Otimização da Mineração de Dados com GPT da OpenAI: Classificação Nutricional de Rações

Benildes Fernandes de Menezes
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Cubatão, SP, Brasil.

Bruna Helena Silva Santos
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Cubatão, SP, Brasil.

Douglas Reis Rodrigues dos Santos
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Cubatão, SP, Brasil.

Eduardo Henrique Gomes
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Cubatão, SP, Brasil.

Marcelo Modesto de Lima Junior
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Cubatão, SP, Brasil.

RESUMO: Este artigo explora a integração da Inteligência Artificial Generativa (IAG), com ênfase no modelo GPT da OpenAI, no campo da mineração de dados. A pesquisa destaca como a IAG, ao avançar na análise de dados não estruturados, supera as limitações das técnicas convencionais de mineração de dados, facilitando a geração de insights inovadores e aprimorados. Focando em um estudo de caso específico - a classificação nutricional de rações para animais de estimação - o artigo demonstra a aplicação prática do GPT no processo de Mineração de Dados dentro do método KDD (Knowledge Discovery in Databases). O uso do GPT não apenas automatiza a categorização e o resumo de textos, mas também abre novas perspectivas para a tomada de decisões e inovação, potencializando a eficiência analítica e a personalização de serviços. Este estudo ilustra como a IAG pode transformar significativamente a maneira como os dados são analisados e interpretados, impulsionando a inovação e a tomada de decisões informadas em diversos campos, redefinindo assim o potencial do big data em múltiplas indústrias.

PALAVRAS-CHAVES: GPT, inteligência artificial generativa, mineração de dados, open AI, k-means.

ABSTRACT: This paper investigates the integration of Generative Artificial Intelligence (GAI), focusing on OpenAI's GPT model, in the field of data mining. The

research emphasizes how GAI, advancing in the analysis of unstructured data, overcomes the limitations of conventional data mining techniques, thereby facilitating the generation of innovative and enhanced insights. Centering on a specific case study - the nutritional classification of pet food - the article demonstrates the practical application of GPT in the Data Mining process within the KDD (Knowledge Discovery in Databases) method. The use of GPT not only automates the categorization and summarization of texts but also opens new perspectives for decision-making and innovation, enhancing analytical efficiency and service personalization. This study illustrates how GAI can significantly transform the way data is analyzed and interpreted, driving innovation and informed decision-making across various fields, thus redefining the potential of big data in multiple industries.

KEYWORDS: GPT, Generative Artificial Intelligence, Data Mining, OpenAI, K-Means.

INTRODUÇÃO

A evolução tecnológica e a proliferação de dados na era digital têm impulsionado a necessidade de abordagens inovadoras na análise de informações. Dentro desse cenário, a integração da Inteligência Artificial Generativa (IAG) no contexto da Mineração de Dados emerge como um campo promissor e transformador. A capacidade da IAG de não apenas identificar padrões, mas também gerar dados sintéticos e revelar insights ocultos apresenta uma abordagem única para extrair valor de conjuntos de dados cada vez mais complexos.

A era digital conferiu à mineração de dados um papel crucial na extração de informações valiosas e padrões significativos de informações. A mineração de dados é um campo interdisciplinar que integra diversas técnicas e desempenha um papel vital na identificação de padrões. (GOODFELLOW, 2014).

Esta abordagem vai além da simples identificação de padrões, permitindo a criação de modelos capazes de gerar novos dados, semelhantes aos existentes. A interseção entre a Inteligência Artificial Generativa e a Clusterização, especialmente por meio do algoritmo K-means, revela um potencial promissor para impulsionar a análise e a interpretação de dados, abrindo portas para descobertas inovadoras e aplicações revolucionárias. (HAN et al., 2011).

Nesta pesquisa, investigamos a fusão entre a Inteligência Artificial Generativa (IAG) e a Mineração de Dados, com ênfase especial na aplicação do algoritmo K-means. O objetivo é entender como essa combinação pode aprimorar a análise e o aproveitamento de grandes volumes de dados, impulsionando a inovação e o

progresso em diversos campos. Ao explorar as aplicações práticas da IAG, incluindo a geração de dados sintéticos, a análise de padrões ocultos e a melhoria da qualidade da clusterização, buscamos oferecer uma visão abrangente do impacto dessa convergência no processo de descoberta de conhecimento.

