2021) propõem um método de geração de linguagem natural para apoiar os alunos de MOOC. Eles usaram o modelo GPT-2 (*Generative Pretrained Transformer*

Figura 1 – Interface da aplicação *mobile* com textos resumidos.



Fonte: (YANG et al., 2013)

2) (RADFORD et al., 2019) para fornecer aos alunos suporte informativo, emocional e comunitário com geração de linguagem natural em fóruns de discussão. Em experimentos, eles mostraram que o modelo GPT-2 poderia fornecer respostas contextuais e de suporte de forma semelhante em comparação com os humanos.

Hooshyar et al. (2018) propõem uma abordagem PCG (*Procedural Content Generation*) baseada em dados que se beneficia de um algoritmo genético e SVM (*Support Vector Machines*) para gerar automaticamente conteúdos de jogos educativos adaptados às habilidades dos indivíduos. Em experimentos, eles mostraram que os usuários obtiveram ganhos de desempenho mais significativos ao jogar conteúdo gerado sob medida para suas habilidades do que jogar conteúdo de jogo não personalizado.

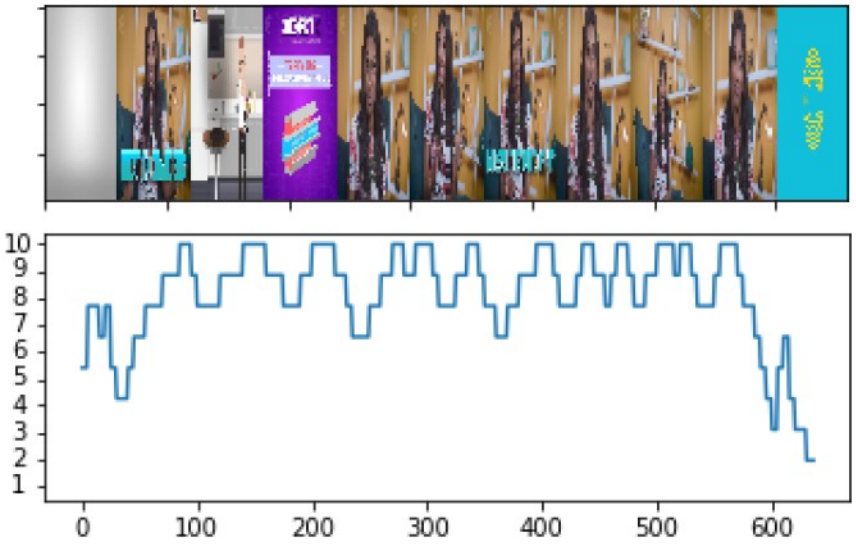
Xu, Smeets e Bidarra (2021) apresentam uma abordagem genérica para geração procedural de problemas matemáticos. Seu processo de geração consiste em duas fases: geração de problemas matemáticos abstratos e geração de texto. Para a geração de problemas matemáticos abstratos, eles propõem um método genérico baseado em modelo que opera em vários níveis e domínios de dificuldade. Além disso, para geração de texto, eles propõem um pipeline de geração de conteúdo textual adaptável em vários idiomas para transformar os problemas matemáticos abstratos gerados em perguntas de texto semanticamente coerentes em linguagem natural. Em experimentos com especialistas, eles mostraram que os problemas matemáticos gerados por sua abordagem são sensíveis e solucionáveis para alunos do ensino fundamental.

Os trabalhos citados nesta seção apresentam como característica a geração automática de OA. No entanto, poucos deles utilizam modelos *Deep Learning* baseado em texto. Desta forma a utilização de modelos *Transformers* surge como uma alternativa para um novo tipo de geração de OA utilizando sumarizadores de texto automáticos.

2.2 Sumarização de Vídeos Educacionais

Ghauri, Hakimov e Ewerth (2020) apresentam uma abordagem de anotar segmentos importantes em vídeo educacionais, para isso eles apresentam um ferramenta de anotação e um novo *dataset* denominado EDUVSUM (*Educational Video Summarization*) com vídeos educacionais, anotados com a importância de cada segmento por um especialista na área de computação, coletados de plataformas populares de *e-learning*, a Figura 2 apresenta um exemplo da classificação por notas de 1-10 dos segmentos de um vídeo do conjunto EDUVSUM. Além disso, apresentam uma arquitetura neural multimodal para a classificação de importância de segmentos a partir de características de estado da arte visuais, textuais (legenda) e de áudio, investigando o impacto de diferentes combinações dos mesmos.

Figura 2 – Vídeo exemplo com as anotações de importância para cada segmento.



Fonte: (GHAURI; HAKIMOV; EWERTH, 2020)

Alrumiah e Al-Shargabi (2022) apresentam uma abordagem de geração de resumos textuais de vídeo-aulas através da aplicação de LDA (*Latent Dirichlet Allocation*), os resumos textuais foram aplicados ao *dataset* EDUVSUM para comparação com a técnica ROGUE (Recall-Oriented Understudy for Gisting Evaluation) e resumos feitos por humanos. O resultado utilizando LDA supera os resultados utilizando TF-IDF (*Term Frequency – Inverse Document Frequency*) e LSA (*Latent semantic analysis*). O trabalho também contribui expandindo o dataset EDUVSUM com resumos textuais para cada vídeo educacional.

Abhilash et al. (2021) em seu trabalho apresentou uma forma de sumarização de vídeos educacionais através da transcrição da fala utilizando sumarização de texto extrativa. Para isso o autor coletou vídeos educacionais de plataformas e tratou a transcrição do texto

removendo *stop words* e pontuações. Utilizando TF-IDF o autor calculou pontuações para cada sentença da legenda e normalizou para o tamanho de cada sentença. As pontuações acima da média eram mantidas, diminuindo o tempo dos vídeos para cerca da metade do tempo. Utilizando a ferramenta ROUGE o autor atingiu resultados de medida-F de 0.805 comparando com o resumo gerado por especialistas.

Outros trabalhos utilizam os *datasets* TvSum (SONG et al., 2015) e SumMe (GYGLI et al., 2014) para a sumarização de vídeos, como por exemplo Ghauri, Hakimov e Ewerth (2021), Zhu et al. (2022) e Apostolidis et al. (2022) que são os modelos de *benchmark* para os mesmos, no entanto esses datasets, embora sejam variados em categorias, duração e tipos de imagem, não são adequados para domínio educacional. Os vídeos educacionais possuem como principal característica o que está sendo falado ou escrito, o que se torna raro nesses *datasets* ou não estão nem presentes em alguns casos.

Outras formas de sumarizar vídeos são evidenciando os *frames* principais. Meng, Yang e Gong (2022) em seu trabalho apresentam uma abordagem de sumarização de vídeos educacionais leve para boa performance em aplicações de tempo real. Considerando vídeos de apresentações de PowerPoint, o autor utiliza do desvio padrão e variância de *frames* nos vídeos para localizar *frames* chave, em seguida o grau de transição subjetiva em vídeo é usado para marcar a importância do conhecimento de cada *frame* chave. Os experimentos realizados demonstraram que a abordagem consegue localizar os *frames* chave de um vídeo educacional que estão alinhados ao tema.

3 Fundamentação Teórica

Neste Capítulo, é descrita a fundamentação teórica das redes neurais e modelos de *Deep Learning* utilizadas nos experimentos da dissertação. A Seção 3.1 e 3.2 apresenta os conceitos de Redes Neurais Artificiais e *Deep Learning* enquanto que a Seção 3.3 apresenta as redes baseadas nos modelos Transformers.

3.1 Redes Neurais Artificiais

De acordo com Haykin (2007), uma rede neural é um computador altamente distribuído, feito de unidades de processamento simples. O mesmo é semelhante ao cérebro, pois suas informações são adquiridas pela natureza através do processo de aprendizagem e a força de comunicação entre os neurônios é utilizada para armazenar as informações obtidas, essa força é chamada de pesos sinápticos.

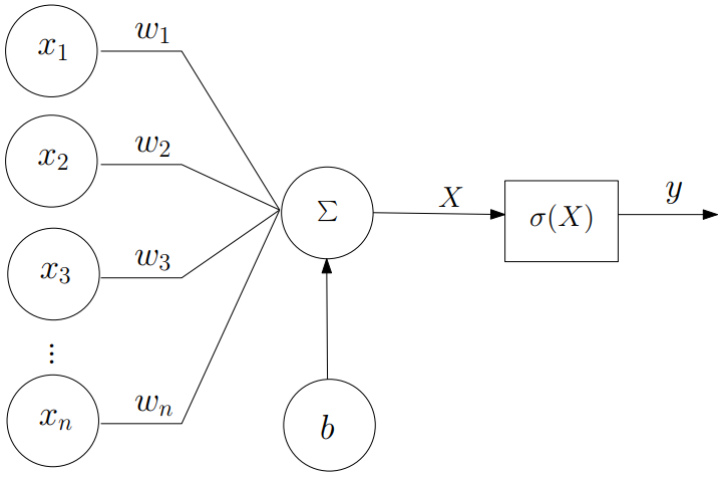
Outra definição de redes neurais artificiais é dada por Gurney (1997), que a define como um conjunto conectado de nós cuja função é baseada em redes neurais biológicas. O processamento de rede funciona conectando esses nós usando pesos que são obtidos e modificados por padrões de treinamento, esse processo é chamado de aprendizado.

As redes neurais artificiais foram desenvolvidas como uma generalização de modelos matemáticos de sistemas nervosos biológicos. Seus nós são chamados de neurônios, e a modelagem matemática dos seus elementos são as sinapses que são representadas como os pesos nas conexões dos nós que modulam o sinal de entrada do neurônio, a característica não linear dos neurônios biológicos é representada por uma função de transferência (ABRAHAM, 2005).

A Figura 3 mostra a arquitetura básica de uma rede neural artificial. É possível notar que um neurônio possui diversas características a partir de sinais de entrada (x_i), que são valores numéricos transmitidos pelos pesos sinápticos de outros neurônios ou pelo sinal inicial. Esses valores são ponderados por pesos (w_i) que devem ser somados a um valor chamado *bias* b. A saída deste neurônio pode ser a entrada de um novo neurônio ou é o valor final podendo ser utilizado em uma função de ativação (σ) que produz uma nova saída (y) (SOUSA, 2018).

A partir da Figura 3 é possível modelar uma equação matemática do funcionamento de um neurônio artificial, a Equação 3.1 apresenta a modelagem desta equação onde os valores x_i, y, w_i, b e φ são, respectivamente, os sinais de entrada e saída do neurônio, o peso a ser aplicado no valor de entrada, o valor *bias* a ser adicionado e a função de ativação no neurônio.

Figura 3 – Arquitetura de uma rede neural artificial.

