Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

ERC4AI: Uma Ferramenta para Classificação de Requisitos Éticos em IA

Elizangela de Freitas Ximenes

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientadora
Prof.a Dr.a Edna Dias Canedo

Brasília
2025

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

ERC4AI: Uma Ferramenta para Classificação de Requisitos Éticos em IA

Elizangela de Freitas Ximenes

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof.a Dr.a Edna Dias Canedo (Orientadora)
CIC/UnB

Prof. Dr. Geraldo Pereira Rocha Filho
CIC/UnB

Prof.a Dr.a Carla Taciana L. L. S. Schuenemann
CIN/UFPE

Prof. Dr. Rodrigo Bonifácio de Almeida
Coordenador do Programa de Pós-graduação em Informática

Brasília, 10 de Fevereiro de 2025

Dedicatória

Dedico este trabalho a todos que, com determinação e paixão, continuam a perseguir seus sonhos.

Agradecimentos

Inicio meus agradecimentos rendendo graças a Deus, que nas horas mais árduas me concedeu força e orientação para perseverar nesta caminhada.

Ao meu esposo, Luiz Eduardo, minha profunda gratidão pelo apoio incondicional, pela compreensão nos momentos de minha ausência e pelo suporte incansável que sempre me ofereceu.

À minha mãe, Maria Joana, que acreditou em mim, não sendo apenas minha grande inspiração, mas também a principal incentivadora dos meus estudos e conquistas.

À Dra. Edna Dias Canedo, cuja orientação e apoio inestimáveis foram fundamentais ao longo desta pesquisa. Sua disponibilidade constante e disposição em ajudar fizeram toda a diferença nessa jornada.

Por último e não menos importante, agradeço a todos os professores e colegas do PPGI, que enriqueceram meu aprendizado e ajudaram a moldar minha trajetória acadêmica.

Resumo

Contexto: A operacionalização dos princípios éticos da IA durante a fase de engenharia de requisitos é fundamental para a criação de sistemas de IA eticamente alinhados. No entanto, a classificação de requisitos éticos – ou seja, a categorização de requisitos com base nos princípios éticos que eles refletem – continua sendo um desafio significativo devido à natureza abstrata dos princípios éticos e à falta de ferramentas práticas para dar suporte aos desenvolvedores. **Objetivo:** Apresentamos o EthicalRequirements4AI, um conjunto de dados que compreende 1.093 requisitos anotados, incluindo requisitos éticos em IA e fora do contexto de ética em IA. Também fornecemos a *Ethical Requirements Classifier for AI* (ERC4AI), um modelo para classificar requisitos com base em princípios éticos de IA. **Método:** Os conjuntos de dados Passenger Flow e PROMISE foram usados para construir o conjunto de dados EthicalRequirements4AI, abrangendo 156 requisitos alinhados com 11 princípios éticos de IA e 937 requisitos não éticos. Três modelos de linguagem – XLM-RoBERTa, BERT e DistilBERT – foram avaliados para tarefas de classificação binária e multirrótulo. **Resultados:** O modelo treinado com BERT, uma arquitetura baseada em Transformers, obteve o melhor desempenho, alcançando um F1-score médio ponderado de 95% na classificação binária e 88% na abordagem multirrótulo. Na configuração multirrótulo, o ERC4AI classificou os requisitos nos princípios éticos analisados, com os seguintes F1-scores: Transparência (78%), Responsabilidade (81%), Confiança (56%), Privacidade (65%), Outros Princípios Éticos (53%) e Não Princípios Éticos 98%). **Conclusão:** O ERC4AI oferece uma nova ferramenta para pesquisadores e desenvolvedores abordarem preocupações éticas no início do desenvolvimento da IA. O conjunto de dados e os modelos estão disponíveis publicamente para promover mais pesquisas e avanços na ética prática da IA. Além disso, foi desenvolvido um protótipo de interface gráfica, permitindo que os usuários insiram requisitos para análise e obtenham classificações automáticas.

Palavras-chave: Ética em Inteligência Artificial, Requisitos Éticos, Processamento de Linguagem Natural, Aprendizado de Máquina

Abstract

Context: Operationalizing AI ethical principles during the requirements engineering phase is critical to building ethically aligned AI systems. However, ethical requirements classification —that is, categorizing requirements based on the ethical principles they reflect—remains a significant challenge due to the abstract nature of ethical principles and the lack of practical tools to support developers. **Objective:** We present EthicalRequirements4AI, a dataset comprising 1,093 annotated requirements, including both ethical requirements in AI and outside the context of AI ethics. We also provide the *Ethical Requirements Classifier for AI* (ERC4AI), a model for classifying requirements based on AI ethical principles. **Method:** The Passenger Flow and PROMISE datasets were used to construct the EthicalRequirements4AI dataset, covering 156 requirements aligned with 11 ethical AI principles and 937 non-ethical requirements. Three language models – XLM-RoBERTa, BERT, and DistilBERT – were evaluated for binary and multi-label classification tasks. **Results:** The model trained with BERT, a Transformers-based architecture, achieved the best performance, achieving a weighted average F1-score of 95% in binary classification and 88% in the multi-label approach. In the multi-label setting, ERC4AI classified the requirements into the analyzed ethical principles, with the following F1-scores: Transparency (78%), Responsibility (81%), Trust (56%), Privacy (65%), Other Ethical Principles (53%), and Non-Ethical Principles (98%). **Conclusion:** ERC4AI provides a new tool for researchers and developers to address ethical concerns early in AI development. The dataset and models are publicly available to promote further research and advances in practical AI ethics. In addition, a prototype graphical interface was developed, allowing users to input requirements for analysis and obtain automatic classifications.

Keywords: Ethics in Artificial Intelligence, Ethical Requirements, Natural Language Processing, Machine Learning

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AM Aprendizado de Máquina.

BERT Bidirectional Encoder Representations from Transformers.

ER Engenharia de Requisitos.

ERC4AI Ethical Requirements Classifier for AI.

ES Engenharia de Software.

GANs Generative adversarial networks.

HLEG on AI High-Level Expert Group on Artificial Intelligence.