MINERAÇÃO DE DADOS

Atualmente, com a crescente digitalização e armazenamento de dados em diversas áreas, a necessidade de extrair informações valiosas e exploradas a partir desses conjuntos de dados tornou-se essencial. É nesse contexto que a Mineração de Dados desempenha um papel crucial. A Mineração de Dados é um campo interdisciplinar que combina técnicas da estatística, inteligência artificial e banco de dados para descobrir padrões, tendências e conhecimentos ocultos em grandes volumes de dados. O processo de Mineração de Dados envolve várias etapas, sendo o KDD (Knowledge Discovery in Databases) uma metodologia abrangente para esse processo de maneira organizada e eficaz formalizado por Fayyad (1996). O KDD é composto por diversas fases interligadas, cada uma com um propósito específico, visando transformar dados brutos em conhecimento útil.

A sigla KDD engloba as seguintes etapas:

Seleção: Nesta fase, os dados relevantes são selecionados a partir de fontes diversas, como bancos de dados, arquivos e sistemas de informações.

Pré-processamento: Os dados brutos são limpos e pré-processados para tratar ruídos, valores ausentes e inconsistências, garantindo que estejam prontos para a análise.

Transformação: Nesta etapa, os dados são transformados em um formato adequado para análise, muitas vezes envolvendo a redução da dimensionalidade ou a criação de novas variáveis.

Mineração de Dados: Esta é a fase central, onde algoritmos de mineração são aplicados aos dados transformados para identificar padrões, relações e informações relevantes.

Avaliação: Os padrões e conhecimentos descobertos são avaliados quanto à sua fidelidade e utilidade, bem como em relação aos objetivos da análise.

Interpretação: Nesta fase, os resultados são interpretados e convertidos em conhecimento acionável. Os padrões descobertos estão relacionados ao contexto do problema em questão.

Consolidação: Finalmente, o conhecimento obtido é consolidado, documentado e apresentado de forma compreensível aos tomadores de decisão.

> O algoritmo KDD não se refere a um algoritmo específico, mas sim a uma metodologia que abrange uma variedade de técnicas e algoritmos de Mineração de Dados. O sucesso do processo KDD depende da escolha adequada dessas técnicas de acordo com os objetivos e características dos dados analisados. Em resumo, a Mineração de Dados através da metodologia KDD é uma abordagem sistemática para explorar grandes conjuntos de dados e extrair conhecimentos valiosos. Ao seguir as etapas do processo KDD, é possível transformar dados em *insights* acionáveis, o que é fundamental para apoiar a tomada de decisões controladas em diversos campos, desde negócios a pesquisa científica. (GOMES, 2019).

INTELIGÊNCIA ARTIFICIAL GENERATIVA

A Inteligência Artificial (IA) tem evoluído constantemente desde sua origem, com a Inteligência Artificial Generativa (IAG) emergindo como uma subárea dinâmica. A IAG foca em criar modelos e algoritmos capazes de gerar dados similares aos reais, diferenciando-se dos sistemas discriminativos de IA que classificam ou diferenciam dados existentes. O objetivo da IAG é produzir resultados indistinguíveis dos dados reais para observadores humanos ou outros modelos de IA, baseando-se em técnicas avançadas de aprendizado de máquina, como redes neurais e deep learning, e apoiada por redes neurais generativas.

Uma abordagem notável na IAG é o uso de Redes Adversariais Generativas (GANs), introduzidas por Goodfellow (2014). Estes sistemas são capazes de gerar hipóteses a partir de dados padronizados e aprender autonomamente, reconhecendo padrões em múltiplas camadas de processamento (MOURA, 2023).

Os modelos de linguagem, especialmente os de grande escala como o GPT (Generative Pre-trained Transformer) da OpenAI (2021), têm se destacado como "few-shot learners", conforme observado por Brown (2020). No âmbito da IAG, essa capacidade se traduz na habilidade do modelo de gerar conteúdo linguístico coerente

e contextualmente relevante a partir de poucos exemplos. O GPT adapta-se rapidamente a novas tarefas, evidenciando seu papel significativo na IA generativa e abrindo caminhos para aplicações em processamento de linguagem natural e áreas relacionadas (BROWN, 2020).