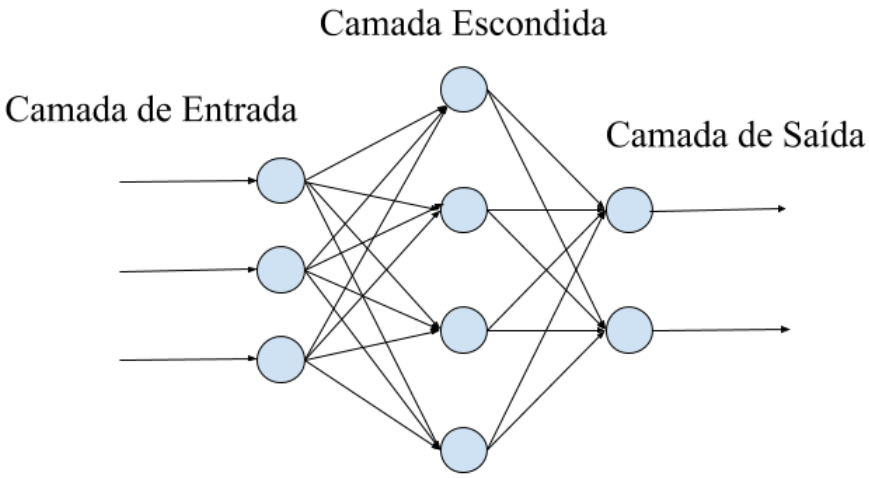


Fonte: (KOVALESKI, 2018).

$$y = \varphi(\sum_{i=1}^{n} x_i w_i + b) \tag{3.1}$$

Um neurônio resolve apenas problemas linearmente separáveis, para os problemas não lineares é necessário associar dezenas de neurônios interconectados em várias camadas para resolver o problema e aumentar a robustez (ABRAHAM, 2005). Esses modelos são denominados de *Multilayer Perceptron* (MLP), a Figura 4 retrata um exemplo de um MLP que possui três tipos de camadas, a camada de entrada em que os sinais codificados são inseridos, a camada escondida onde os dados são processados e transferidos para a camada seguinte, a camada de saída, da qual resulta a classificação final para o sistema (SOUSA, 2018).

Figura 4 – *Multilayer Perceptron*.



Fonte: O autor, adaptado de (ABRAHAM, 2005).

O treinamento das redes neurais para resolver uma tarefa específica se dá pela atualização do valor dos pesos e do *bias*, pois os mesmos são os responsáveis por ditar a saída da rede para determinado problema. O principal método para treinamento e, portanto, a atualização dos pesos é o algoritmo *back-propagation* (RUMELHART; HINTON; WILLIAMS, 1986).

3.1.1 *Back-Propagation* e Método do Gradiente Descendente

De acordo com MARUMO (2018) o *back-propagation* é um algoritmo supervisionado para o treinamento principalmente de redes neurais artificiais, podendo ser utilizado em outros modelos auto-regressivos. Em sua primeira fase, para frente (*forward*), são gerados valores de saída da rede para um dado sinal de entrada, e na segunda fase, para trás (*backward*), são utilizados esses valores de saída e os valores desejados (a anotação original para o sinal de entrada) para calcular o erro e minimizar o mesmo a partir da atualização dos pesos sinápticos. Os erros são calculados por uma função de custo, onde a saída da rede é comparada com a saída real e os ajustes podem ser feitos utilizando algoritmos de otimização, como por exemplo o método do gradiente descendente(BRAGA, 2000).

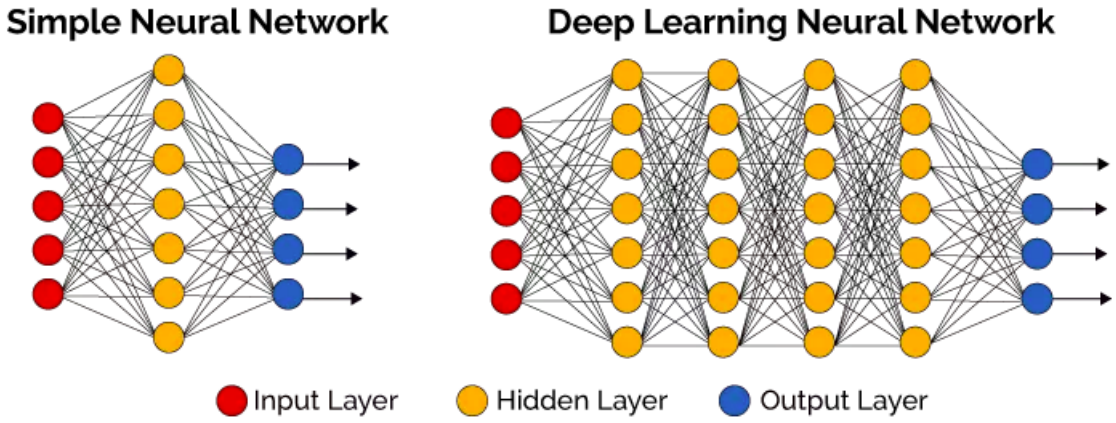
O método do gradiente descendente busca encontrar o valor de mínimo local, visto que este valor é diretamente proporcional ao valor do erro. Assim o algoritmo *back-propagation* procura minimizar o erro obtido pela rede ajustando pesos e limiares para que eles correspondam às coordenadas dos pontos mais baixos da superfície de erro, essa direção é encontrada calculando a derivada do função de custo indicando a direção em que a mesma decresce (BRAGA, 2000). A função de custo determina o peso dos erros, é ela que define a forma de penalizar os resultados preditos que se distanciam do valores reais.

3.2 Deep Learning

Deep Learning ou Aprendizagem Profunda é a denominação para redes neurais mais robustas, ou seja, com muitas camadas intermediarias. O *Deep Learning* se baseia em aprender diferentes níveis de abstração de uma tarefa, utilizando para isso várias camadas adicionais para processamentos não-lineares (DENG; YU, 2014).

Com isso, *Deep Learning* trata-se de um aprendizado de redes neurais com arquiteturas com várias camadas intermediárias conseguindo extrair melhor as características de um conjunto de dados, no entanto essa robustez de vária camadas acaba incidindo em um tempo de processamento maior que das redes neurais clássicas. A Figura 5 apresenta uma comparação de uma rede neural simples com uma *Deep Learning*.

Figura 5 – Duas arquiteturas de redes neurais: A esquerda uma rede neural simples e a direita uma rede de *Deep Learning*.



Fonte: (Favio Vázquez, 2017)

De acordo com Deng e Yu (2014) existem três principais razões para a popularidade da *Deep Learning* atualmente:

1. O aumento drástico do poder computacional por unidades de processamento gráficos (GPU), aumentando a capacidade computacional através do paralelismo dos cálculos.
2. O aumento significativo do tamanho dos dados utilizados para treinamento
3. Os recentes avanços em aprendizado de máquina e o processamento de sinais que facilitaram a construção de redes mais complexas realizando tarefas que antes não eram possíveis.

A seção a seguir apresenta um tipo de rede neural profunda que é utilizado nos experimentos realizados neste trabalho para o processamento de linguagem natural.

3.3 Redes Transformers

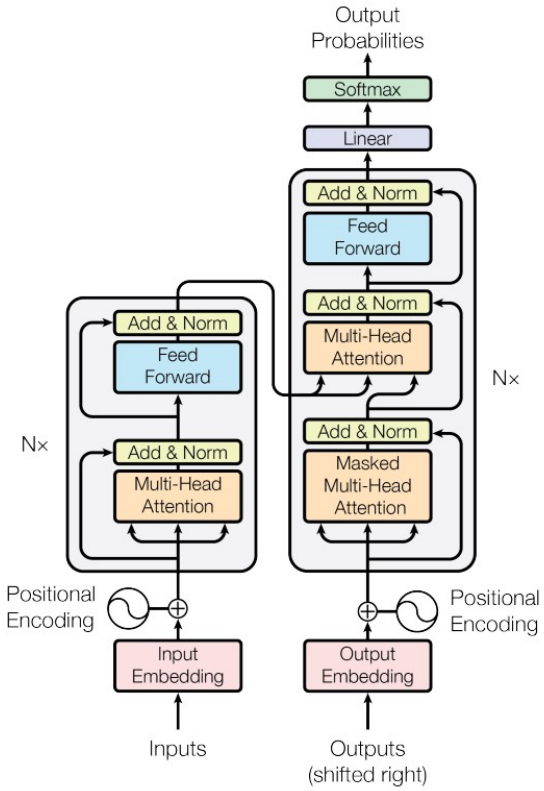
Em problemas do tipo *sequence-to-sequence* (A entrada é uma sequência e a saída é outra sequência), como no caso do processamento de linguagem natural na tarefa de traduzir ou resumir textos, existe a necessidade de transformar os textos em uma representação numérica para aplicação de diversas técnicas de aprendizagem de máquina. Como esse tipo de tarefa possui variáveis dependentes entre si (como por exemplo uma sentença textual, onde as palavras possuem relações entre si), as redes neurais mais utilizadas são as Redes Neurais Recorrentes (RNN), que são um tipo de rede com um *loop* para que a informação persista. Associado com codificadores (que produz uma saída única para a sequência de entrada) e decodificadores (que utiliza a saída do decodificador para produzir uma nova

sequência) se tornou possível resolver problemas do tipo *sequence-to-sequence* No entanto essas redes, além de serem difíceis de treinar, tinham uma grande limitação para lidar com longas sequências, a habilidade de reter elementos iniciais era perdida quando inserido novos elementos na rede (Eduardo Muñoz, 2020).

Desta forma, surgiu um novo modelo de rede: o modelo *Attention* (Atenção), que ao invés de olhar apenas para os últimos elementos da sequência (como no caso das RNN) em cada passo do decodificador que forma a sentença de saída é avaliado o estado correspondente no codificador. Desta forma o modelo *Attention* consegue extrair as partes mais importantes de uma sentença antes de produzir uma nova. Ainda sim essa abordagem possui a limitação de precisar processar um elemento por vez de uma sequência, aumentando muito o custo computacional (Eduardo Muñoz, 2020).

A rede *Transformers* surgiu para resolver esse problema. Vaswani et al. (2017) apresenta uma arquitetura de rede utilizando apenas o modelo *Attention*, eliminando as redes recorrentes da mesma, alcançando boa performance e diminuindo o tempo de treinamento. Desta forma, a arquitetura da rede se baseia em codificadores e decodificadores com modelos de *self-attention* calculando a importância dos elementos das palavras através de produto escalar simples. Isso é feito na sequência inteira de entrada como também em combinações de diferentes tamanhos da sequência (*Multi-Head Attention*). A Figura 6 apresenta a arquitetura da rede *Transformers*.

Figura 6 – Arquitetura da rede *Transformers*.



Fonte: (VASWANI et al., 2017)

Este uso diferenciado do modelo *Attention* possibilitou avanços na pesquisa na área de processamento de linguagem natural, pois além de serem mais rápidos de serem treinados os resultados também foram superiores às redes recorrentes (Eduardo Muñoz, 2020). Os modelos considerados estado da arte atualmente utilizam vários modelos *Transformers* internamente ou variantes do mesmo, como por exemplo o modelo BERT (DEVLIN et al., 2018).

