KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data

Autores:	André Eusébio - 127600
	Filipe Pereira - 131531
Docentes:	Prof.ª Doutora Ana Almeida
	Prof.ª Doutora Isabel Machado Alexandre
Unidade Curricular:	Conhecimento e Raciocínio em Inteligência Artificial
Curso:	Mestrado em Inteligência Artificial
Ano letivo:	2024/2025

Conteúdo

1 Objetivo

Como projeto final da unidade curricular de "Conhecimento e Raciocínio em Inteligência Artificial", foi-nos proposto analisar um artigo ciêntifico acerca de um dos temas discutidos durante as aulas ou fortemente relacionado com esta unidade currcicular.

O artigo escolhido foi o **"KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data"** de ADDI AIT-MLOUK (investigador do departamento de ciências da computação da "Umeå University"na Suécia) e LILI JIANG (investigadora no departamento de bases de dados e de sistemas de informação no "Max-Planck-Institut für Informatik"na Alemanha)[1]. Este é um artigo que mostra o estado da arte no mundo dos *chatbots* atual e propõe um *chatbot* baseado em bases de conhecimento para responder às perguntas dos utilizadores.

2 Introdução

O avanço da tecnologia tem transformado de forma significativa a maneira como interagimos com a informação. Neste contexto, o artigo *"KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data"* apresenta um marco relevante nesta evolução. O KBot é descrito como uma solução inovadora que combina inteligência artificial, aprendizagem automática (*machine learning*) e dados interligados (*Linked-Data*) para criar um chatbot capaz de compreender e responder a perguntas em linguagem natural com elevada eficiência.

O principal objetivo do KBot é enfrentar os desafios que ainda limitam muitos sistemas de chat: interpretar corretamente a intenção do utilizador, integrar múltiplas bases de conhecimento e operar em diferentes línguas. Este sistema supera as capacidades dos assistentes virtuais convencionais, ao incorporar dados de fontes *open-source*, como DBpedia e Wikidata, além de explorar o conjunto de dados *myPersonality*, reconhecido pela vasta quantidade de informações sociais e psicológicas.

O KBot distingue-se pela sua capacidade de realizar consultas analíticas complexas, como estatísticas demográficas ou traços de personalidade. A arquitetura modular do sistema permite uma adaptação eficiente do *chatbot* a novas áreas, tanto na educação como no campo da investigação científica.

Com uma interface intuitiva e uma abordagem escalável, o KBot não só facilita o acesso a dados complexos, como também demonstra como a integração de dados semânticos pode transformar o nosso quotidiano. Num cenário caracterizado por uma abundância de informação disponível na internet e em bases de conhecimento *open-source*, soluções como esta oferecem um acesso mais eficiente, interativo e relevante aos dados, o que representa um avanço significativo na forma como gerimos informação e conhecimento.

3 Arquitetura e Tecnologias do KBot

A arquitetura do KBot é estruturada de forma modular, permitindo que cada módulo tenha uma função específica no processo de interação com o utilizador e no processamento das perguntas.