As aplicações da IA generativa são diversas e incluem a criação de arte gerada por computador, a geração de texto, a criação de músicas e até mesmo a melhoria da realidade virtual. No entanto, uma aplicação menos óbvia, mas igualmente impactante, é a utilização da IA generativa na Mineração de Dados. (VASWANI et al., 2017).

IA GENERATIVA NA MINERAÇÃO DE DADOS

A aplicação do GPT na fase de Mineração de Dados dentro do método KDD representa um avanço significativo, considerando a habilidade do modelo em analisar e interpretar grandes volumes de dados linguísticos. Esta capacidade é crucial, pois permite uma análise mais aprofundada e contextualizada dos dados, essencial para extrair padrões significativos e insights valiosos.

A eficácia do GPT na Mineração de Dados é particularmente notável, dada a sua habilidade em processar e analisar informações complexas. Nesta fase do KDD, a precisão na análise de dados é fundamental para garantir a qualidade dos insights gerados. A flexibilidade do GPT em se adaptar a novas tarefas com ajustes mínimos destaca seu potencial como uma ferramenta poderosa para automatizar e otimizar a Mineração de Dados, contribuindo significativamente para a descoberta de conhecimento em conjuntos de dados extensos e complexos.

A Inteligência Artificial Generativa oferece oportunidades empolgantes no contexto da Mineração de Dados e do método KDD. Ao aplicar técnicas generativas, podemos melhorar a qualidade dos dados, descobrir padrões ocultos e melhorar a interpretação dos resultados. A IA generativa está se tornando uma ferramenta valiosa para analistas de dados e cientistas de dados em várias indústrias, e seu potencial continua a se expandir à medida que novas técnicas e abordagens são desenvolvidas.

À medida que a IA generativa continua a evoluir, é fundamental considerar questões éticas e de privacidade, especialmente quando se trata de geração de dados sintéticos. No entanto, quando usada com responsabilidade, a IA generativa

tem o potencial de revolucionar a forma como exploramos e compreendemos grandes volumes de dados, impulsionando a inovação e a tomada de decisões informadas em diversas áreas.

ESTUDO DE CASO: CLASSIFICAÇÃO NUTRICIONAL DE RAÇÕES

O presente estudo de caso foi conduzido com um grupo de alunos do curso superior de Análise e Desenvolvimento de Sistemas. O objetivo central do projeto era desenvolver uma ferramenta digital que auxiliasse proprietários de animais domésticos na seleção otimizada de rações, ponderando aspectos relativos à saúde e economia. Nesse contexto, a elaboração de um modelo analítico baseado nos níveis de garantia presentes em rações para pets foi identificada como primordial.

A equipe desejava classificar as rações em três categorias: A (alta qualidade), B (qualidade média) e C (baixa qualidade). Para aprimorar a análise e a tomada de decisão, o grupo formulou perguntas específicas para serem respondidas pelo modelo GPT (Generative Pre-trained Transformer):

1. *Como determinar a qualidade de uma ração seca para cães com base em seus níveis de garantia?*
2. *Como realizar a mineração de dados em Python para este propósito?*
3. *Como classificar rações para pets usando o algoritmo K-Means?*
4. *Como utilizar um arquivo CSV de rações para pets para criar uma classificação com o K-Means em Python?*
5. *Como desenvolver uma IA usando TensorFlow para classificar rações de animais domésticos em níveis A, B ou C, baseando-se na saúde e qualidade nutricional?*

ANÁLISE DOS RESULTADOS

Em resposta aos questionamentos formulados, o GPT ofereceu orientações abrangentes de todas as etapas do processo KDD, desde a coleta de dados por meio de Web Scraping referentes aos níveis de garantia até a aplicação do algoritmo K-means na fase de mineração de dados do Processo KDD. Com base nessas orientações, foram identificados grupos de rações classificados como tipos A, B e C. No âmbito deste trabalho, focaremos nossa análise unicamente na fase mineração

de dados dentro do processo de KDD. Após a definição dos atributos-chave, buscou-se o auxílio do modelo GPT para orientar o processo de mineração de dados utilizando a linguagem de programação python, como mostrado na figura 1.