3.3.1 BERT

BERT (*Bidirectional Encoder Representations from Transformers*) é um recente tipo de arquitetura de rede para tarefas de processamento de linguagem natural, como *Question-Answering*, inferência de palavras e resumo de textos. Sua premissa é de utilizar o treinamento bidirecional (da esquerda para direita de uma sequência e vice-versa) em redes *Transformers*. Desta forma, essa nova arquitetura de rede consegue aprender o contexto das palavras e inferir em momentos de ambiguidade (Rani Horev, 2018).

O objetivo principal de tarefas de processamento de linguagem natural é entender a linguagem da forma que ela é falada naturalmente, um grande diferencial do modelo BERT é que o mesmo foi treinado em um grande conjunto de dados sem rótulo (A página

da Wikipédia em inglês e outros *datasets*), assim o modelo consegue entender melhor o significado das palavras principalmente pelo contexto que elas se encontram, ou seja, as palavras que estão a sua volta (Ben Lutkevich, 2020). Desta forma, com a rede já pré-treinada é possível retreina-la para se adaptar a diversas tarefas específicas de linguagem natural, essa tarefa é denominada *Transfer Learning*.

3.3.2 DistilBERT

DistilBERT (*Distiled BERT*) é uma variação do modelo BERT com menos parâmetros. Considerando que os modelos de processamento de linguagem natural vem atingindo níveis de processamento elevados, colocar os mesmos em uma aplicação ou para operar em dispositivos móveis pode ser desafiador. O modelo DistilBERT surge com a premissa de ser uma rede pré-treinada com menos parâmetros que o BERT mas que alcance desempenho semelhante em uma grande variedade de tarefas em contra-parte (SANH et al., 2019).

O modelo DistilBERT utiliza da técnica *Knowledge distillation* (HINTON; VINYALS; DEAN, 2015) que se baseia em um modelo compacto - o estudante - é treinado especificamente para reproduzir o comportamento de um modelo maior - o professor. A ideia principal é treinar o modelo para que seu palpite generalize melhor as outras respostas, não focando na resposta principal, desta forma prevenindo o modelo de ter muita certeza em um problema de classificação. Desta forma os autores conseguiram mostrar que é possível diminuir o tamanho do modelo em 40% e mantendo uma taxa de 97% da sua capacidade de entendimento da linguagem além de ser 60% mais rápido que o modelo BERT (SANH et al., 2019).

3.3.3 RoBERTa

RoBERTa (*Robustly Optimized BERT-Pretraining Approach*), como o próprio nome diz, é uma otimização robusta do modelo BERT, tanto em parâmetros da rede como em *dataset* de treinamento. O time de Inteligência Artificial do *Facebook* em seu artigo Liu et al. (2019) citam que o modelo pré-treinado BERT foi de maneira significativa sub-treinado, além de ter algumas escolhas não muito boas de arquitetura. Desta forma, os autores criaram o modelo RoBERTa como uma extensão do modelo original BERT.

Uma das alterações do modelo RoBERTa foi no *dataset* em que o mesmo foi treinado. Observou-se que o modelo BERT quando treinado em *datasets* maiores incrementa drasticamente sua performance. Desta forma o modelo RoBERTa foi treinado em um conjunto vasto de dados de cerca de 160 GB de texto (LIU et al., 2019). O conjunto é composto pelos seguintes *datasets*:

- **BookCorpus (ZHU et al., 2015) + Wikipedia em Inglês**: Esse é o conjunto original utilizado no modelo BERT (16GB).

- **CC-NEWS**: Este foi coletado de uma porção em inglês do conjunto *CommonCrawl News* [1]. O *dataset* possui cerca de 63 milhões de notícias em inglês extraídas entre setembro de 2016 e fevereiro de 2019. (76 GB).
- **OpenWebText**: Recriação de código aberto do conjunto de dados WebText descrito em Radford et al. (2019). (38GB).
- **STORIES**: Um conjunto de dados introduzido em Trinh e Le (2018) contendo um subconjunto de *CommonCrawl News* filtrados para corresponder ao estilo de história de Esquemas de Winograd. (31 GB).

Além de ter sido treinado em um *dataset* vasto, o modelo RoBERTa foi treinado por mais tempo em comparação com o modelo BERT (LIU et al., 2019).

Trabalhos anteriores, como o de Ott et al. (2018) mostraram que os modelos Transformer e BERT são adequados para grandes tamanhos de treinamento em lote. Eles tornam a otimização mais rápida e pode melhorar o desempenho da tarefa final quando ajustado corretamente, desta forma o modelo RoBERTa foi treinado em lotes maiores, o que inclusive facilita o processo de treinamento em paralelo (LIU et al., 2019).

Outro ponto é a criação da máscara no pré-treinamento. O objetivo da criação da máscara no pré-treinamento do BERT é essencialmente mascarar alguns *tokens* (representação das palavras) de cada sequência aleatoriamente e, em seguida, prever esses *tokens*. No entanto, na implementação original do BERT, as sequências são mascaradas apenas uma vez no pré-processamento. Isso implica que o mesmo padrão de mascaramento é usado para a mesma sequência em todas as etapas de treinamento (Rohan Jagtap, 2020).

Para evitar isso, na re-implementação do BERT, os autores duplicaram os dados de treinamento 10 vezes para que cada sequência fosse mascarada em 10 padrões diferentes. Isso foi treinado por 40 épocas, ou seja, cada sequência foi treinada para os mesmos padrões de mascaramento 4 vezes (LIU et al., 2019). Esse tipo de máscara dinâmica também foi utilizado como padrão na implementação do modelo RoBERTa. Unindo todos esses pontos, o modelo RoBERTa, em tarefas de *benchmark*, conseguiu uma melhoria na performance de 2-20% a mais que o modelo BERT (Rohan Jagtap, 2020).

[1] https://commoncrawl.org/2016/10/newsdataset-available

4 Geração de OA Através de Sumarizadores de Texto

Neste capítulo, é descrito o processo de geração de objetos de aprendizagem a partir da sumarização automática de textos. Desta forma o capítulo está dividido em seções onde na Seção 4.1 tem-se o contexto da geração de objetos de aprendizagem. Já na Seção 4.2 tem-se a descrição da metodologia para a realização do trabalho e, por fim, na Seção 4.3 tem-se a descrição do trabalho realizado e a descrição dos resultados obtidos no mesmo.

4.1 Contexto

No campo da educação existem várias formas de disseminar conhecimento, desde as mais tradicionais, como o ensino presencial com tutores em sala de aula, até as que utilizam de tecnologia para passar conhecimento ao estudante. Existem recursos digitais como texto, vídeo, *software*, entre outros que podem ser usados e reutilizados para educação. Esses recursos são denominados Objetos de Aprendizagem (BRAGA; MENEZES, 2014).

Atualmente, a criação desses objetos de aprendizagem se trata de um processo manual e que exige a presença de um especialista, o que pode transformar todo o processo oneroso. Para resumir um texto, o profissional além de precisar ter um mínimo de conhecimento na área necessita ler todo o texto do qual precisará resumir. Uma solução seria a geração desses objetos de aprendizagem através de sumarizadores de textos devido à capacidade dos mesmos de sintetizar um conteúdo, tanto para um tutor como para um estudante, facilitando o entendimento do assunto, a geração de questões e até a resolução de questões sobre determinado assunto.

Alguns autores já utilizaram de modelos de geração de texto para criação de objetos de aprendizagem. Liu et al. (2012) utilizou textos provenientes da Wikipedia e grafos estruturados para a geração semi-automática de questões para auxiliar na escrita acadêmica. No entanto, pouco se encontra quanto a geração de textos resumidos com conteúdo educativo relativos a um certo tema.

Esta etapa do trabalho tem como objetivo utilizar sumarizadores de textos baseados em modelos de *Deep Learning* como solução para a geração automática de objetos de aprendizagem. Almeja-se, através da avaliação de especialistas, tutores e estudantes, estimar a viabilidade do uso de sumarizadores para a geração de textos com conteúdo e gramática aceitáveis. Como contribuição, deseja-se também mostrar a viabilidade do uso de tradutores automáticos para a realização do procedimento na língua portuguesa, devido aos sumarizadores serem específicos para a língua inglesa.

4.2 Método

O objetivo, nesta etapa do projeto, é verificar a viabilidade de utilizar sumarizadores para a geração de objetos de aprendizagem, dessa forma como método deseja-se gerar resumos de textos em português provenientes da Wikipédia. No entanto, existem alguns fatores limitantes para utilizar esta abordagem, como a falta de sumarizadores para a língua portuguesa, o que cria a necessidade de utilizar tradutores automáticos na implementação. Outro fator é o limite de palavras no dicionário dos modelos de sumarização, exigindo que se divida o texto em várias partes para gerar pequenos resumos. Há ainda outro fator limitante que é a necessidade de formatação inicial do texto oriundo da Wikipédia, como descrição de imagens, referências e links fora de contexto com o conteúdo.

Assim, o algoritmo proposto para geração de objetos de aprendizagem é ilustrado na Figura 7, e seus passos são descritos a seguir:

1. Buscar, através de API (*Application Programming Interface*), um conteúdo educacional da página da Wikipédia sobre determinado tema em português.
2. O conteúdo passa, então, por um processo de limpeza para manter apenas a parte textual, removendo caracteres especiais da Wikipédia, referências de outros artigos, descrição de imagens e transformação de listas em texto amplo.
3. O texto agora limpo passa por uma etapa de estruturação, particionando o mesmo em pequenos textos de no máximo 1100 caracteres, assim evitando que o sumarizador extrapole sua capacidade de geração de modelo.
4. Cada partição é traduzida automaticamente para a língua inglesa (na qual os sumarizadores operam) e, por fim, passa pelo sumarizador, que gera um modelo e resume cada partição em um texto menor.
5. A última etapa apenas junta as partições e realiza a tradução de volta para a língua portuguesa.

Para realizar as traduções foi utilizada a biblioteca para Python *googletrans*[1] que implementa uma API para o tradutor da Google. Os sumarizadores são originários da plataforma HuggingFace [2] que é uma plataforma *Open-Source* com vários modelos de geração de texto utilizando *Deep Learning* e processamento de linguagem natural.