IA Inteligência Artificial.

IEEE Institute of Electrical and Electronic Engineers.

LLM Large Language Models.

MIT Massachusetts Institute of Technology.

NLG Geração de Linguagem Natural.

NLU Compreensão da Linguagem Natural.

OECD Society of the Organisation for Economic Co-operation and Development.

PLN Processamento de Linguagem Natural.

RE Requisitos Éticos.

RE4AI Requirements Engineering for Artificial Intelligence.

RFs Requisitos Funcionais.

RNFs Requisitos Não Funcionais.

Capítulo 1

Introdução

Este Capítulo aborda o contexto para o desenvolvimento deste trabalho, as motivações que impulsionaram a pesquisa, uma visão geral da solução proposta e a metodologia adotada. Além disso, é apresentada a organização deste documento.

1.1 Contextualização

A Inteligência Artificial (IA) tem despertado interesse tanto na comunidade científica quanto na indústria, impulsionando inovações tecnológicas que permitem às máquinas realizarem tarefas anteriormente exclusivas aos seres humanos [3, 4]. Técnicas emergentes como *Bidirectional Encoder Representations from Transformers* (BERT) [5], *ChatGPT* [6, 7] e as *Generative Adversarial Networks* (GANs) [8] revolucionaram o campo do Processamento de Linguagem Natural (PLN), incluindo tarefas como a compreensão e geração de textos, além de permitirem a criação de conteúdo multimídia [9, 10, 11].

Com o aumento da adoção de sistemas baseados em IA em setores como saúde, educação e finanças, os benefícios que essas tecnologias proporcionam à sociedade tornam-se cada vez mais evidentes. Exemplos incluem diagnósticos médicos aprimorados [12], educação personalizada [13], e detecção de fraudes financeiras [14]. No entanto, juntamente com esses benefícios, surgem desafios éticos complexos, especialmente relacionados ao uso mal-intencionado ou antiético da IA [15]. Roy e O'Neil [16] expressaram preocupação com sistemas de IA que são tendenciosos e injustos, e que podem ser usados para reforçar o *status quo*:

> "Você vê o paradoxo? Um algoritmo processa uma série de estatísticas e chega a uma probabilidade de que uma determinada pessoa pode ser uma má contratação, um tomador de empréstimo arriscado, um terrorista ou um professor miserável. Essa probabilidade é destilada em uma pontuação, que pode virar a vida de alguém de cabeça para baixo" [16, p. 10].

Diante deste cenário, promover a confiabilidade nos sistemas de IA torna-se necessário não apenas para assegurar sua aceitação pela sociedade, mas também para fomentar um desenvolvimento tecnológico responsável. Para Floridi et al. [17], a incorporação da ética na IA pode oferecer uma "dupla vantagem", ao permitir que as organizações aproveitem o valor social das tecnologias de IA, identificando oportunidades, ao mesmo tempo que antecipam e minimizam erros e riscos potencialmente inaceitáveis. Trabalhos como de Li et al. [18] abordam a confiabilidade da IA focando nas etapas de preparação de dados, design, desenvolvimento, implantação, operação, monitoramento e governança dentro do ciclo de vida destes sistemas. No entanto, uma etapa parece ter sido menos considerada: a engenharia de requisitos. A definição clara de requisitos é determinante para o desenvolvimento bem-sucedido de sistemas [19], inclusive os de IA, pois orienta todas as fases subsequentes do ciclo de vida.

Portanto, à medida que a definição de requisitos se torna relevante para a confiabilidade e a ética em sistemas de IA, enfrenta-se o desafio prático de classificação desses requisitos de forma eficaz. Embora seja viável classificar manualmente um número limitado de requisitos, essa abordagem torna-se impraticável à medida que a quantidade de requisitos aumenta para centenas ou milhares [1]. Como apontado por Quba et al. [20], classificar manualmente requisitos é uma tarefa trabalhosa, demorada e custosa. Nesse sentido, o uso de um classificador automatizado apresenta vantagens, como a aplicação de critérios uniformes para classificar requisitos, o que pode levar a maior precisão e consistência em comparação com abordagens exclusivamente manuais [21, 22, 23].

A classificação automatizada de outros tipos de requisitos, como os Funcionais (RFs) e Não Funcionais (RNFs) foi amplamente explorada [1, 24, 25, 26, 27, 28, 29, 30]. No entanto, há uma lacuna significativa em relação à classificação de requisitos éticos em sistemas de IA. Este trabalho se propõe a preencher essa lacuna introduzindo um modelo de classificação específico para identificar requisitos éticos em projetos de IA, usando técnicas de PLN.

O modelo realiza tanto a classificação binária, distinguindo entre Princípio Ético e Não Princípio Ético, quanto a classificação multirrótulo, associando os requisitos éticos a diferentes princípios éticos em IA [31]. Dessa forma, um requisito pode estar vinculado a um ou mais princípios, como transparência, privacidade e responsabilidade , permitindo uma análise mais granular dos dados.

O modelo proposto permitirá que os desenvolvedores avaliem o escopo e a consistência dos requisitos éticos desde os estágios iniciais do projeto, contribuindo para a priorização dos requisitos e fornecendo suporte para estágios subsequentes, como design e codificação no ciclo de vida de desenvolvimento de software, assim como ocorre na classificação de requisitos RFs e RNFs [25]. A ausência de uma metodologia clara para identificar requisitos

éticos pode dificultar o cumprimento de códigos éticos estabelecidos, como os propostos pela Comissão Europeia [32], que visam proteger a sociedade de potenciais riscos de IA ao mesmo tempo em que apoiam a inovação no setor. No melhor do nosso conhecimento, a literatura atual não apresenta um trabalho que tenha desenvolvido um classificador específico para a classificação de requisitos éticos em IA, conforme proposto por nosso estudo. Este achado sugere que nossa abordagem pode representar uma contribuição relevante a este campo pesquisa.