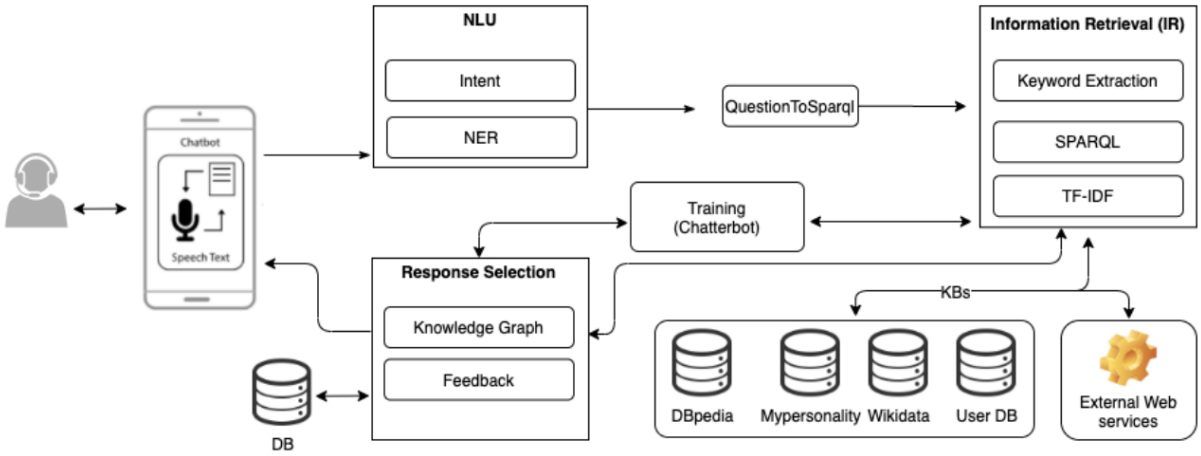

Figura 1: Kbot Overview (fonte DOI: 10.1109/ACCESS.2020.3016142)

Essa estrutura oferece várias vantagens:

- **Escalabilidade:** A arquitetura permite fácil expansão com a adição de novos módulos, como bases de conhecimento ou novos tipos de consultas.
- **Adaptabilidade:** O KBot pode ser configurado para lidar com diversos domínios de conhecimento, desde perguntas simples, como *analytical queries* (ex: "Qual é a capital da França?"), até questões complexas envolvendo dados analíticos ou comportamentais, como as do *dataset* myPersonality.
- **Integração Multidomínio:** Cada módulo interage com uma ou mais fontes de dados, como DBpedia, Wikidata ou dados sociais (ex: myPersonality), permitindo que o KBot consulte e integre informações de várias fontes simultaneamente.

3.1 Módulo de Processamento de Linguagem Natural (NLU)

Este módulo é responsável por interpretar as entradas do utilizador (consultas em linguagem natural) e de as converter em informações estruturadas que o sistema possa processar. Ele é composto por três sub-módulos principais:

3.1.1 Deteção da Língua

A deteção de idioma é o primeiro passo no processamento do *input* do utilizador. Antes de poder compreender o conteúdo da pergunta, o sistema precisa de identificar corretamente a língua na qual esta foi feita, especialmente em sistemas multilingues como o KBot. Este processo é realizado com a ajuda da biblioteca `langdetect`.

- **Teoria**: A biblioteca `langdetect` utiliza um algoritmo probabilístico que analisa o texto do *input* e atribui uma probabilidade para diferentes línguas. Com base nessas probabilidades, o sistema escolhe o idioma mais provável do *input*.

- **Exemplo**: Se o utilizador fizer a pergunta "Qual é a capital da França?", o sistema irá identificar automaticamente que a pergunta foi feita em português e, dependendo da configuração das bases de conhecimento e de dados, procurará a resposta em português ou em inglês.

3.1.2 Intenção

Após detetar o idioma do `input`, o próximo passo é entender qual a **intenção** por detrás da pergunta. Este processo permite ao KBot compreender o objetivo do utilizador, o que é essencial para determinar qual o tipo de informação que deve ser fornecida.

- **Teoria:** O KBot utiliza um modelo de aprendizagem automática, que neste caso é o SVM (*Support Vector Machine*), para classificar a intenção do `input`. O SVM é um algoritmo de classificação supervisionada que utiliza classificação linear ou não linear de modo a separar os dados em diferentes classes, com base nas características extraídas.

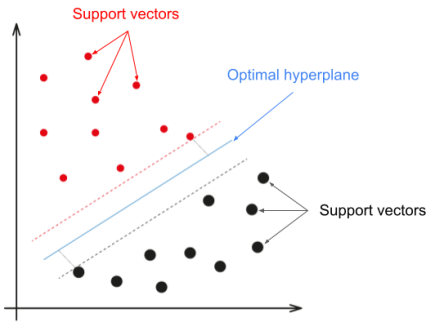

Figura 2: Representação gráfica do modelo SVM (fonte: techslang.com)

  - **Treino:** O modelo SVM é treinado com um conjunto de exemplos de perguntas, onde cada pergunta já está associada à sua intenção. Por exemplo, a pergunta 'Qual é a capital da França?' é associada à intenção 'informação geográfica'.
  - **Classificação:** Quando o utilizador faz uma pergunta, o KBot analisa as palavras-chave e utiliza o modelo treinado para identificar qual a intenção por detrás da pergunta. As intenções podem incluir perguntas factuais, definição de termos, perguntas analíticas, etc.

- **Importância:** A classificação de intenções é fundamental porque permite que o KBot compreenda o tipo de resposta que deve fornecer. Com esta classificação correta, o KBot pode consultar as bases de conhecimento de forma mais precisa.