Para verificar a viabilidade do uso de sumarizadores como objetos de aprendizagem, um tema da área de computação foi selecionado com o intuito de fazer uma avaliação da qualidade de conteúdo educacional dos mesmos e da qualidade da gramática do texto

[1] https://pypi.org/project/googletrans/
[2] https://huggingface.co/

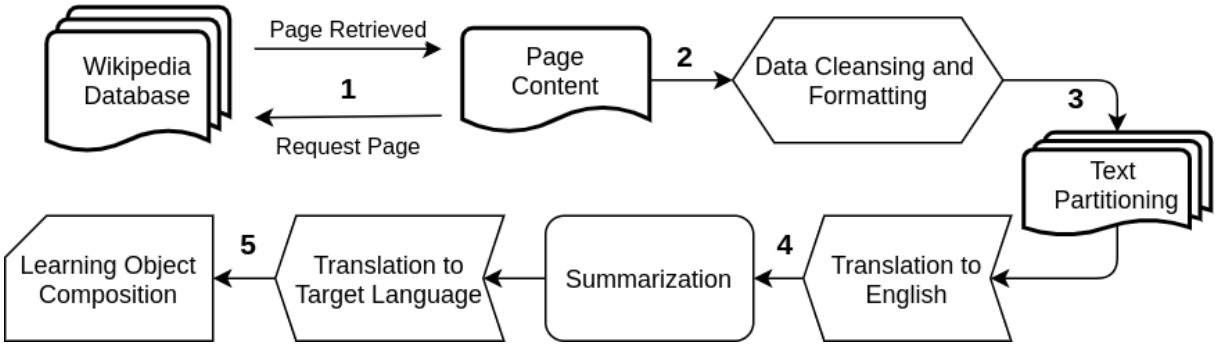


Figura 7 – Etapas da método proposto de geração de objetos de aprendizagem

resumido. O tema selecionado foi da página Software da Wikipedia [3], de tal forma que o texto original está dividido em vários tópicos e cada tópico possui algumas listas de itens, sendo assim um bom estudo de caso para realizar o resumo.

Na plataforma HuggingFace foram selecionados os três modelos *Deep Learning* de sumarizadores mais utilizados pelos usuários para resumir os textos para fins de comparação.

- **Modelo 1 - *sshleifer/distilbart-cnn-12-6***[4]: Modelo de sumarização desenvolvido por Sam Shleifer que utiliza a versão destilada do modelo BART (LEWIS et al., 2019), com 174 mil downloads era o mais baixado no momento da realização dos experimentos.
- **Modelo 2 - *facebook/bart-large-cnn***[5]: Modelo de sumarização desenvolvido pela equipe do Facebook que utiliza o modelo BART (LEWIS et al., 2019), com 139 mil downloads era o segundo mais baixado no momento da realização dos experimentos.
- **Modelo 3 - *google/pegasus-cnn_dailymail***[6]: Modelo de sumarização desenvolvido pela equipa de IA da Google utilizando o modelo PEGASUS (ZHANG et al., 2020), com 55 mil downloads era o terceiro mais baixado no momento da realização dos experimentos.

4.3 Experimento

Para verificar a viabilidade do uso de sumarizadores como objetos de aprendizagem, foi construído um formulário contendo o resumo de cada tópico da página da Wikipedia, a partir de cada um dos três sumarizadores citados, desta forma foram gerados 15 textos no

[3] https://pt.wikipedia.org/w/index.php?title=Softwareoldid=60974962
[4] https://huggingface.co/sshleifer/distilbart-cnn-12-6
[5] https://huggingface.co/facebook/bart-large-cnn
[6] https://huggingface.co/google/pegasus-cnn$_dailymail$

total. O formulário consiste inicialmente no usuário autorizar o termo de consentimento livre e esclarecido, depois preencher algumas informações pessoais, especialmente sobre sua formação. Segue o formulário com a avaliação de cada um dos 15 textos resumidos, tanto analisando seu conteúdo educacional quanto a sua gramática. Essas avaliações utilizam uma nota de 1 a 5, sendo 1 um texto mais pobre e 5 de mais qualidade. Em cada etapa foi apresentado o texto original com o resumo gerado por cada sumarizador de forma comparativa. O formulário não apresenta de qual algoritmo cada resumo pertence, para impossibilitar o viés dos participantes, além de serem apresentados em ordem aleatória.

Outra pergunta feita aos voluntários da pesquisa foi ao final de cada seção, qual dos três textos, no critério de cada um, era comparado como o melhor e o pior texto entre os três. Para cada uma das questões do formulário, o voluntário tinha a opção não-obrigatória de justificar a sua avaliação, em texto livre. O público alvo dessa pesquisa são os estudantes de cursos na área da computação, formados e pós-graduados em computação e professores da computação, devido ao conteúdo extraído ser relacionado a área da computação.

4.3.1 Resultados

Como primeiro resultado da pesquisa pelo formulário, tem-se na Figura 8 o gráfico da avaliação dos estudantes dos cursos de computação e áreas afins. Foram coletadas respostas de 12 estudantes que ainda não estão formados na área. A Figura apresenta a média da nota de cada um, junto com o intervalo de confiança pela variância.

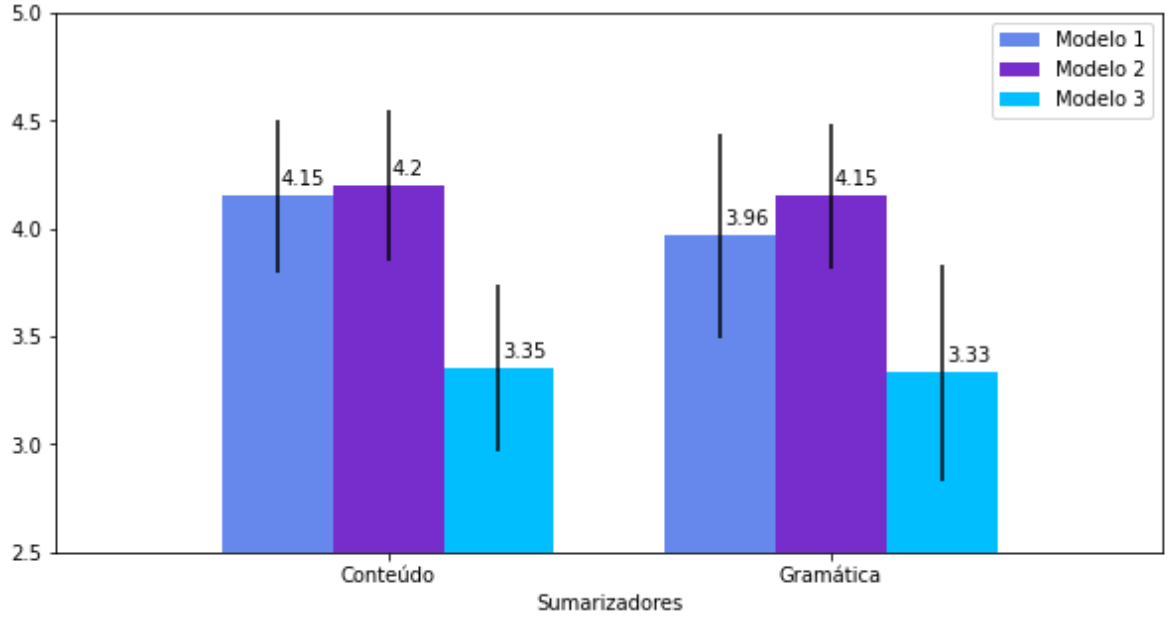


Figura 8 – Avaliação de estudantes quanto a qualidade dos textos gerados. Na esquerda, a avaliação de conteúdo e na direita, a avaliação de gramática

Nota-se que quanto ao conteúdo, os sumarizadores do Modelo 1 e do Modelo

2 obtiveram notas e variância semelhantes, tendo um desempenho melhor que o do Modelo 3 de acordo com os estudantes participantes, que obteve média de $3,35$. Quanto à qualidade da gramática dos sumarizadores, os estudantes elegeram o Modelo 2 como melhor sumarizador de conteúdo educativo, sobressaindo no desempenho comparado com os outros dois.

A justificativa dos estudantes quanto a nota de cada sumarizador, nota-se que o sumarizador do Modelo 3 obteve um desempenho inferior devido que algumas partes do texto estão desconexas e sem sentido, além de possuírem strings "<N>" soltas pelo texto, que poderiam ser suprimidas. Enquanto que os textos do Modelo 1 e do Modelo 2 tiveram menos reclamações no campo de justificativa, comparado com o do Modelo 3. A Tabela 1 e 2 apresenta a votação de cada estudante sobre qual seria o melhor e o pior sumarizador de conteúdo. É possível notar que em quase todos os casos, os estudantes elegeram o Modelo 2 como o melhor texto. Apenas no texto 3, o Modelo 1 teve um desempenho superior aos demais.

Tabela 1 – Votação do melhor texto pelos estudantes.

	Texto 1	**Texto 2**	**Texto 3**	**Texto 4**	**Texto 5**	**Total**
Modelo 1	3	3	**8**	5	3	22
Modelo 2	**6**	**7**	4	**6**	**8**	**31**
Modelo 3	3	2	0	1	1	7

Tabela 2 – Votação do pior texto pelos estudantes.

	Texto 1	**Texto 2**	**Texto 3**	**Texto 4**	**Texto 5**	**Total**
Modelo 1	2	3	2	2	5	14
Modelo 2	3	0	0	0	1	4
Modelo 3	**7**	**9**	**10**	**10**	**6**	**42**

Quanto à avaliação de textos com qualidade inferior, houve unanimidade para eleger os textos sumarizados pelo Modelo 1. No entanto, as justificativas dadas pelos alunos foram poucas, em sua maioria se referindo ao texto ser confuso. Interessante notar que a Tabela 2 também elege com sobra os textos do Modelo 2 como os melhores. Observando as justificativas dos alunos, o motivo parece ter sido o que menos errou gramaticalmente e teve uma leitura mais fluida para os estudantes de computação.

O segundo público utilizado na pesquisa é de pessoas formadas em cursos de computação, alunos de pós-graduação e professores de áreas afins à computação. Totalizando 13 respostas, os resultados da média da avaliação encontram-se na Figura 9.

Inicialmente, de acordo com a Figura 9, nota-se que as avaliações do Modelo 1 e do Modelo 2 decresceram comparada com as notas atribuídas pelos estudantes de cursos da computação, enquanto a nota, tanto do conteúdo como da gramática, do Modelo 3 cresceu

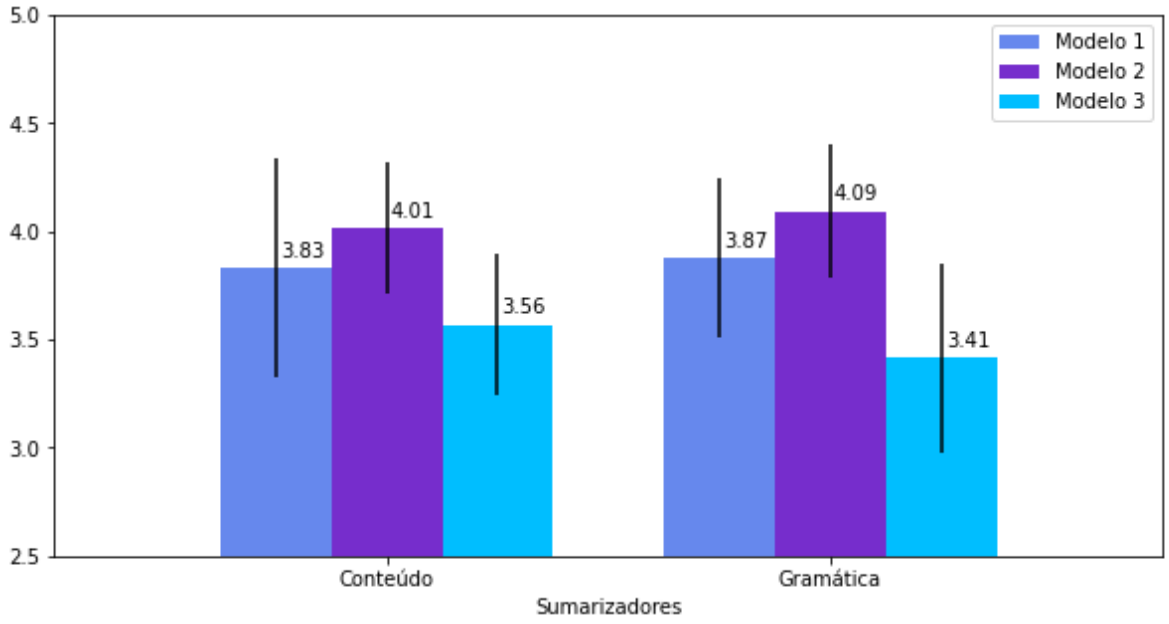


Figura 9 – Avaliação de graduados quanto a qualidade dos textos gerados. Na esquerda, a avaliação de conteúdo e na direita, a avaliação de gramática.

pouco, embora ainda se manteve, de acordo com os graduados, com desempenho inferior comparado aos outros sumarizadores.