1.2 Problema de Pesquisa

Considerando a crescente importância dos princípios éticos no desenvolvimento de sistemas de IA, torna-se fundamental explorar abordagens que os operacionalizem de forma eficaz. Nesse contexto, a disponibilidade de um conjunto de requisitos com anotações estruturadas — organizadas segundo uma classificação baseada em princípios éticos —, aliada aos avanços nas técnicas de PLN, abre novas possibilidades para a análise e implementação de diretrizes éticas em projetos de IA.

Para investigar como esses recursos podem contribuir para a promoção de práticas éticas no desenvolvimento de sistemas de IA, este trabalho de pesquisa se orienta pelas seguintes questões:

- RQ.1: Como o conjunto de dados de requisitos com anotações estruturadas tem o potencial de contribuir para a operacionalização de princípios éticos no desenvolvimento de sistemas de IA?
- RQ.2: Como as técnicas de PLN podem auxiliar na classificação de requisitos éticos em projetos de IA?

1.3 Justificativa

Os princípios éticos como um conjunto de diretrizes sugiram para garantir que as aplicações de IA sejam desenvolvidas e utilizadas de maneira responsável [33, 34]. Em resposta a isso, organizações têm estabelecido comitês compostos por especialistas em IA, que frequentemente recebem a missão de elaborar documentos de políticas voltados para este foco [35, 36].

Destaca-se pela atuação, a Comissão Europeia *High-Level Expert Group on Artificial Intelligence* (HLEG on AI), com a proposição da diretriz *Ethics guidelines for Trustworthy AI* [37], que tem por objetivo lidar com os impactos, desenvolvimentos e implicações éticas da IA. Além disso, a Comissão Europeia elaborou a Lei de IA da União Europeia [32],

visando regular sistemas e modelos de IA considerados de alto risco. Essa legislação impõe rigorosos requisitos regulamentares às organizações atuantes no mercado europeu, estabelecendo penalidades severas para o descumprimento, com o intuito de proteger a sociedade contra os riscos potenciais da IA, ao mesmo tempo em que apoia a inovação no setor.

Com um grupo de especialistas, a *Society of the Organisation for Economic Co-operation and Development* (OECD)[1] listou em suas pautas o compromisso com uma IA confiável e centrada no ser humano. Este compromisso foi endossado por diversos países, incluindo o Brasil. O *Advisory Council on the Ethical Use of Artificial Intelligence and Data*[2] foi estabelecido pelo governo de Singapura para lidar com questões éticas e legais relacionadas ao uso da IA e dados.

O *Select Committee on Artificial Intelligence of the UK House of Lords*, comitê especial estabelecido pela Câmara dos Lordes do Reino Unido, foi criado para considerar as implicações econômicas, éticas e sociais da crescente utilização da IA, que teve como contribuição o relatório *AI in the UK: ready, willing and able?* [38].

Os esforços não restringem-se apenas às instituições governamentais. Entidades como o *Institute of Electrical and Electronic Engineers* (IEEE) lançou o documento intitulado *IEEE Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, que fornece diretrizes sobre ética e autonomia na IA [39]. A OpenAI[3], conhecida por suas contribuições na área de pesquisa de aplicações em IA, tem uma missão declarada para garantir que a IA beneficie a humanidade, e como consequência, tem escrito sobre suas abordagens éticas. Empresas consolidadas como a Google[4] e Microsoft[5] lançaram seus próprios princípios de IA que delineiam o compromisso em desenvolver e usar a IA de uma maneira que beneficie a sociedade e evite causar dano ou reforçar preconceitos e injustiças. Além disso, essas duas empresas em conjunto com o Amazon[6], Apple[7], Facebook[8] se juntaram para estudar e formular melhores práticas na pesquisa e desenvolvimento de IA[9].

Apesar da disponibilidade de diretrizes éticas, há uma falta de orientação prática, e os desenvolvedores normalmente não sabem como operacionalizar tais princípios abstratos e

[1] https://www.oecd.org/going-digital/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm

[2] https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/archived/imda/press-releases/2018/composition-of-the-advisory-council-on-the-ethical-use-of-ai-and-data

[3] `https://openai.com/charter`

[4] `https://ai.google/responsibility/principles/`

[5] `https://www.microsoft.com/en-us/ai/principles-and-approach`

[6] `https://aws.amazon.com/pt/generative-ai/`

[7] `https://www.apple.com/br/`

[8] `https://ai.meta.com/`

[9] https://partnershiponai.org/resources/

de alto nível [40, 41, 42]. A operacionalização dos princípios éticos de IA deve ocorrer desde os primeiros estágios do desenvolvimento de software, ou seja, engenharia de requisitos, em vez de uma reflexão tardia [40]. Dentro desta etapa, a classificação de requisitos é importante para o gerenciamento eficaz de projeto, pois pode melhorar a utilização de recursos, documentação e reduzir o retrabalho [1].

A criação de um modelo classificador de requisitos éticos pode ser uma abordagem relevante para automatizar a identificação e categorização desses requisitos em sistemas de IA. Utilizando técnicas de PLN e AM, esse modelo pode reduzir o esforço manual e mitigar riscos de omissões ou interpretações inadequadas. Além disso, classificadores automatizados podem apoiar a conformidade com regulamentações e boas práticas, promovendo o desenvolvimento de sistemas de IA mais transparentes e responsáveis.

1.4 Objetivo Geral

O objetivo geral deste trabalho é o desenvolvimento de um classificador automatizado, denominado *Ethical Requirements Classifier for AI (ERC4AI)*, que utilizará técnicas de Aprendizado de Máquina (AM) para identificar requisitos baseados em princípios éticos em IA.

1.4.1 Objetivos Específicos

Para atingir o objetivo geral deste trabalho, os seguintes objetivos específicos foram definidos:

- Identificar requisitos éticos em IA nas bases de dados textuais de requisitos de software, disponíveis em repositórios públicos. A escolha da base de dados PROMISE [1] deve-se à sua composição por requisitos de sistema tradicionais, os quais, conforme indicado por Carleton et al. [43], apresentam uma interseção significativa com requisitos baseados em IA. Já a base *Passenger Flow* [44] foi selecionada por ser uma fonte de dados especificamente de requisitos éticos em IA.