De acordo com as diretrizes fornecidas pelo modelo GPT em resposta à segunda pergunta, O GPT apresentou os seguintes detalhes:

TABELA 1. Resposta do GPT

Resposta
A mineração de dados em Python pode ser realizada usando várias bibliotecas e técnicas disponíveis. Aqui está um exemplo básico de como você pode realizar a mineração de dados em Python.

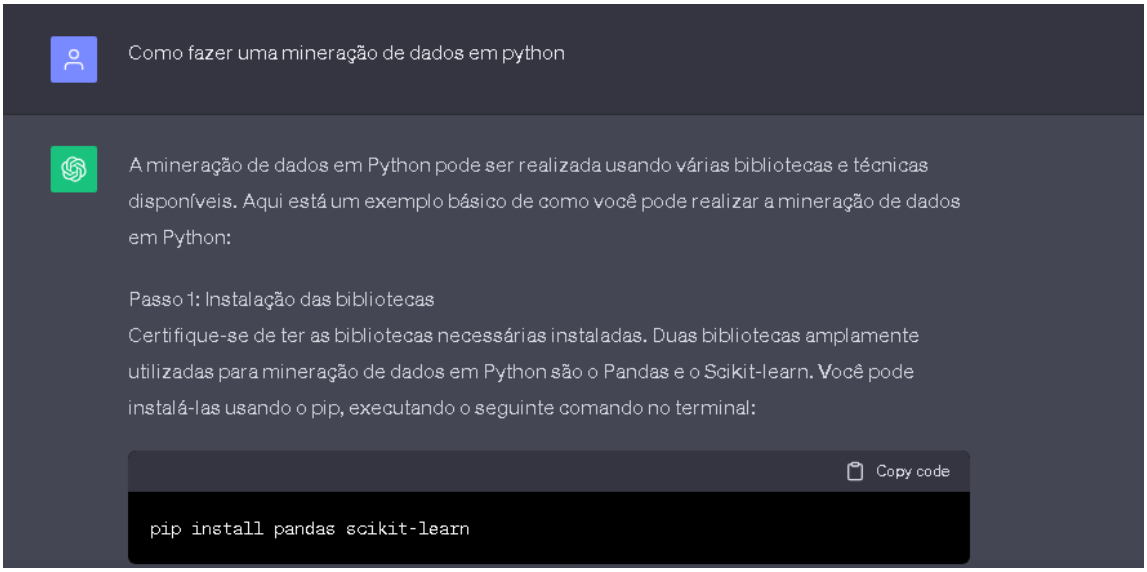

FIGURA 1. Trecho de código para instalação da biblioteca necessária para realizar a mineração de dados fornecida pelo GPT.

Após a instalação das bibliotecas necessárias, o próximo passo envolveu a utilização de uma base de dados já submetida ao pré-processamento. O objetivo era importar esses dados para realizar a mineração, visando gerar insights relevantes, conforme ilustrado na Figura 2.