O mesmo padrão da Figura 8 se manteve na avaliação da Figura 9, o sumarizador do Modelo 2 foi o mais bem avaliado e, diferente do anterior, teve valores menores de variância, indicando uma concordância dos avaliadores quanto ao conteúdo e a gramática, desta forma se manteve com média próxima a 4. Diferente das justificativas dos alunos, aquelas dadas pelos graduados possuíam mais elogios quanto à qualidade do conteúdo do texto e da gramática. A Tabela 3 apresenta algumas das justificativas das avaliações dadas pelos graduados e alunos de computação.

Houve nas justificativas novamente uma quantidade de avaliadores que se referiram ao problema do "<N>" que apareciam na maioria dos textos do Modelo 3. Provavelmente o principal fator da nota baixa na avaliação desse sumarizador se deu por esse fato, porém o mesmo pode ser resolvido com um simples pós-processamento do texto de saída do mesmo.

A Tabela 4 e 5 apresenta a votação de cada graduado e professor da computação de qual seria o melhor e o pior sumarizador de conteúdo, de acordo com suas opiniões pessoais. É fácil verificar novamente nessa Tabela a preferência dos avaliadores quanto ao texto do Modelo 2, lembrando que os textos estavam randomizados e sem descrição de qual era o sumarizador.

Interessante notar que novamente o Modelo 2 foi a preferência como melhor

Tabela 3 – Justificativa das avaliações dos candidatos sobre cada sumarizador.

	Alunos	**Professores e Graduados**
Modelo 1	Pequenos erros de concordância e o uso de termos não usados geralmente como "Idiomas de programação"	Qualidade gramatical muito boa
	Duplicidades são desnecessárias.	Para alguém que não é estudante da área o texto mostra-se confuso e com pedaços de ideias que não se comunicam.
Modelo 2	Uso incorreto do ponto final do meio da frase: "análise econômica, análise de requisitos. Especificação(...)"	Em termos gramaticais o texto está muito bom
	O primeiro parágrafo restringe uma definição que poderia ser mais abrangente e completa.	A primeira parte do texto não está clara e parece não fazer muito sentido
Modelo 3	Não é tão claro quanto os textos anteriores, mas informa o necessário	Presença de expressões sem sintaxe (). Refere-se à uma sequência "por esta sequência [...]", no primeiro parágrafo, mas não especifica qual.
	O <N>é confuso	O texto demonstra clareza no repasse da informação

Tabela 4 – Votação do melhor texto pelos graduados e professores.

	Texto 1	**Texto 2**	**Texto 3**	**Texto 4**	**Texto 5**	**Total**
Modelo 1	**6**	**7**	4	2	4	23
Modelo 2	**6**	6	**7**	**9**	**8**	**36**
Modelo 3	1	0	2	2	1	6

Tabela 5 – Votação do pior texto pelos graduados e professores.

	Texto 1	**Texto 2**	**Texto 3**	**Texto 4**	**Texto 5**	**Total**
Modelo 1	3	2	2	4	5	16
Modelo 2	4	1	4	1	1	11
Modelo 3	**6**	**10**	**7**	**8**	**7**	**38**

sumarizador. No entanto, para os professores e graduados, houve empate e também preferência pelo Modelo 1 no texto 1 e 2, algo que foi diferente do que foi mostrado pelos estudantes.

Durante a pesquisa, algumas pessoas tiveram problemas para responder o formulário, levando mais que o tempo necessário para finalizar. Com isso, algumas pessoas não responderam os últimos textos com a mesma atenção que deram para os textos iniciais. Dessa forma, pode-se dizer que em alguns pontos o Modelo 1 e o Modelo 2 tiveram desempenho bem similar, devido a pouca diferença nos três primeiros textos.

Era esperado que os resultados do Modelo 2 fossem superiores ao do Modelo 1, haja vista que o Modelo 1 utiliza o DistilBART, uma versão destilada, ou seja mais leve e com menos parâmetros, do modelo BART, o que é utilizado no Modelo 2. Dessa forma, pela rede de aprendizagem do Modelo 2 ser mais robusta, a mesma consegue gerar textos mais sucintos, embora necessite de mais tempo de processamento. O Modelo 3 utiliza o PEGASUS, que consegue obter bons desempenhos mesmo que treinado com poucos exemplos. Contudo, ele também não é tão robusto como o modelo BART, o que pode ser um dos motivos do Modelo 3 ter ficado abaixo na avaliação em todos os requisitos do formulário.

Embora os textos precisassem passar por duas etapas de tradução automática, os sumarizadores foram capazes de disseminar o conteúdo educativo de maneira satisfatória em alguns casos. Assim, este trabalho começa a trilhar um caminho para a geração de objetos de aprendizagem mais sofisticados através de uma estruturação por tópicos ou ainda trabalhar com o pós-processamento na saída desses sumarizadores.

5 Sumarização de vídeos educacionais

Neste capítulo, é descrito o processo da sumarização de vídeos educacionais através de sumarizadores de texto. Desta forma, o mesmo está dividido em seções onde na Seção 5.1 tem-se o contexto da geração de objetos de aprendizagem, na Seção 5.2 tem-se a descrição da metodologia para a realização do trabalho e, por fim, na Seção 5.3 tem-se a descrição dos resultados obtidos da metodologia descrita.

5.1 Contexto

Com o avanço da tecnologia e da facilidade de acesso das pessoas, a Web vem testemunhando o crescimento de vídeos educacionais. Vídeos educacionais se apresentam de diversas formas, com diferentes durações, conteúdo e tipo de apresentação, podendo ser a gravação de uma aula, uma apresentação em *PowerPoint* ou outros diversos formatos de vídeo.

Com o aumento da quantidade de vídeos disponíveis, alguns de longa duração, observou-se um padrão de comportamento dos usuários, muitos evitam de assistir um vídeo inteiro e "pulam"para o instante do vídeo de seu interesse, esse comportamento é denotado como *skim through* (POTAPOV et al., 2014). Este comportamento, embora específico, denota a importância de termos formas de “resumir” ou anotar a importância de segmentos em um vídeo educacional de forma a poupar tempo do espectador que deseja encontrar um momento chave de algum vídeo.

A sumarização de vídeos pode ser definida como a conversão de um vídeo longo (de acordo com o contexto) em um vídeo mais curto que contém os segmentos essenciais e, desta forma, permite um espectador entender o conteúdo do vídeo e menor tempo (GHAURI; HAKIMOV; EWERTH, 2021). A escolha dos segmentos para o resumo pode ser feita através da anotação da importância, em valores numéricos, de cada segmento do vídeo em relação ao conteúdo abordado no vídeo. O processo de sumarização pode ser complexo e de diferentes abordagens, alguns utilizam características visuais (SONG et al., 2015), outros utilizam histórico de exibição e tempo de visualização por segmento (MUBARAK; CAO; AHMED, 2021), tem-se também o uso de diversas características como áudio, vídeo, e transcrição da fala (GHAURI; HAKIMOV; EWERTH, 2020).

No contexto educacional, algo que se destaca em vídeos é a fala do orador. Normalmente esse tipo de vídeo é a gravação de uma aula ou explicação de algum tópico, assim, esses vídeos possuem em quase sua totalidade de tempo a fala do orador e, consequentemente, alguns desses vídeos possuem a transcrição da fala em legendas.

Essas legendas são anotadas com seu tempo de surgimento e duração, evidenciando segmentos existentes em um vídeo educacional. Considerando isso, temos na literatura formas de resumir textos de maneira eficiente utilizando *Deep Learning* através de redes *Transformers* (OLIVEIRA et al., 2021), no Capítulo 4 mostrou-se que é possível compor objetos de aprendizagem através de resumo de textos educacionais mantendo o conteúdo mais importante.

Desta forma é possível correlacionar o resumo da transcrição do texto com os segmentos do vídeo educacional para anotar sua importância quanto ao conteúdo. Alguns trabalhos apresentam uma sumarização de vídeo utilizando legenda para a geração de resumos textuais (ALRUMIAH; AL-SHARGABI, 2022), tem-se também trabalhos que realizam a sumarização do vídeo a partir da transcrição do texto através da técnica TF-IDF (*Term Frequency - Inverse Document Frequency*) (ABHILASH et al., 2021)

Esta etapa do trabalho tem como experimento verificar a viabilidade do uso de sumarizadores de texto, utilizando redes de *Deep Learning*, para a sumarização de vídeos educacionais através da anotação da importância dos segmentos de cada legenda.

5.2 Método

Este trabalho tem como objetivo verificar a viabilidade de utilizar sumarizadores de texto para a tarefa de sumarização de vídeos educacionais através das legendas, com isso anotando os segmentos importantes em um vídeo e compondo um objeto de aprendizagem. O método proposto para alcançar o resultado se baseia na correlação da legenda original do vídeo e o texto resumido gerado pelo sumarizador. No entanto existem alguns fatores limitantes como a necessidade dos vídeos terem legenda disponível, o vídeo educacional precisa ser focado na fala do interlocutor para que tenha o que transcrever além de possuir poucos momentos sem fala pois nestes instantes não teremos informação para avaliar o contexto do vídeo.

O método proposto para sumarização de vídeos educacionais é ilustrado nos passos a seguir:

1. Obter a legenda original da fala (no idioma inglês) do vídeo a ser sumarizado com seu *timestamp* anotado.
2. O conteúdo da legenda é unido em um texto amplo do qual passa pelo processo de limpeza removendo caracteres especiais.
3. O texto agora limpo (sem caracteres especiais) passa por uma etapa de estruturação, particionando o mesmo em pequenos textos de no máximo 3500 caracteres, assim evitando que o sumarizador extrapole sua capacidade de geração de modelo.

4. Cada partição passa pelo processo de sumarização através do modelo *transformers*: *facebook/bart-large-cnn* que obteve bom desempenho na sumarização de textos com conteúdo educacional (OLIVEIRA et al., 2021).