- Preparar as bases de dados PROMISE [1] e *Passenger Flow* [44], atribuindo rótulos que identifiquem requisitos éticos em IA, quando necessário. Esta tarefa é parte essencial da etapa de desenvolvimento do modelo classificador.

- Desenvolver modelos classificadores utilizando algoritmos de AM, como BERT[10], XLM-RoBERTa[11] e DistilBERT[12].

[10] `https://huggingface.co/google-bert/bert-large-uncased`
[11] `https://huggingface.co/FacebookAI/xlm-roberta-large`
[12] `https://huggingface.co/distilbert/distilbert-base-multilingual-cased`

Avaliar a performance dos modelos desenvolvidos, empregando as métricas de desempenho Precisão [2], *Recall* [2], *F1-Score* [2], *Weighted AVG [45] e Macro Average (AVG) [2, 46]*, a fim de identificar e selecionar o modelo com a melhor eficácia.

1.5 Resultados Esperados

Esta pesquisa visa alcançar os seguintes resultados:

- Desenvolver um modelo classificador de requisitos éticos para sistemas de IA, utilizando técnicas de PLN e AM.
- Avaliar o desempenho do modelo classificador por meio de métricas de desempenho (Precisão [2], *Recall* [2], *F1-Score* [2], *Weighted AVG [45] e Macro AVG [2, 46]*), investigando sua eficácia na identificação de requisitos éticos.

1.6 Contribuições do Trabalho

Esta pesquisa apresenta as seguintes contribuições:

- Disponibilizar em repositório público o conjunto de dados rotulados com requisitos éticos em IA, visando apoiar pesquisadores e profissionais no desenvolvimento de modelos classificadores para aplicações futuras.
- Disponibilizar classificadores de texto flexíveis — em versão binária e multirrótulo —, que possam ser adaptados e reutilizados em diferentes contextos e projetos relacionados à análise de requisitos éticos para sistemas de IA.
- Realizar uma análise comparativa de modelos de linguagem bem estabelecidos, como XLM-RoBERTa [47], BERT [5] e DistilBERT [48], tanto para abordagens de classificação binária quanto multirrótulo, com o objetivo de identificar o modelo mais adequado para a classificação de requisitos éticos, considerando seu desempenho em tarefas de classificação.

1.7 Metodologia de Pesquisa

Nesta pesquisa, é adotada uma abordagem multifásica alinhada com o *Cross Industry Standard Process for Data Mining* (CRISP-DM) [49], um padrão da indústria para mineração de dados que detalha as etapas do projeto, suas tarefas e interconexões. Esta metodologia, empregada com adaptações em estudos anteriores como os de Gil et al. [50]

e Solano et al. [51], oferece flexibilidade ao permitir avanços e revisões entre as fases do projeto. A abordagem seguida é ilustrada na Figura 1.1.

Na *Fase 1*, foram concentrados esforços em uma revisão bibliográfica sobre requisitos éticos em IA. Este estudo envolveu a análise de artigos acadêmicos, visando compreender as principais questões que envolvem este tema, e as estratégias atuais para abordá-las. Essa revisão de literatura contribuiu para identificar lacunas no processo de classificação de requisitos éticos em IA.

Na *Fase 2*, a obtenção de textos de requisitos de sistemas foi realizada em bases de dados públicas, com o propósito de construir um conjunto de dados de requisitos éticos em IA para desenvolvimento do ERC4AI. Dentre as bases consideradas, têm-se a *PROMISE* [1], escolhida por conter requisitos de sistema tradicionais, os quais apresentam uma interseção com requisitos baseados em IA [43]. Adicionalmente, a base *Passenger Flow* [44] foi escolhida por sua especificidade em requisitos éticos em IA.

Na *Fase 3*, as bases de dados *Passenger Flow* [44] e *PROMISE* [1] foram submetidas à análise e rotulação. O processo começou com a base *Passenger Flow*, focando na compreensão dos requisitos. Foi identificado que, apesar de cada requisito ético em IA estar ligado a um princípio ético específico, alguns requisitos também refletiam princípios adicionais. Essa observação, baseada nos princípios e questões éticas discutidos por Ryan e Stahl [31] e documentados na Seção 2.1.1, motivou uma reavaliação e reclassificação dos requisitos em novas categorias éticas de IA, quando necessário. Posteriormente, a base *PROMISE* [1] foi analisada para verificar requisitos convencionais que poderiam ser enquadrados como éticos em IA, utilizando uma abordagem de rotulação similar à aplicada em *Passenger Flow* [44]. Para assegurar a precisão e imparcialidade, foram incluídos mais participantes no processo de rotulação. Cada participante avaliou os requisitos considerados éticos nesta pesquisa, sem classificações prévias, para evitar o viés de confirmação [52], que é a tendência de buscar, interpretar ou priorizar informações que confirmem crenças ou hipóteses preexistentes, ignorando evidências contrárias. Além disso, foi realizada uma análise do grau de concordância entre as rotulações dos diferentes avaliadores, utilizando o coeficiente AC1 de Gwet [53], a fim de avaliar a consistência e a confiabilidade do processo. Por fim, os dados rotulados por diferentes avaliadores foram consolidados em um único conjunto, chamado *EthicalRequirements4AI*, para a criação do modelo ERC4AI.

Na *Fase 4*, ocorreu o pré-processamento e padronização dos dados. Essa fase incluiu a remoção de múltiplos espaços, pontuação, números; assim como a estruturação dos rótulos (binário e multirótulo) para a etapa de treinamento do ERC4AI.

Na *Fase 5*, ocorreu a implementação do processo de AM, utilizando técnicas de PLN para identificar automaticamente as características mais relevantes nos textos de requisitos que possuíam indicativos da presença ou ausência de padrões de princípios éticos em

IA. Como apresentado em estudos [5, 54, 48, 47], essa estratégia envolve a construção de um vetor de características, onde cada palavra é analisada dentro de seu contexto específico, permitindo a compreensão das nuances e do significado subjacente em diferentes cenários de uso. Esse processo foi realizado na plataforma *Google Colaboratory* [13], um serviço baseado em nuvem para execução de código em *Python*[14]. Foram consideradas as métricas de Precisão [2], *Recall* [2], *F1-Score* [2], *Weighted AVG* [45] e Macro AVG [2, 46], permitindo uma análise qualitativa dos acertos e falhas do modelo.