- **Exemplo:** Se o utilizador perguntar "Quem foi Alan Turing?", o sistema classifica esta pergunta como uma intenção de "obter informações sobre uma pessoa". Com base nisso, o KBot procurará informações sobre Alan Turing nas KB's e DB's.

3.1.3 Reconhecimento de Entidades Nomeadas (NER)

O último passo no processo de NLU é o **Reconhecimento de Entidades Nomeadas (NER)**. Após classificar a intenção, o KBot identifica as entidades-chave que estão presentes na pergunta. As entidades são os elementos principais que precisam ser extraídos para que a pergunta seja processada corretamente nas bases de conhecimento.

- **Como funciona:** O **NER** utiliza técnicas de **processamento de linguagem natural (NLP)** para identificar e classificar as entidades no texto. As entidades podem ser pessoas, locais, datas, organizações ou outras informações importantes. O sistema usa o contexto das palavras próximas (**Clusters**) de modo a classificar corretamente estas entidades e entender o seu papel na consulta.
  - **Exemplo de entidades:** Na pergunta "Quem é o presidente de Portugal?", as entidades seriam "presidente"(como o tipo de informação procurada) e "Portugal"(como o local).
- **Importância:**Identificar as entidades corretas é crucial para o KBot gerar a **SPARQL query** certa, direcionada às bases de conhecimento corretas. Sem um bom reconhecimento de entidades, o KBot poderia construir consultas incorretas ou imprecisas.

Estes três passos permitem que o KBot entenda o contexto completo da pergunta do utilizador, o que é essencial para a criação de **SPARQL queries** para fornecer respostas relevantes e coerentes.

3.2 Módulo de Criação de SPARQL Queries

O **SPARQL** é uma linguagem de consulta usada para recuperar e manipular dados armazenados no formato **RDF (*Resource Description Framework*)**. A *SPARQL query* permite que o sistema recupere dados estruturados a partir de várias fontes de **linked-data**, como a **DBpedia**, a **Wikidata** e entre outros.

O módulo traduz as intenções do utilizador e as entidades extraídas pela NLU numa SPARQL query que é depois executada na base de conhecimento. Por exemplo, para a pergunta: `"Quais são os filmes dirigidos por Christopher Nolan?"`. Neste caso, a NLU identifica a intenção de `obter filmes por diretor` e extrai a entidade `Christopher Nolan`. A partir disso, o módulo gera a seguinte SPARQL query:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbr: <http://dbpedia.org/resource/>