5. Cada legenda é correlacionada a sua respectiva partição resumida, e através de técnicas de similaridade de texto é obtida uma nota de similaridade com valores de 0-1.

6. Os segmentos sem legendas são anotados com valor mínimo (0), assim cada segmento com legenda ou sem legenda terá seu valor de importância associado.

As etapas 4 e 5 são o ponto chave para a anotação de importância dos segmentos, a ideia principal é que nos resumos obtidos na etapa 4 estejam os pontos mais importantes da fala transcrita, logo os pontos mais importantes daquela aula. Desta forma na etapa 5, através de um algoritmo de similaridade de textos, assinala-se uma nota de 0-1 para aquele segmento. A Figura 10 ilustra as etapas 4 e 5 descritas anteriormente.

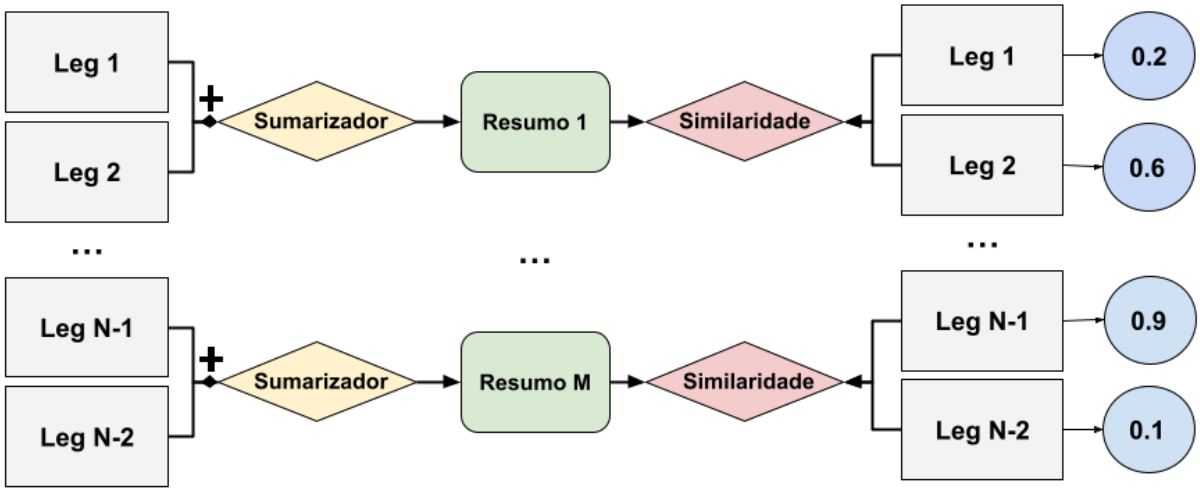


Figura 10 – Etapas da proposta de geração de resumos e obtenção de importância a partir da similaridade de textos.

EDUVSUM

Para a realização dos experimentos foi selecionado o *dataset* EDUVSUM (Educational Video Summarization) (GHAURI; HAKIMOV; EWERTH, 2020) que se trata de um conjunto de vídeos do domínio educacional para treinamento de sumarização. O mesmo contém vídeos de três plataformas populares de *e-learning*: Edx, YouTube e TIB AV-Portal cobrindo os mais variados temas de Ciência da Computação. O *dataset* contém 98 vídeos em inglês, e suas respectivas legendas, cada um com anotações em segmentos (de 5 segundos) de importância para o autor do artigo, que possui *background* acadêmico em

Ciência da Computação. As notas para cada segmento dos vídeos estão em uma escala de 1 a 10, valores maiores indicam maior importância daquele específico segmento em termos de informação para o tópico do vídeo (GHAURI; HAKIMOV; EWERTH, 2020).

O autor do *dataset* EDUVSUM dividiu o mesmo em um conjunto de treinamento (83 vídeos) e teste (15 vídeos) para treinamento e teste de técnicas de sumarização. A abordagem adotada para este trabalho utiliza uma rede pré-treinada para gerar o resumo dos textos das legendas, no entanto a nota é gerada para cada tempo de legenda a partir da similaridade de cada legenda com o resumo obtido, que variam de 0 a 1. Para comparação com os valores reais anotados pelo especialista é necessário adequar a saída do valor de similaridade às notas do autor.

Para isso, o método escolhido neste trabalho para calcular a similaridade de textos foi o SentenceTransformers[1], um *framework* que provê acesso a redes pré-treinadas derivadas do BERT capazes de calcular a similaridade de sentenças através da similaridade do cosseno (REIMERS; GUREVYCH, 2019). A vantagem deste algoritmo de similaridade de textos é a capacidade que o *framework* provê de treinar novamente as redes para o contexto da tarefa. Desta forma, é possível ajustar a saída da rede para as anotações de importância que o autor assinalou no conjunto de treino.

Com isso, foram selecionadas quatro redes pré-treinadas para a tarefa de treinar a similaridade com a anotação da importância do segmento de acordo com a nota do especialista. Os modelos selecionados para isso são:

- **Modelo 1 - *cross-encoder/stsb-roberta-base***[2]: Modelo treinado no conjunto *STS benchmark*[3] para predizer a similaridade de dois textos em valores de 0 a 1 utilizando o modelo RoBERTa, derivado do BERT.
- **Modelo 2 - *sentence-transformers/msmarco-roberta-base-v3***[4]: Semelhante ao Modelo 1 porém treinado no dataset MS MARCO, um *dataset* com documentos de pesquisa e perguntas reais do buscador Bing[5] (NGUYEN et al., 2016).
- **Modelo 3 - *sentence-transformers/msmarco-bert-base-dot-v5***[6]: Modelo BERT treinado também com o *dataset* MS MARCO.
- **Modelo 4 - *sentence-transformers/msmarco-distilbert-base-v4***[7]: Modelo DistilBERT, derivado do BERT, treinado também com o *dataset* MS MARCO.

[1] https://www.sbert.net/
[2] https://huggingface.co/cross-encoder/stsb-roberta-base
[3] http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark
[4] https://huggingface.co/sentence-transformers/msmarco-roberta-base-v3
[5] https://www.bing.com/
[6] https://huggingface.co/sentence-transformers/msmarco-bert-base-dot-v5
[7] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4

Desta forma, foram extraídos do conjunto de treinamento do EDUVSUM os 7591 pares de textos de similaridade com suas notas variando de 1 a 10. As notas foram normalizadas para valores de 0-1, para se ajustar as saídas das redes, e o conjunto foi dividido em 6591 pares de texto para treinamento e 1000 pares para validação. As redes então foram treinadas por 64 épocas (devido ser a quantidade mais que necessária para não haver mais melhoras no desempenho) com as notas do Autor para cada par de texto, em seguida foram escolhidas duas épocas para cada modelo considerando as que tiveram a maior acurácia no conjunto de validação e as que tiveram a melhor acurácia balanceada, que evita performances infladas de acurácia em conjunto de dados desbalanceados através de um ajuste da precisão e da sensibilidade (MOSLEY, 2013), a Equação 5.1 apresenta a forma que a acurácia balanceada é calculada.

$$\texttt{balanced-accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{5.1}$$

Depois de obtidas as melhores redes de cada arquitetura, as mesmas foram utilizadas no conjunto de teste para predizer a similaridade de cada legenda com o texto resumido, em seguida os valores são discretizados de 1-10 para cada *frame* do vídeo e a nota por segmento é pela média das notas dos *frames*.

5.3 Resultados

Os primeiros resultados são referentes a predição da similaridade dos modelos para os textos resumidos e as legendas no conjunto de teste, os mesmos podem ser vistos na Tabela 6, na qual tempo o tempo de treinamento em minutos, a acurácia (Acc), a acurácia balanceada (Acc balan.) e o erro absoluto médio (EAM) das notas preditas com as notas originais. Para cada modelo foram escolhidas duas redes com maior desempenho em acurácia (acc) e acurácia balanceada (b_acc), discriminados pelo tipo.

Tabela 6 – Desempenho das redes de similaridade para as notas do EDUVSUM

Modelo	Nome	Tempo treino	Tipo	Acc %	Acc balan.%	EAM
1	**stsb-roberta-base**	624 min	**acc**	26,49	**12,81**	**1,38**
			b_acc	24,74	11,42	1,41
2	**msmarco-roberta-base-v3**	354 min	**acc**	**26,54**	12,12	1,39
			b_acc	25,46	11,18	1,39
3	**msmarco-bert-base-dot-v5**	615 min	**acc**	25,00	11,50	1,44
			b_acc	24,74	12,42	1,48
4	**msmarco-distilbert-base-v4**	**178 min**	**acc**	25,98	11,45	1,42
			b_acc	23,87	11,62	1,51

É possível notar na Tabela que o modelo que conseguiu melhor adaptar para as notas dadas pelo Autor do *dataset* foi o Modelo 2 (*msmarco-roberta-base-v3*) no qual

conseguiu atingir a acurácia de 26,54%, seguido pelo Modelo 1 com acurácia de 24,49%, os desempenhos são semelhantes também quando comparado o EAM. O Modelo 3 teve desempenho levemente inferior quanto a acurácia e o EAM em comparação com os modelos roBERTa, isso era esperado devido ao modelo BERT ser o estado da arte e o modelo roBERTa derivado do BERT com performance cerca de 2-20% maior que o BERT[8]. Existe também uma leve diferença no modelo 1 e 2 quanto ao dataset do quais foram treinados, o *dataset* **msmarco** é assimétrico, ou seja, foi treinado para encontrar similaridade em textos de tamanhos diferentes, que se assemelha com os casos treinados nesse experimento e apresentando uma leve vantagem em acurácia quanto ao *stsb*.

A seguir, na Tabela 7, temos o resultado do processo de sumarização do conjunto de teste do EDUVSUM. Seguindo a mesma metodologia de resultados, temos a acurácia, erro médio por *frame* e erro médio por segmento de 5 segundos. Como as redes tinham saídas no formato sigmóide, não foi possível obter os resultados Top-2 e Top-3 de acurácia. O Modelo 5 são os dois melhores resultados no experimento original do Autor no conjunto EDUVSUM.

Tabela 7 – Desempenho dos Modelos na anotação de importância dos segmentos em vídeos do conjunto EDUVSUM.

Modelo	**Nome**	**Tipo**	**Acc %**	**EAM**	
				med_{fra}	med_{seg}
1	**stsb-roberta-base**	**acc**	24,83	1,52	1,47
		b_acc	22,54	1,55	1,51
2	**msmarco-roberta-base-v3**	**acc**	25,08	1,52	1,47
		b_acc	25,02	**1,49**	**1,45**
3	**msmarco-bert-base-dot-v5**	**acc**	**26,53**	1,52	1,48
		b_acc	25,25	1,59	1,54
4	**msmarco-distilbert-base-v4**	**acc**	25,20	1,55	1,49
		b_acc	23,85	1,60	1,52
5	**VGG-16**	**h2**	26,26	1,60	1,57
		h3	25,55	1,51	1,49

Nota-se inicialmente que todos modelos treinados conseguiram alcançar um desempenho semelhante ao da rede treinada pelo autor (VGG-16), tendo sido até superado pelo Modelo 3 com 26.53% de acurácia no conjunto de teste e pelo Modelo 2 quanto ao EAM por *frame* e por segmento. Vale ressaltar que o método do artigo original utiliza de características visuais (imagem do vídeo), textuais (legendas) e de áudio (som) (GHAURI; HAKIMOV; EWERTH, 2020). Considerando que a abordagem descrita neste trabalho utiliza apenas a sumarização da transcrição da fala dos vídeos educacionais, os resultados foram satisfatórios.