Na *Fase 6*, o resultado do ERC4AI foi avaliado, comparando-o com outros estudos e discutida suas contribuições práticas e acadêmicas. Também foram abordadas as limitações do trabalho, com foco nas ameaças à validade e nas estratégias adotadas para realizar a mitigação delas durante o processo de desenvolvimento.

Figura 1.1: Fases de desenvolvimento da pesquisa

1.8 Estrutura do Trabalho

Esta dissertação de mestrado está organizada em 6 capítulos. A seguir, serão apresentados os demais capítulos e seus respectivos resumos.

- **Capítulo 2:** apresenta a fundamentação teórica necessária para compreensão do problema de pesquisa. Neste contexto, também são introduzidos os trabalhos correlatos identificados durante a revisão de literatura.

[13] `https://colab.research.google.com/`
[14] `https://www.python.org/`

- **Capítulo 3:** descreve as etapas do desenvolvimento do modelo *Ethical Requirements Classifier for AI (ERC4AI)*, abrangendo os métodos e práticas empregadas neste processo.
- **Capítulo 4:** detalha a análise comparativa das métricas e resultados dos modelos de IA desenvolvidos, para seleção do modelo de melhor desempenho nos contextos binário e multirótulo.
- **Capítulo 5:** discute as contribuições e ameaças para validade deste trabalho em relação aos resultados alcançados.
- **Capítulo 6:** descreve as principais conclusões desta pesquisa e aborda possíveis direções para trabalhos futuros.

Capítulo 2

Revisão de Literatura

Este capítulo explora a Ética em IA, adentra no campo do Processamento de Linguagem Natural, e ao final, ressalta-se a distinção deste trabalho em comparação com outros identificados na literatura que abordam as questões éticas em IA.

2.1 Ética em IA

Apesar dos progressos reconhecidos da IA como uma inovação disruptiva, há um contrassenso entre profissionais da área, gerando assim, um debate em torno do seu valor [55]. Tais controvérsias têm sido evidentes desde os primeiros registros da criação desta tecnologia.

No *Massachusetts Institute of Technology*(MIT), durante a década de 1970, Weizenbaum desenvolveu um dos primeiros experimentos em IA, que simulava um terapeuta. Esse programa foi batizado de ELIZA [56], um fenômeno desde o início. Observando a resposta do público ao programa, o cientista percebeu que muitas pessoas se sentiam confortáveis em se comunicar com computadores, além de chegar ao ponto de considerá-lo um psicoterapeuta eficaz. Para Weizenbaum essa ideia era "perversa", levando-o a se questionar como as pessoas percebiam os computadores e a IA, sejam elas do público em geral ou especialistas da área. Ele se perguntava se o pensamento humano poderia ser totalmente transformado em formalismo lógico [57]. Wiener [58], um matemático notável e pioneiro na cibernética, também reconheceu o potencial que a computação tinha de provocar mudanças éticas significativas na sociedade.

O renomado físico teórico Stephen Hawking via a IA como uma força potencialmente milagrosa e catastrófica, enfatizando que possivelmente ela poderia ser "o último [evento em nossa história], a menos que aprendamos como evitar os riscos"[1]. Elon Musk, empre-

[1]https://time.com/4278790/smart-people-ai/

sário e fundador da SpaceX[2] e CEO da Tesla Motors[3], é conhecido por chamar a IA de "nossa maior ameaça existencial".[4]

Os dilemas éticos com o passar do tempo despertaram reflexões acerca do desenvolvimento, implementação e uso das tecnologias digitais [59]. Conforme Ryan e Stahl [60] mencionou, essas tecnologias são extremamente maleáveis e sujeitas à diferentes interpretações. Portanto, podem ser utilizadas para diversos fins, que nem sempre refletem a intenção original de seus criadores, tornando o impacto da IA ainda mais significativo.

Segundo Vakkuri et al. [40], o debate sobre a ética em IA intensificou-se na última década, impulsionado pelo avanço tecnológico nesta área. Isso levou à formulação de princípios-chave, hoje reconhecidos como questões centrais da ética em IA. Este campo concentra-se no estudo de questões éticas relacionadas à forma como a IA pode ou deve ser desenvolvida e implementada de maneira responsável e justa, explorando princípios, regras, diretrizes, políticas e regulamentos associados [61].

2.1.1 Princípios Éticos

Nos últimos anos, têm surgido diretrizes éticas que consolidam princípios fundamentais, com o propósito de encorajar os desenvolvedores a adotá-los, na medida do possível, como base para a construção de aplicações em IA [62]. Segundo Cerqueira et al. [33], grande parte dos estudos relativos à ética em IA tende a se concentrar na parte conceitual do assunto, envolvendo a compilação, apresentação e avaliação de diretrizes éticas e dos princípios que as norteiam. Jobin et al. [35] destacaram uma convergência significativa nos princípios fundamentais dos documentos orientativos sobre este tema. Contudo, a despeito das semelhanças, há divergências em interpretações, justificativas para sua importância, domínios e atores envolvidos, bem como as estratégias de implementação. Todos esses pontos podem ocasionar o risco de repetição desnecessária, sobreposição, confusão e ambiguidade [41].

Diversos autores [41, 63, 64] têm dedicado esforços no sentido de sintetizar os princípios éticos, visando assegurar que os avanços e implementações em IA ocorram de maneira não apenas responsável, mas também justa. Neste trabalho, adotaremos o estudo de Ryan e Stahl [31], reconhecido por sua relevância não apenas para a comunidade acadêmica e política diretamente engajadas nesta temática, mas também para o conjunto de usuários que buscam orientação no desenvolvimento ou implementação de sistemas de IA. Na sequência, serão apresentados de forma geral, os 11 princípios abordados por Ryan e Stahl [31] e as questões éticas associadas a eles.