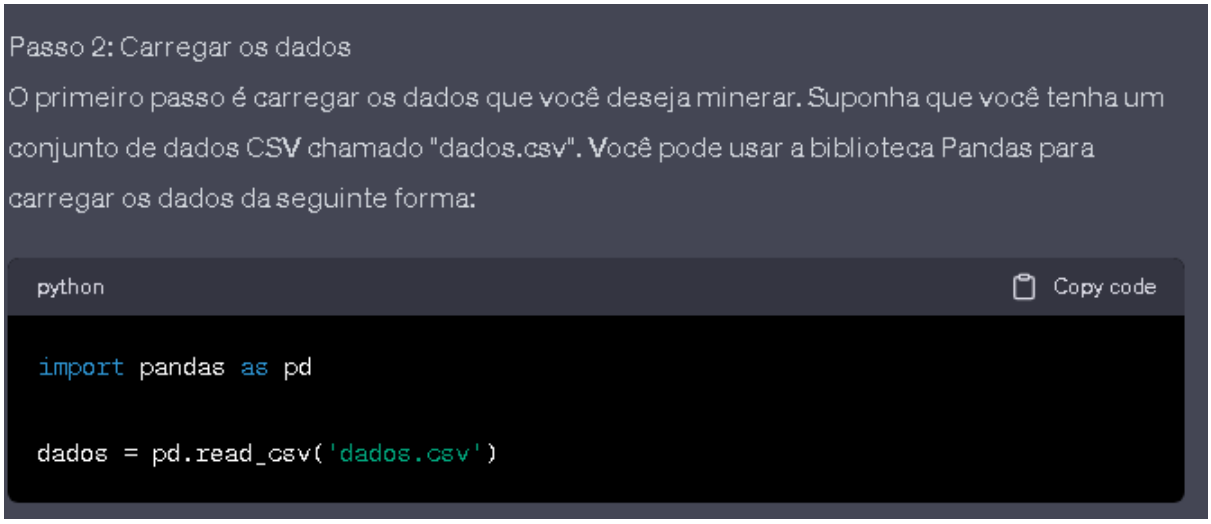

FIGURA 2. Trecho de código para a importação da biblioteca pandas, na qual será utilizada para realizar o carregamento dos dados da base de dados que está em um arquivo CSV.

Em seguida, o GPT apresentou duas abordagens para a categorização de dados. A primeira sugestão envolve a aplicação de técnicas de aprendizado de máquina supervisionado, utilizando a classificação por meio de uma árvore de decisão. A segunda sugestão propõe o uso do aprendizado de máquina não supervisionado para agrupar os dados, empregando o algoritmo K-means, conforme ilustrado nas figuras 3 e 4.

Aprendizado de máquina supervisionado - Classificação usando árvore de decisão:

python Copy code

```python
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Separar os dados em recursos (X) e rótulos (y)
X = dados_codificados.drop('target', axis=1)
y = dados_codificados['target']

# Dividir os dados em conjunto de treinamento e conjunto de teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_

# Criar e treinar o modelo de classificação
modelo = DecisionTreeClassifier()
modelo.fit(X_train, y_train)

# Fazer previsões com o modelo treinado
previsoes = modelo.predict(X_test)

# Calcular a precisão do modelo
precisao = accuracy_score(y_test, previsoes)
print("Precisão: ", precisao)
```

FIGURA 3. Trecho de código para realizar a classificação de dados a partir de aprendizado de máquina supervisionado.

Aprendizado de máquina não supervisionado - Agrupamento usando o algoritmo K-means:

python Copy code

```python
from sklearn.cluster import KMeans

# Criar o modelo de agrupamento
modelo = KMeans(n_clusters=3)

# Aplicar o modelo aos dados
modelo.fit(X)

# Obter os rótulos dos grupos para cada exemplo
rotulos = modelo.labels_
```

FIGURA 4. Trecho de código para realizar agrupamento de dados utilizando o algoritmo K-Means.

A seção do código destinada ao agrupamento de dados foi utilizada para categorizar as rações de animais de estimação com base em um arquivo CSV que já passou pela fase inicial do método KDD, ou seja, o pré-processamento, no qual foram removidos dados que poderiam interferir na mineração de dados. Durante esse processo, o cluster zero (0), representa rações de alta qualidade, o cluster um (1), se posiciona em um ponto intermediário e o cluster dois (2), engloba rações de baixa qualidade.

Esses agrupamentos podem ser observados claramente na figura 5, onde foi possível gerar um gráfico depois da clusterização a partir da plataforma Google Colab.