Para uma melhor visualização dos resultados, a Figura 11 a apresenta a predição do modelo 3 (que apresentou o melhor resultado na Tabela 7), e o valor original anotado pelo

[8] https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8

especialista para os segmentos de 5 segundos em dois vídeos, no primeiro gráfico temos um caso em que a predição obteve uma alta acurácia balanceada (35,2%) e no segundo caso temos uma predição com baixa acurácia balanceada (16,6%).

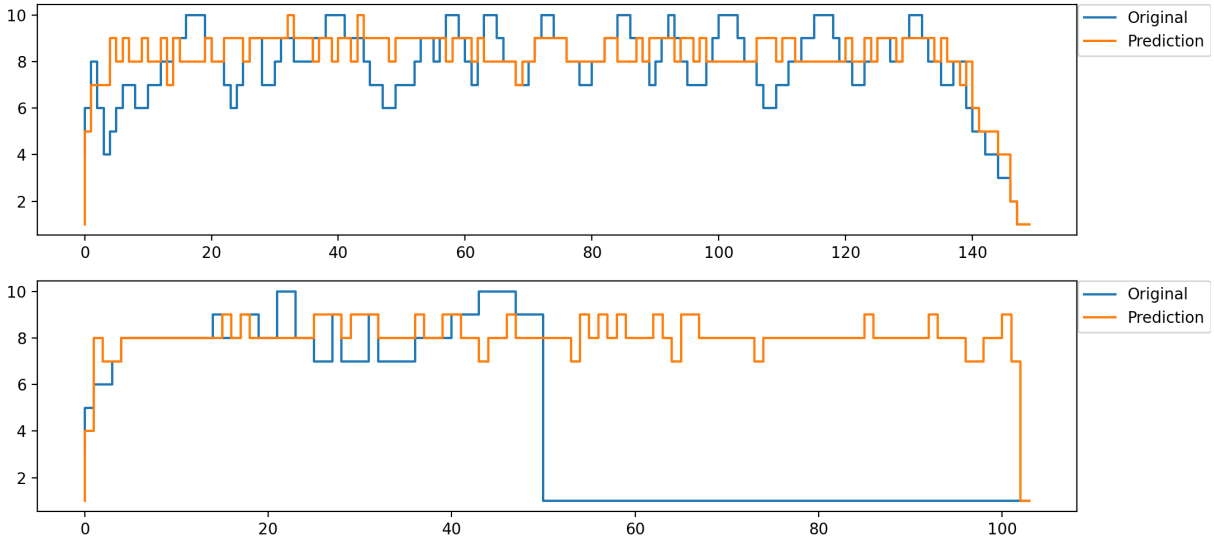


Figura 11 – Predições do modelo 3 para dois vídeos. Superior com alta acurácia balanceada (35,2%), Inferior com baixa acurácia balanceada (16,6%)

Nota-se que, no caso onde a onde a acurácia foi relativamente alta, o padrão da predição se comportou de maneira semelhante ao da anotação original, e mesmo não acertando em boa parte do tempo do vídeo, os resultados se mantiveram próximos, inclusive no início e final onde a predição acompanhou o valor original. Considerando que as redes utilizam apenas a similaridade da legenda com o resumo da transcrição do vídeo e nenhum caráter temporal é adicionado, o desempenho na predição desses pontos é satisfatório. Por outro lado, na predição que teve baixa acurácia podemos perceber uma discrepância na metade final do vídeo, onde o especialista avaliou que o restante dos segmentos possuem valor 1, enquanto que a predição se manteve na média entre 7-9, sendo essa a principal causa para a baixa acurácia nesse vídeo. Isso denota a necessidade do conjunto EDUVSUM precisar de mais avaliadores, para aumentar o grau de confiança na avaliação dos segmentos.

Devido a necessidade de ter a transcrição da fala, essa abordagem tem desempenho melhor em vídeos que possuem uma grande quantidade de diálogos, como em vídeos de aulas e apresentações. Tendo em vista isso, uma alternativa para melhores resultados seria a inclusão de outras características dos vídeos (como imagem e som) na metodologia apresentada de sumarização da transcrição da fala do interlocutor do vídeo.

Para uma análise mais profunda dos resultados é necessário enriquecer o *dataset* EDUVSUM com a avaliação de mais especialistas do domínio aumentando a confiabilidade das anotações e também diminuir a escala de notas de avaliação para valores de 1-5, desta

forma diminuindo a variedade de opções para o anotador e inserindo a opção de um valor médio (3).

6 Conclusão

Este trabalho tem como objetivo verificar a viabilidade do uso de redes baseadas em modelos *Deep Learning Transformers* para a geração de objetos textuais e de vídeo educacionais utilizando a sumarização automática de texto como cerne principal. Para isso, o trabalho foi dividido em dois experimentos onde cada um verificou um tipo de OA específico.

6.1 Geração Automática de Objetos de Aprendizagem

No primeiro experimento realizou-se o uso de sumarizadores automático de textos baseados em modelos *Transformers* para a geração de Objetos de Aprendizagem textuais. Nele foi feito o resumo de verbetes de textos em português provenientes da Wikipédia através da estruturação do texto, tradução para a língua inglesa e sumarização utilizando os principais modelos da plataforma *Hugging Face* na época da realização do experimento. Para avaliar o resultado, foi necessário realizar um experimento com especialistas através de um formulário onde os mesmos tinham que avaliar a qualidade de conteúdo e de gramática do objeto de aprendizagem gerado. Foram selecionados estudantes e graduados da computação para avaliar e comparar a qualidade dos textos gerados por três modelos de sumarizadores.

Dois modelos se sobressaíram quanto a avaliação de conteúdo e gramática, obtendo bons resultados com média em torno de 4 em uma escala de 1 a 5. Muito se deve ao fato de ambos utilizarem um dos modelos de sumarização mais robustos, principalmente o Modelo 2, do qual utiliza o modelo BART em que já se esperava desempenho melhor que o Modelo 1 com sua arquitetura DistilBART. Interessante notar que mesmo com o processo de tradução para outra língua, e no domínio da educação, os modelos ainda assim se destacaram. O Modelo 3 foi o que obteve menor avaliação e evidenciou uma necessidade de pós-processamento para a remoção de possíveis erros que atrapalhem a legibilidade do conteúdo.

Dessa forma é possível notar uma viabilidade do uso de sumarizadores para a geração de objetos de aprendizagem baseados em texto, contribuindo também na demonstração da viabilidade de utilizar sumarizadores da língua inglesa para a língua portuguesa através de tradutores automáticos.

Este experimento de geração automática de OA resultou em um *paper* intitulado de *Automatic Generation of Learning Objects Using Text Summarizer Based on Deep Learning Models* publicado no XXXII Simpósio Brasileiro de Informática na Educação

(SBIE 2021) com Qualis B1 (OLIVEIRA et al., 2021).

6.2 Sumarização de Vídeos Educacionais

No segundo experimento realizou-se a sumarização de vídeos educacionais através da classificação de importância dos segmentos de um vídeo educacional. Nele fez-se o uso de um sumarizador de texto baseado em modelo *Deep Learning* para resumir a transcrição da fala de um vídeo, e, através deste resumo, correlacionar com o tempo proveniente de seu segmento e classificar uma importância do mesmo para o conteúdo geral do vídeo. Para isso fez-se o uso do *dataset* EDUVSUM com 98 vídeos educacionais com anotações de importância feito por um especialista da área para cada segmento de 5 segundos. Utilizando algoritmos de similaridade de texto baseado em modelos pré-treinados de *Deep Learning Transformers* foi possível correlacionar cada legenda com o resumo gerado através do treino com a classificação feita pelo especialista no conjunto de teste.

O resultado do experimento demonstrou a viabilidade do uso de sumarizadores de texto para a classificação de segmentos em vídeos devido ao resultado ter sido semelhante ao melhor modelo multimodal realizado no artigo Ghauri, Hakimov e Ewerth (2020), tendo sido superado pelo Modelo BERT pré-treinado no *dataset* MSMarco com 26,53% de acurácia e superado também pelo modelo RoBERTa pré-treinado no mesmo *dataset* na tarefa de regressão, onde o erro absoluto médio das notas ficaram 1,49 por *frame* e 1,45 por segmento de 5 segundos. Considerando que apenas o componente textual dos vídeos foram utilizados, sem informações de vídeo, áudio ou temporal, os resultados foram satisfatórios.

Vale ressaltar que esse tipo de sumarização é de nicho, específico para vídeos com grande quantidade de fala (como aulas, seminários e palestras) tendo em vista que os segmentos sem fala são dadas notas mínimas de importância. Outro ponto importante é a necessidade de aumentar a quantidade de avaliações de especialistas no *dataset* EDUVSUM para maior confiabilidade das notas e a diminuição da escala de notas para valores com menor granularidade, como de 1-5.

Os resultados obtidos com o segundo experimento deste trabalho foram satisfatórios por terem superados os modelos estado da arte desenvolvidos pelo autor em Ghauri, Hakimov e Ewerth (2020) para o conjunto EDUVSUM, melhorando tanto em valores de acurácia (26,53%) como também em EAM por *frame* e por segmento (1,49 e 1,45) utilizando apenas sumarizadores de texto para a tarefa, desta forma sendo uma contribuição a abordagem de sumarização de vídeos utilizando modelos Transformers.

6.3 Trabalhos Futuros

A metodologia proposta nos dois experimentos possuem bastante espaço para melhorias, desta forma, algumas sugestões de melhoria estão listadas a seguir:

Geração Automática de Objetos de Aprendizagem

- Comparação dos resumos gerados pelos modelos por resumos de especialistas da área para melhor validação dos resultados.
- Pós-processamento na saída dos modelos e geração de tópicos através dos elementos existentes nos verbetes da Wikipedia.
- Geração de Objetos de Aprendizagem do tipo perguntas e respostas através do conteúdo educacional extraído da Wikipedia utilizando modelos de *Deep Learning*.

Sumarização de Vídeos Educacionais

- Publicar o experimento em forma de artigo para uma conferência para avaliação por pares da metodologia e resultados.
- Avaliar a metodologia desse experimento em outros *datasets* ou utilizando uma pesquisa subjetiva com especialistas quanto a qualidade do vídeo resumido.
- Expandir o *dataset* EDUVSUM com a avaliação de mais especialistas.
- Unir à metodologia proposta outros componentes existentes nos vídeos educacionais (como o áudio e a imagem) com o intuito de melhorar a performance em vídeos com pouca fala.
- Verificar a viabilidade do uso de tradutores automáticos para a sumarização de vídeos educacionais em outras línguas diferentes do inglês.