[2]https://www.spacex.com/
[3]https://www.tesla.com/
[4]https://time.com/4278790/smart-people-ai/

1. Transparência

A transparência pode ser vista sob duas perspectivas, tanto em relação à tecnologia de IA em si, quanto às organizações que a desenvolvem e utilizam. Ela é essencial pois protege requisitos como os direitos humanos fundamentais, a privacidade, a dignidade, a autonomia e o bem-estar. Este princípio implica clareza nos objetivos, benefícios, danos e resultados potenciais do uso da IA, possibilitando que os usuários façam escolhas conscientes a respeito do compartilhamento de seus dados e da utilização desta tecnologia[31].

- **Explainabilidade:** Envolve o monitoramento ativo para assegurar que os resultados sejam precisos e confiáveis. As organizações devem documentar e reproduzir suas decisões para possíveis auditorias, garantindo que a IA seja compreensível para entidades de auditoria externa [31].

- **Explicabilidade:** Organizações que trabalham com IA devem ser capazes de explicar dados de entrada e saída, funções algorítmicas e seu objetivo de fazê-lo. Isso deve ocorrer de maneira compreensível, assegurando que dados e decisões sejam rastreáveis e reprodutíveis, para promover segurança e permitir a identificação e a análise das causas raízes em situações de danos ou problemas [31].

- **Compreensibilidade:** Devem ser adotados métodos de monitoramento adequados pela organizações de IA para dados e algoritmos, assegurando que as decisões e ações tomadas pela IA sejam compreensíveis para os humanos. É imperativo que as organizações possam entender e serem capazes de explicar o funcionamento e as decisões tomadas por suas tecnologias de IA, quando possível [31].

- **Interpretabilidade:** As entidades que adotam a IA necessitam compreender os mecanismos decisórios de suas tecnologias, inclusive aquelas com características de "caixa-preta", assegurando a existência de uma supervisão humana para evitar e tratar potenciais danos. Em setores de risco acentuado como saúde, justiça criminal e assistência social, é necessário reavaliar o uso extensivo da IA [31].

- **Comunicação:** Os usuários finais devem ser informados de forma clara e precisa sobre as intenções e resultados das tecnologias de IA, para evitar manipulações e coerções. As empresas que desenvolvem e utilizam esta tecnologia devem comunicar de maneira compreensível possíveis falhas e danos. Tal comunicação pode necessitar de adaptações conforme o setor e o público. Além disso, é fundamental que as organizações mantenham os governos atualizados sobre seus avanços e expectativas de atingir objetivos específicos [31].

- **Divulgação:** A IA deve ser desenvolvida e usada para coletar o mínimo possível de dados pessoais e garantir que quaisquer dados coletados sejam anonimizados, criptografados e processados de forma segura, sendo isso demonstrável a auditores externos. É necessário que a IA seja submetida a auditorias internas e externas, sendo que as organizações possam explicar e justificar seu uso [31].

- **Apresentação:** Empresas devem assegurar e mostrar que seus dados são precisos e atuais, mantendo transparência e monitoramento constante da qualidade dos mesmos. Desenvolvedores de IA precisam tornar seus códigos de ética acessíveis a autoridades, usuários e, quando possível, ao público, implementando revisões. Ademais, é de extrema importância que usuários saibam quando estão interagindo com um sistema de IA, e não com uma pessoa [31].

2. Justiça e Equidade

O tema da discriminação e os desfechos injustos provenientes de algoritmos têm ganhado destaque tanto na mídia quanto no meio acadêmico. É compreensível que tópicos envolvendo justiça, igualdade e equidade tenham sido frequentemente explorados nas diretrizes éticas. Durante o processo de criação de sistemas de IA, os especialistas da área devem determinar os graus de justiça e equidade que podem ser incorporados. Ainda que os criadores de IA possuam valores individuais, é de suma importância que seja evitado incorporar vieses historicamente injustos nos algoritmos que desenvolvem [31].

- **Consistência:** Organizações devem assegurar a coleta, análise e uso de dados amostrais precisos e representativos para prevenir danos na tomada de decisões pela IA. É primordial estabelecer procedimentos que identifiquem e minimizem imprecisões, garantindo alta qualidade de dados, realizando auditorias algorítmicas externas e mantendo diálogo constante com usuários finais e *stakeholders* impactados [31].

- **Inclusão:** A IA deve evitar tornar-se um instrumento de exclusão social, tendo como foco particular a inclusão de grupos sub-representados e vulneráveis. Os dados utilizados precisam ser representativos e inclusivos do público-alvo. Organizações de IA devem não só mitigar problemas de exclusão, mas também fomentar a participação ativa de mulheres e minorias no desenvolvimento destas tecnologias [31].

- **Igualdade:** A IA deve evitar danos e promover a igualdade individual, valendo-se da diversidade em equipes e dados, e adotar medidas contra prejuízos relacionados a sexismo e preconceitos de gênero [31].

- **Equidade:** Em projetos envolvendo IA, é essencial capacitar os indivíduos, assegurar oportunidades iguais e recompensas justas, a fim de promover um uso equitativo e imparcial na sociedade [31].

- **Não-viés:** As organizações devem se dedicar a identificar, abordar e atenuar preconceitos injustos, examinando-os durante todo o desenvolvimento e eliminando-os quando identificados. A atenção deve ser voltada para os dados de treinamento e preconceitos humanos e algorítmicos [31].

- **Não-discriminação:** A IA deve ser criada visando uma aplicação universal, prevenindo discriminação com base em atributos como gênero, raça e idade. É crucial implementar mecanismos que previnam e ajustem resultados que se revelem discriminatórios [31].

- **Diversidade:** É essencial incorporar perspectivas de uma variedade ampla de interessados e incentivar uma multiplicidade de opiniões ao longo de todas as fases de uso da IA [31].

- **Pluralidade:** Criadores de IA devem valorizar perspectivas diversificadas, prevenindo a padronização de comportamentos e práticas. É crucial que as organizações não apenas modifiquem seus "modelos de pipeline", mas também garantam representatividade a todos os integrantes, fomentando uma cultura de inclusão refletida nesta tecnologia [31].