SELECT ?filme
WHERE {
  ?filme dbo:director dbr:Christopher_Nolan .
  ?filme dbo:title ?titulo.
}
```

Listing 1: Exemplo de uma SPARQL Query

Quando executada numa base de conhecimento como a *DBpedia*, esta *query* retorna uma lista de filmes dirigidos por Christopher Nolan, como *Inception*, *Interstellar* e *The Dark Knight*.

3.3 Módulo de Recuperação de Informação

Este módulo é responsável por executar as SPARQL *queries* nas bases de conhecimento e receber os resultados relevantes. A query é enviada para fontes como a **DBpedia**, a **Wikidata** e a base de dados **myPersonality**, para obter as informações solicitadas. Além disso, o módulo também processa, organiza e formata as respostas, garantindo que sejam apresentadas ao utilizador de forma clara, estruturada e objetiva.

Uma das técnicas usadas no módulo para melhorar a relevância das respostas é o **TF-IDF** (*Term Frequency-Inverse Document Frequency*). O **TF-IDF** é utilizado para avaliar a importância das palavras nas respostas recebidas, ajudando a identificar as informações mais relevantes. O processo envolve duas partes principais:

- **TF (*Term Frequency*):** Mede a frequência com que uma palavra aparece num documento ou resposta. Quanto mais vezes uma palavra aparecer, maior será a sua importância dentro daquele documento específico.
- **IDF (*Inverse Document Frequency*):** Avalia a raridade de uma palavra em todo o conjunto de dados. Palavras que aparecem em poucos documentos têm um **IDF** mais alto, o que as torna mais significativas.

O **TF-IDF** combina essas duas métricas para calcular um valor que determina a importância de uma palavra num conjunto de documentos. Assim, o KBot pode selecionar as palavras mais relevantes para uma *query*, melhorando a precisão e a qualidade das respostas fornecidas ao utilizador.

3.4 Módulo de Apresentação de Respostas

Este módulo é responsável por apresentar as respostas ao utilizador. Quando o KBot recupera informações das KB's e das DB's, estas são organizadas e formatadas de maneira que sejam fáceis de entender e visualmente atraentes.

Para isso, são usados **painéis de conhecimento** (*knowledge panels*) para exibir as respostas. Dependendo da pergunta, a resposta pode incluir elementos adicionais, como mapas interativos para perguntas geográficas, ou gráficos para exibir dados estatísticos.

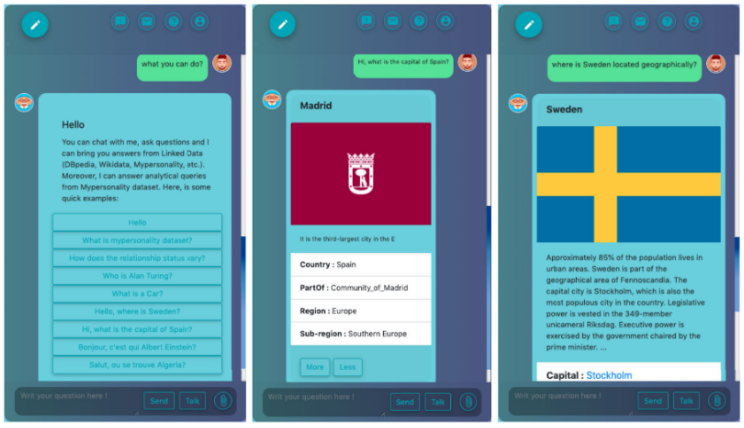

Figura 3: Respostas do KBot (Fonte DOI: 10.1109/ACCESS.2020.3016142)

3.5 *Feedback* do Utilizador e Treino Contínuo no KBot

O *feedback* dos utilizadores é usado de modo a melhorar continuamente o desempenho. As *queries* e o *feedback* são armazenados e analisados para identificar padrões, lacunas na compreensão e respostas ineficazes. Esses dados são usados para:

- **Ajustar modelos de aprendizagem automática:** *Retraining* do classificador de intenções (SVM) para melhorar a precisão.
- **Corrigir falhas:** Refinar a interpretação de intenções e a geração de *queries* SPARQL.
- **Expandir capacidades:** Adaptar-se a novas intenções, línguas e bases de conhecimento.

Este processo garante que o KBot possa evoluir com as interações, oferecendo respostas cada vez mais personalizadas.

3.6 Integração de Múltiplas Fontes de Dados e Escalabilidade

O sistema apresenta uma alta flexibilidade quanto às fontes de dados que pode consultar e, também, planeado de modo a suportar um grande volume de *queries* simultaneamente, sem comprometer o desempenho. Além de bases de conhecimento tradicionais como **DBpedia** e **Wikidata**, também é possível integrar dados de outras fontes, como o dataset **myPersonality**, e até mesmo fontes externas, como **APIs**.

3.6.1 DBpedia, Wikidata e myPersonality

Entre as fontes de dados mais importantes estão:

- **DBpedia e Wikidata:** Algumas das maiores fontes `open-source` semânticos e de *linked-data* disponíveis na internet. Facilitam a recuperação de informações sobre uma vasta gama de tópicos, desde pessoas, lugares e até conceitos mais abstratos.
- **myPersonality:** *Dataset* que contém dados psicológicos e sociais de milhões de utilizadores do *Facebook*. Proporcionando uma capacidade enorme de responder a perguntas analíticas, como tendências políticas ou comportamentos baseados em dados sociodemográficos. A integração deste recurso amplia significativamente as funcionalidades, permitindo responder a questões relacionadas com traços de personalidade, estado civil, entre outros.