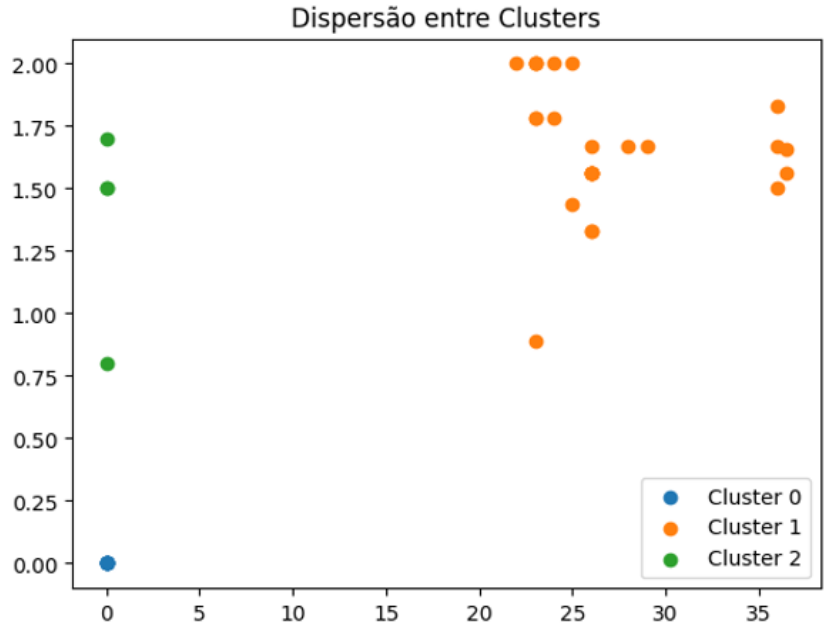

FIGURA 5. Gráfico de dispersão entre clusters a respeito das rações pets.

CONSIDERAÇÕES FINAIS

Este estudo de caso, focado na classificação nutricional de rações para animais de estimação, ilustra o potencial da Inteligência Artificial Generativa (IAG), particularmente o modelo GPT, na otimização do processo de Mineração de Dados dentro do método KDD. A aplicação prática do GPT demonstrou ser não apenas viável, mas também eficiente, fornecendo orientações detalhadas e práticas para a coleta, seleção, limpeza e análise de dados.

Os resultados obtidos neste projeto reforçam a capacidade da IAG de adaptar-se a diferentes desafios. A eficácia do modelo GPT na categorização e análise de

dados nutricionais de rações para pets, utilizando o algoritmo K-means, destaca a utilidade prática da IA generativa em contextos específicos de mineração de dados.

No entanto, este estudo também sublinha a importância de pesquisas adicionais na intersecção entre IAG e KDD. Aspectos como a ética da automação, a qualidade e a relevância dos insights gerados, bem como a eficácia a longo prazo do uso de modelos de IA generativa em processos de KDD, são áreas que merecem atenção contínua.

Além disso, este estudo evidencia a crescente relevância da IAG no campo da mineração de dados, particularmente na contribuição de modelos como o GPT e na interação com técnicas de clusterização, como o K-means. A IA generativa está redefinindo os paradigmas tradicionais de análise de dados, abrindo novas perspectivas e possibilidades para a descoberta de conhecimento.

Em resumo, os insights obtidos neste estudo de caso reforçam o valor da IAG como uma ferramenta poderosa na mineração de dados, capaz de transformar significativamente a maneira como os dados são analisados e interpretados, impulsionando a inovação e a tomada de decisões informadas em diversos campos.

REFERÊNCIAS

BROWN, TOM et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877-1901, 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. AI Magazine, [S.l.], v. 17, n. 3, p.37-54. 1996.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., & BENGIO, Y. Generative adversarial nets. **In Advances in neural information processing systems** (pp. 2672- 2680). 2014.

GOMES, JOSIR; MEDEIROS PIMENTA, RICARDO; SCHNEIDER, MARCO; Mineração De Dados Na Pesquisa Em Ciência Da Informação: Desafios E Oportunidades. **Encontro Nacional De Pesquisa Em Ciência Da Informação**. v. 1, n. 1, 2019.

HAN, J., PEI, J., & KAMBER, M. **Data mining: concepts and techniques**. Elsevier. 2011.

MOURA, M. V. A Inteligência Artificial Generativa como autora de invenções patenteáveis: um estudo analítico do" Caso DABUS". 2023. Disponível em: https://repositorio.animaeducacao.com.br/handle/ANIMA/36309. Acesso em 15 ago. 2021.

OPENAI. GPT-3.5 ChatGPT. Disponível em: https://www.openai.com/chatgpt. Acesso em: 22 de agosto de 2023.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I. 2017. Attention is all you need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**. 6000–6010.