- **Acessibilidade:** Organizações precisam preservar os direitos dos proprietários dos dados, incluindo o acesso à informação, correção ou deleção de dados armazenados sobre eles e explicações claras e gratuitas quando decisões são tomadas a seu respeito [31].

- **Reversibilidade:** Organizações que usam IA devem claramente definir se as decisões tomadas por algoritmos, como a recusa de um empréstimo, são reversíveis e garantir que exista uma opção de revisão e correção em casos de resultados prejudiciais [31].

- **Remediar:** A adoção de medidas preventivas é essencial para prontamente identificar e resolver possíveis prejuízos na utilização da IA [31].

- **Reparação:** Em casos de eventos nocivos ou injustos resultantes da utilização de IA, é essencial que as partes afetadas tenham acesso a medidas de reparação eficazes e transparentes, e de forma ágil [31].

- **Contestação:** As organizações precisam apoiar e proteger aqueles que têm preocupações éticas, garantindo sua segurança ao se manifestarem [31].
- **Acesso e distribuição:** As organizações precisam assegurar que suas tecnologias sejam equitativas e acessíveis para uma ampla variedade de grupos de usuários na sociedade. A IA deve estar acessível para aqueles frequentemente em desvantagem social, como indivíduos com problemas de visão, dislexia ou mobilidade [31].

3. Não-maleficência

A não-maleficência, basicamente, significa não causar dano ou evitar prejudicar outros. Na ética da IA, a maioria das diretrizes destaca fortemente a importância de proteger os cidadãos, assegurando a segurança da IA e tomando medidas preventivas e corretivas quando necessário [31].

- **Segurança:** A IA precisa ser robusta e protegida durante todo o seu ciclo de vida, operando devidamente sem perigos significativos para a segurança. As organizações precisam garantir uma segurança cibernética eficiente, integrando e avaliando medidas preventivas na arquitetura da IA, além de endereçar prontamente qualquer vulnerabilidade ou falha identificada por pesquisadores de segurança [31].
- **Proteção:** Desenvolvedores e usuários de IA devem garantir que a tecnologia respeite os direitos humanos e seja segura, avaliando riscos e aplicando fortes medidas de segurança. A IA precisa passar por testes e garantia de qualidade durante toda a sua implantação, com gestão e procedimentos adequados para qualquer quebra de segurança [31].
- **Dano:** Ao longo do desenvolvimento, é essencial avaliar e documentar os objetivos e os impactos previstos da IA, garantindo uma revisão constante dos seus efeitos. É imprescindível que as organizações incentivem uma "responsabilidade algorítmica" e procedam com cautela ao criar aplicações de IA que possam ter consequências negativas. Qualquer tecnologia do tipo que venha substituir funções humanas deve assegurar uma redução de danos antes de ser introduzida no mercado, prevenindo causar lesões ou aflições emocionais intensas às pessoas [31].
- **Precaução:** Os desenvolvedores de IA devem possuir compreensão profunda sobre as operações e possíveis efeitos desta tecnologia, mantendo as medidas de segurança devidamente registradas, assim como permitir a pausa ou término por intervenção humana diante de possíveis riscos. Assessoria de especialistas legais, éticos e analistas políticos pode ser buscada para auxilio nesse processo [31].

- **Prevenção:** Durante toda a sua vida útil, um sistema de IA deve ser não apenas controlável e gerenciável, mas também robusto e confiável, protegendo-se contra ataques e manipulações. Deve haver ênfase proativa na prevenção de acidentes, tal como situações críticas sempre que possível [31].
- **Integridade:** É imperativo que ataques direcionados à IA não ponham em risco a integridade física e mental dos indivíduos, assegurando a confiabilidade dos sistemas internos. Em circunstâncias de falha, a IA deve agir de maneira segura, como por exemplo, desligando-se ou ativando um modo de segurança [31].
- **Não-Subversão:** Os sistemas de IA devem ser empregados para respeitar e aprimorar a vida dos cidadãos, ao invés de minar os processos sociais e cívicos essenciais à saúde da sociedade [31].

4. Responsabilidade

A preocupação moral sobre quem assume a responsabilidade em contextos de IA é crucial, especialmente quando há temores de empresas ofuscando a culpa em sistemas autônomos ou semi-autônomos. O risco de uma "lacuna de responsabilidade", onde não é evidente quem é o culpado pode surgir devido à autonomia da IA [31].

- **Responsabilidade/*Accountability:*** Organizações e criadores de IA precisam assumir responsabilidade pelas consequências negativas advindas de dados de baixa qualidade e pelo impacto global de seus sistemas, mantendo a transparência e a responsabilidade por meio de auditorias e avaliações. Qualquer dano causado pela IA deve ser atribuído a uma entidade legal, não à ferramenta causadora do dano [31].
- **Responsabilidade/*Liability:*** É indispensável diferenciar desenvolvedores e usuários organizacionais de sistemas IA para atribuição correta de responsabilidade legal em casos de falhas e danos. A responsabilidade final, especialmente para efeitos adversos de sistemas autônomos, pode ser estabelecida por meio de uma manutenção adequada de registros e documentação [31].
- **Agir com Integridade:** Organizações de IA devem assegurar a integridade dos dados em todas as etapas de uso e comunicar erros ou violações de segurança às entidades pertinentes [31].

5. Privacidade

A privacidade refere-se à proteção e ao controle dos dados pessoais e informações sensíveis dos indivíduos durante a coleta, o processamento e o uso desses dados por sistemas de

IA. Este princípio tornou-se um foco crucial, especialmente no desenvolvimento e uso da IA, devido à quantidade de dados pessoais utilizados. Para isso, é fundamental aderir aos regulamentos atuais de proteção de dados e incorporar princípios de privacidade desde a concepção de um projeto [31].

- **Informação Pessoal ou Privada:** Para garantir a privacidade, organizações de IA devem adotar medidas de proteção de dados, assegurar consentimento informado e proporcionar ao usuário acesso e controle sobre suas informações, seguindo simultaneamente normas de privacidade e priorizando a qualidade dos dados. Além disso, é essencial promover uma cultura voltada à esse foco, optando por aplicações de IA desenvolvidas sob essa perspectiva [31].