Referências

ABHILASH, R. K.; ANURAG, C.; AVINASH, V.; UMA, D. Lecture video summarization using subtitles. In: SPRINGER. *2nd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*. [S.l.], 2021. p. 83–92. Citado 2 vezes nas páginas 17 e 36.

ABRAHAM, A. Artificial neural networks. *handbook of measuring system design*, Wiley Online Library, 2005. Citado 2 vezes nas páginas 19 e 20.

ALLAHYARI, M.; POURIYEH, S.; ASSEFI, M.; SAFAEI, S.; TRIPPE, E. D.; GUTIERREZ, J. B.; KOCHUT, K. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*, 2017. Citado na página 12.

ALRUMIAH, S. S.; AL-SHARGABI, A. A. Educational videos subtitles' summarization using latent dirichlet allocation and length enhancement. *cmc-computers materials & continua*, v. 70, n. 3, p. 6205–6221, 2022. Citado 2 vezes nas páginas 17 e 36.

APOSTOLIDIS, E.; BALAOURAS, G.; MEZARIS, V.; PATRAS, I. Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames. In: *Proceedings of the 2022 International Conference on Multimedia Retrieval*. [S.l.: s.n.], 2022. p. 407–415. Citado na página 18.

ARAÚJO, R. D.; BRANT-RIBEIRO, T.; FREITAS, R. S. de; DORÇA, F. A.; CATTELAN, R. G. Autoria automática de objetos de aprendizagem a partir de captura multimídia e associação a estilos de aprendizagem. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2014. v. 25, n. 1, p. 229. Citado na página 11.

Ben Lutkevich. *BERT language model*. 2020. Disponível em: <https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model>. Acesso em: 5 de agosto de 2022. Citado na página 25.

BRAGA, A. d. P. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: Livros Técnicos e Científicos, 2000. Citado na página 21.

BRAGA, J.; MENEZES, L. Introdução aos objetos de aprendizagem. *Objetos de Aprendizagem*, v. 1, p. 19–40, 2014. Citado na página 27.

CHOWDHARY, K. Natural language processing. *Fundamentals of artificial intelligence*, Springer, p. 603–649, 2020. Citado na página 11.

DENG, L.; YU, D. *Deep Learning: Methods and Applications*. [S.l.], 2014. Disponível em: <https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/>. Citado 2 vezes nas páginas 21 e 22.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Citado na página 24.

Eduardo Muñoz. *A Guide to the Encoder-Decoder Model and the Attention Mechanism*. 2020. Disponível em: <https://towardsdatascience.com/attention-is-all-you-need-discovering-the-transformer-paper-73e5ff5e0634>. Acesso em: 5 de agosto de 2022. Citado 2 vezes nas páginas 23 e 24.

Favio Vázquez. *Deep Learning made easy with Deep Cognition*. 2017. Disponível em: <https://becominghuman.ai/deep-learning-made-easy-with-deep-cognition-403fbe445351>. Acesso em: 3 de agosto de 2022. Citado na página 22.

GHAURI, J. A.; HAKIMOV, S.; EWERTH, R. Classification of important segments in educational videos using multimodal features. *arXiv preprint arXiv:2010.13626*, 2020. Citado 6 vezes nas páginas 17, 35, 37, 38, 40 e 44.

GHAURI, J. A.; HAKIMOV, S.; EWERTH, R. Supervised video summarization via multiple feature sets with parallel attention. In: IEEE. *2021 IEEE International Conference on Multimedia and Expo (ICME)*. [S.l.], 2021. p. 1–6s. Citado 2 vezes nas páginas 18 e 35.

GURNEY, K. An introduction to neural networks. 1997. Citado na página 19.

GYGLI, M.; GRABNER, H.; RIEMENSCHNEIDER, H.; GOOL, L. V. Creating summaries from user videos. In: SPRINGER. *European conference on computer vision*. [S.l.], 2014. p. 505–520. Citado na página 18.

HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman Editora, 2007. Citado na página 19.

HINTON, G.; VINYALS, O.; DEAN, J. Distilling the knowledge in a neural network (2015). *arXiv preprint arXiv:1503.02531*, v. 2, 2015. Citado na página 25.

HOOSHYAR, D.; YOUSEFI, M.; WANG, M.; LIM, H. A data-driven procedural-content-generation approach for educational games. *Journal of Computer Assisted Learning*, Wiley Online Library, v. 34, n. 6, p. 731–739, 2018. Citado na página 16.

KOVALESKI, P. de A. *Implementação de Redes Neurais Profundas para Reconhecimento de Ações em Vídeos*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2018. Citado na página 20.

KURDI, G.; LEO, J.; PARSIA, B.; SATTLER, U.; AL-EMARI, S. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, Springer, v. 30, n. 1, p. 121–204, 2020. Citado na página 15.

LEWIS, M.; LIU, Y.; GOYAL, N.; GHAZVININEJAD, M.; MOHAMED, A.; LEVY, O.; STOYANOV, V.; ZETTLEMOYER, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. Citado na página 29.

LI, C.; XING, W. Natural language generation using deep learning to support mooc learners. *International Journal of Artificial Intelligence in Education*, Springer, p. 1–29, 2021. Citado na página 15.

LIU, M.; CALVO, R. A.; ADITOMO, A.; PIZZATO, L. A. Using wikipedia and conceptual graph structures to generate questions for academic writing support. *IEEE Transactions on Learning Technologies*, IEEE, v. 5, n. 3, p. 251–263, 2012. Citado na página 27.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. Citado 2 vezes nas páginas 25 e 26.

MARUMO, F. S. Deep learning para classificação de fake news por sumarização de texto. 2018. Citado na página 21.

MENG, Q.; YANG, K.; GONG, Y. A lightweight video summarization method considering the subjective transition degree for online educational screen content videos. In: *Proceedings of the 4th International Symposium on Signal Processing Systems*. [S.l.: s.n.], 2022. p. 81–88. Citado na página 18.

MOSLEY, L. A balanced approach to the multi-class imbalance problem. *Doctor of Philosophy Thesis, Iowa State University of Science and Technology, USA*, 2013. Citado na página 39.

MUBARAK, A. A.; CAO, H.; AHMED, S. A. Predictive learning analytics using deep learning model in moocs' courses videos. *Education and Information Technologies*, Springer, v. 26, n. 1, p. 371–392, 2021. Citado na página 35.

NGUYEN, T.; ROSENBERG, M.; SONG, X.; GAO, J.; TIWARY, S.; MAJUMDER, R.; DENG, L. Ms marco: A human generated machine reading comprehension dataset. In: *CoCo@ NIPs*. [S.l.: s.n.], 2016. Citado na página 38.

OLIVEIRA, L. M. R.; BUSSON, A. J. G.; SALLES, S. N. Carlos de; SANTOS, G. N. dos; COLCHER, S. Automatic generation of learning objects using text summarizer based on deep learning models. In: SBC. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2021. p. 728–736. Citado 3 vezes nas páginas 36, 37 e 44.

OTT, M.; EDUNOV, S.; GRANGIER, D.; AULI, M. Scaling neural machine translation. *CoRR*, abs/1806.00187, 2018. Disponível em: <http://arxiv.org/abs/1806.00187>. Citado na página 26.

POTAPOV, D.; DOUZE, M.; HARCHAOUI, Z.; SCHMID, C. Category-specific video summarization. In: SPRINGER. *European conference on computer vision*. [S.l.], 2014. p. 540–555. Citado na página 35.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. et al. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 9, 2019. Citado 2 vezes nas páginas 16 e 26.

Rani Horev. *BERT Explained: State of the art language model for NLP*. 2018. Disponível em: <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>. Acesso em: 5 de agosto de 2022. Citado na página 24.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. Citado na página 38.

ROCHA, O. R.; ZUCKER, C. F.; GIBOIN, A.; LAGARRIGUE, A. Automatic generation of questions from dbpedia. *International Journal of Continuing Engineering Education and Life Long Learning*, Inderscience Publishers (IEL), v. 30, n. 3, p. 276–294, 2020. Citado na página 15.

Rohan Jagtap. *RoBERTa: Robustly Optimized BERT-Pretraining Approach*. 2020. Disponível em: <https://medium.com/dataseries/roberta-robustly-optimized-bert-pretraining-approach-d033464bd946>. Acesso em: 6 de agosto de 2022. Citado na página 26.

RÜDIAN, S.; HEUTS, A.; PINKWART, N. Educational text summarizer: Which sentences are worth asking for? *DELFI 2020–Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik eV*, Gesellschaft für Informatik eV, 2020. Citado na página 15.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, Nature Publishing Group, v. 323, n. 6088, p. 533–536, 1986. Citado na página 21.

SANH, V.; DEBUT, L.; CHAUMOND, J.; WOLF, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. Citado na página 25.

SANTOS, R. E.; SOUZA, E. P.; CORREIA-NETO, J. S.; MAGALHÃES, C. V.; VILAR, G. Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático. *Revista de Sistemas e Computação-RSC*, v. 4, n. 2, 2015. Citado na página 11.

SONG, Y.; VALLMITJANA, J.; STENT, A.; JAIMES, A. Tvsum: Summarizing web videos using titles. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 5179–5187. Citado 2 vezes nas páginas 18 e 35.

SOUSA, G. G. B. *Deep Learning para a Detecção e Classificação de Pneumonia por Radiografias do Tórax*. 44 p. Monografia (Graduação) — Ciências da Computação, Universidade Federal do Maranhão, São Luís, 2018. Citado 2 vezes nas páginas 19 e 20.

TAROUCO, L. M. R.; COSTA, V. M. d.; AVILA, B. G.; BEZ, M. R.; SANTOS, E. F. d. Objetos de aprendizagem: teoria e prática. Evangraf, 2014. Citado na página 11.

TRINH, T. H.; LE, Q. V. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018. Citado na página 26.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado 3 vezes nas páginas 12, 23 e 24.

XU, Y.; SMEETS, R.; BIDARRA, R. Procedural generation of problems for elementary math education. *International Journal of Serious Games*, v. 8, n. 2, p. 49–66, 2021. Citado na página 16.

YANG, G.; CHEN, N.-S.; SUTINEN, E.; ANDERSON, T.; WEN, D. et al. The effectiveness of automatic text summarization in mobile learning contexts. *Computers & Education*, Elsevier, v. 68, p. 233–243, 2013. Citado 2 vezes nas páginas 15 e 16.

ZHANG, J.; ZHAO, Y.; SALEH, M.; LIU, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2020. p. 11328–11339. Citado na página 29.

ZHU, W.; HAN, Y.; LU, J.; ZHOU, J. Relational reasoning over spatial-temporal graphs for video summarization. *IEEE Transactions on Image Processing*, IEEE, v. 31, p. 3017–3031, 2022. Citado na página 18.

ZHU, Y.; KIROS, R.; ZEMEL, R.; SALAKHUTDINOV, R.; URTASUN, R.; TORRALBA, A.; FIDLER, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 19–27. Citado na página 25.