3.6.2 Arquitetura Escalável

A arquitetura foi desenvolvida de modo a ser escalável e suportar execução distribuída, o que melhora o tempo de resposta, permitindo consultas paralelas a múltiplas fontes de dados e garantindo respostas rápidas, mesmo com muitos utilizadores.

4 Análise critica artigo

4.1 Alternativas ao Uso de SVM e TF-IDF em Processamento de Linguagem Natural

O KBot utilizou o modelo SVM combinado com o TF-IDF, mas existem outras técnicas que poderiam ser exploradas para melhorar a análise e o processamento de linguagem natural (**NLP**). Algumas alternativas incluem:

- **BERT (*Bidirectional Encoder Representations from Transformers*)**: O BERT compreende o contexto completo das palavras, melhorando a capacidade do *chatbot* de entender intenções complexas e gerar respostas precisas.
- **Redes Neurais + *Word Embeddings***: Utilizar redes neurais com *word embeddings* permite ao *chatbot* capturar a semântica das palavras, proporcionando interações mais naturais e fluídas.
- ***Deep Learning* (CNNs + RNNs)**: A combinação de CNNs para extração de características locais e RNNs para capturar sequências temporais melhora a compreensão e o contexto nas conversas.
- **"Naive Bayes"+ TF-IDF**: Método simples e eficiente para categorizar intenções e respostas rápidas, adequado para *chatbots* com vocabulário limitado e interações diretas.
- **"XGBoost"+ TF-IDF**: Utiliza o XGBoost para melhorar a precisão na classificação de intenções e sentimentos, sendo ideal para *chatbots* que precisam processar grandes volumes de dados rapidamente.

4.2 Dependência de Bases de Conhecimento *Open-Source*

A dependência do KBot em bases de conhecimento **open-source** levanta preocupações significativas quanto à veracidade das informações fornecidas. Essa limitação é corroborada por estudos recentes no campo dos *chatbots* e sistemas de diálogo. *Chatbots* baseados em conhecimento estruturado podem enfrentar desafios ao lidar com informações complexas ou em constante mudança, comprometendo a precisão das respostas (Ignaczuk & Ignaczuk, 2022)[4].

4.3 Problemas éticos

O uso do conjunto de dados *myPersonality*, apesar de ser uma valiosa ferramenta para pesquisa, suscita preocupações éticas significativas, particularmente no que diz respeito à privacidade e ao consentimento informado dos utilizadores. Este *dataset*, criado a partir de dados do *Facebook*, inclui informações altamente sensíveis e identificáveis, como traços de personalidade e interações sociais dos participantes. Estudos destacam que muitos utilizadores não estavam cientes da recolha e do uso subsequente de seus dados para fins de pesquisa, o que gera preocupações relacionadas à transparência e à adequação do consentimento obtido (Zimmer, 2010)[7].

Além disso, a divulgação do conjunto de dados com terceiros exacerba os riscos de violação da privacidade. A reutilização de dados pessoais para finalidades não previstas inicialmente é

frequentemente associada à exploração de lacunas regulatórias e à falta de proteção adequada, comprometendo a confiança pública na ciência (Metcalf & Crawford, 2016)[5]. O escândalo do Cambridge Analytica, por exemplo, mostrou como dados de redes sociais podem ser usados para manipular comportamentos em larga escala, evidenciando a necessidade de regulamentações mais rigorosas na proteção de dados pessoais (Zwitter, 2014)[8].

Portanto, enquanto o "myPersonality"pode ser valioso para estudos em psicologia e inteligência artificial, o seu uso deve ser rigorosamente avaliado sob a ótica ética, garantindo que os princípios de privacidade, consentimento e uso responsável dos dados sejam respeitados.

4.4 Escalabilidade e Desafios Técnicos no Uso de Múltiplas Bases de Conhecimento

Embora o KBot demonstre flexibilidade no uso de múltiplas bases de conhecimento, o artigo não aborda os desafios técnicos associados ao crescimento do volume de dados e à complexidade das consultas. Problemas como a gestão de memória, eficiência de busca e resposta em tempo real tornam-se críticos à medida que os dados aumentam exponencialmente ("European Journal of Computer Science and Information Technology", 2021)[2].

Além disso, integrar fontes heterogéneas com dados inconsistentes requer *pipelines* robustos para limpeza e normalização, essenciais para evitar a degradação do desempenho (Shwe & Aritsugi, 2024)[6].

4.5 Abordagem Multilingue

A abordagem multilingue apresentada no artigo carece de uma avaliação rigorosa sobre a precisão e a eficácia do sistema em diferentes idiomas. Embora a implementação de suporte a múltiplas línguas seja um avanço relevante, a eficácia do sistema pode variar significativamente, especialmente em línguas menos representadas. Questões como a disponibilidade limitada dos dados de treino de algumas línguas, diferenças semânticas e estruturais entre línguas e a capacidade do sistema de lidar com diferentes pronúncias ou expressões mais regionais são desafios que merecem atenção.

Estudos comparativos entre o desempenho do *chatbot* em línguas com maior e menor suporte digital poderiam oferecer *insights* valiosos, além de destacar limitações que poderiam ser resolvidas com abordagens específicas, como o uso de *transfer learning* ou de *embeddings* multilingues, como o M-BERT (Barbon & Akabane, 2022)[3]. Dessa forma, uma análise mais detalhada sobre o impacto do multilinguismo no desempenho e na acessibilidade do sistema contribuiria para validar a proposta e ampliar seu alcance.

4.6 Limitações na Avaliação do Modelo

Embora o KBot apresente resultados sobre a precisão das respostas dadas, não há uma avaliação específica sobre o texto gerado pelo módulo de áudio-para-texto e a sua precisão.

A ausência de testes com uma diversidade de utilizadores é uma limitação significativa, pois impede a compreensão de como o sistema lida com diferentes perfis de fala, incluindo variações culturais, linguísticas e de estilos de interação.

Para validar o modelo de forma mais completa, é necessário conduzir testes com um público diversificado, abrangendo diferentes faixas etárias e níveis de ensino. Estes testes são indispensáveis para verificar o desempenho do sistema em cenários reais e garantir que ele seja eficaz em contextos variados.

5 Conclusão

Em conclusão, ao analisarmos o Kbot, desenvolvido em 2020, e levando em consideração os avanços exponenciais nas tecnologias de *Machine Learning* e Inteligência Artificial nos últimos anos, podemos afirmar que o Kbot representa uma abordagem sólida para *chatbots* mais simples. Ele faz um uso eficaz dos dados disponíveis na internet, sendo uma ferramenta prática e acessível.

No entanto, com o rápido progresso nas técnicas de Processamento de Linguagem Natural e em modelos mais avançados, como o BERT, o Kbot podia beneficiar de atualizações para explorar melhor as complexas interações contextuais e melhorar a compreensão das intenções dos utilizadores, algo que modelos mais recentes já conseguem fazer de maneira mais eficiente. Assim, o Kbot serve como uma boa base, mas há espaço para evolução à medida que novas tecnologias continuam a ser desenvolvidas.

É importante alertar que não existe uma versão do KBot disponível ao público, logo todas as considerações que foram feitas neste artigo foram apenas com base na descrição fornecida pelos autores do artigo original.

6 Referências

[1] Addi Ait-Mlouk e Lili Jiang. "KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data". Em: *IEEE Access* PP (ago. de 2020), pp. 1–1. DOI: 10.1109/ACCESS.2020.3016142.

[2] Unknown Author. "European Journal of Computer Science and Information Technology". Em: *European Journal of Computer Science and Information Technology* (jun. de 2021). DOI: 10.37745/ejcsit.2013.

[3] Rafael Silva Barbon e Ademar Takeo Akabane. "Towards Transfer Learning Techniques — BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study". Em: *Sensors* 22.21 (out. de 2022), p. 8184. DOI: 10.3390/s22218184. URL: https://doi.org/10.3390/s22218184.

[4] Carolina Ignaczuk e Carolina Ignaczuk. *Base de conhecimento open source: prós e contras*. Jul. de 2022. URL: https://conteudo.movidesk.com/base-conhecimento-open-source/.

[5] Jacob Metcalf e Kate Crawford. "Where are human subjects in Big Data research? The emerging ethics divide". Em: *Big Data Society* 3.1 (jun. de 2016). DOI: 10.1177/2053951716650211.

[6] Thanda Shwe e Masayoshi Aritsugi. "Optimizing Data Processing: A comparative study of big data platforms in edge, fog, and cloud layers". Em: *Applied Sciences* 14.1 (jan. de 2024), p. 452. DOI: 10.3390/app14010452.

[7] Michael Zimmer. ""But the data is already public": on the ethics of research in Facebook". Em: *Ethics and Information Technology* 12.4 (jun. de 2010), pp. 313–325. DOI: 10.1007/s10676-010-9227-5.

[8] Andrej Zwitter. "Big Data ethics". Em: *Big Data Society* 1.2 (jul. de 2014). DOI: 10.1177/2053951714559253.