6. Beneficência

A beneficência, que envolve realizar ações para promover o bem, esta é por vezes subvalorizada nas discussões da ética em IA, sendo presumido que a tecnologia por si só, proporcionará vantagens. É crucial enfatizar a beneficência nas diretrizes de ética em IA, assegurando que ela contribua para o bem-estar individual e coletivo, promovendo paz e bem comum social [31].

- **Bem-Estar:** As organizações de IA precisam assegurar o bem-estar e o desenvolvimento individual, garantindo que suas tecnologias sejam apropriadas e não inibam o crescimento pessoal nem o acesso a recursos essenciais. Deve promover o bem-estar humano, capacitar indivíduos globalmente e ser empregada para realçar e apoiar o trabalho dos profissionais de saúde, melhorando os cuidados e apoiando o bem-estar dos pacientes [31].

- **Paz:** Organizações devem se empenhar para prevenir conflito armamentista em armas autônomas, que podem ser letais. Se a IA representar uma ameaça à paz, é de suma importância que as entidades responsáveis cooperem com os governos a fim de atenuar possíveis conflitos [31].

- **Bem Social:** A IA deve aprimorar as oportunidades vantajosas para a sociedade, com organizações da área cultivando um ecossistema industrial robusto, fundamentado na cooperação e concorrência saudável, sem gerar conflitos com aqueles que não utilizam tais tecnologias [31].

- **Bem Comum:** A IA deve ser criada para contribuir para o bem comum, assim como servir às pessoas. As organizações necessitam avaliar os benefícios e malefícios da tecnologia, considerando estratégias de mitigação de danos para assegurar o bem-estar geral da sociedade [31].

7. Liberdade e Autonomia

Desenvolvedores necessitam identificar e otimizar cenários onde a IA possa ameaçar liberdades humanas. É primordial garantir que a utilização da IA não comprometa as liberdades dos usuários finais, resguardando-os de perigos como rastreamento, censura e vigilância, que podem ameaçar suas liberdades de locomoção, expressão e associação, respectivamente. Organizações devem certificar que os usuários sejam devidamente informados e não sejam enganados ou manipulados pela tecnologia, assegurando que possam exercer sua autonomia [31].

- **Consentimento:** A utilização de dados pessoais deve ser explicitamente acordada antes de seu uso. Se esses dados forem reutilizados, devem seguir os requisitos acordados originalmente. O manuseio dos dados não deve ser feito de forma que o proprietário considere inapropriado e deve sempre alinhar-se com as expectativas e consentimento da pessoa, sendo usado apenas para fins legítimos [31].
- **Escolha:** A IA deve proteger a habilidade dos usuários de fazerem escolhas sobre suas próprias suas vidas e preservar a liberdade e autonomia humana. Não deve limitar secretamente e ilegitimamente suas opções e conhecimentos [31].
- **Auto-Determinação:** É preciso balancear o poder de decisão dado aos sistemas autônomos pelo usuário, especialmente quando esta escolha é restringida pelo sistema. Organizações de IA não devem manipular a decisão das pessoas, em particular as mais suscetíveis a abusos [31].
- **Liberdade/*Liberty:***As organizações que lidam com IA devem assegurar que sua tecnologia proteja as liberdades individuais. Isso inclui garantir liberdades como expressão, reunião e movimento durante o desenvolvimento das aplicações [31].
- **Empoderamento:** A IA deve ser empregada para potencializar e fortalecer nossos direitos humanos, não para limitá-los ou violá-los. Indivíduos devem ter o direito de contestar decisões que possam prejudicar suas liberdades [31].

8. Confiança

A confiança, essencial para interações humanas e o funcionamento da sociedade, está emergindo como requisito crucial para a implementação ética da IA, com diretrizes focando neste princípio fundamental [31].

- **Confiabilidade/*Trustworthiness:*** As organizações precisam provar sua confiabilidade, assegurando que suas tecnologias sejam seguras e eficazes. A confiança é

construída cumprindo promessas, garantindo o funcionamento adequado dos sistemas e protegendo dados de forma responsável, enquanto se promove transparência e responsabilidade [31].

9. Sustentabilidade

A sustentabilidade, é um tema relevante, que vem sendo discutido devido ao seu impacto em diversos campos, incluindo a IA. As Organizações devem assegurar a sustentabilidade ambiental, incorporando considerações ecológicas em suas decisões e incentivando a adoção de energia sustentável e eficiente, além de apoiar a proteção da biodiversidade [31].

- **Meio Ambiente (Natureza):** As organizações devem utilizar a IA elaborada com uma perspectiva ambientalmente consciente e identificar danos ecológicos que ultrapassem limites aceitáveis, pois não deve causar prejuízos à biodiversidade [31].
- **Energia:** A IA deve ser usada de forma ecológica e eficiente, minimizando emissões de gases de efeito estufa e protegendo a biodiversidade [31].
- **Recursos(Energia):** A IA deve ser desenvolvida visando o uso eficiente de energia e recursos, priorizando materiais renováveis e minimizando desperdícios e o uso de materiais escassos, com uma avaliação constante do seu impacto ambiental ao longo de todo seu ciclo de vida [31].

10. Dignidade

A IA deve respeitar e preservar a dignidade humana, reconhecendo o valor inerente dos indivíduos e assegurando que seus direitos sejam sempre respeitados. Os criadores e usuários de IA devem assegurar que a tecnologia preserve a integridade física e mental, identidade e necessidades das pessoas que a utilizam. Além disso, deve ser deixado explícito que a interação está sendo realizada com uma máquina, não com um humano. Este é o único princípio cuja questão ética relacionada tem o mesmo nome que o próprio princípio [31].

11. Solidariedade

A IA deve fomentar relações sociais sólidas, evitando a propagação de notícias falsas e invasões de privacidade. Seu desenvolvimento e uso devem potencializar interações sociais e fortalecer laços intergeracionais, enquanto preserva e promove a solidariedade, protegendo estruturas sociais e identidades culturais, e responsabilizando organizações por impactos negativos [31].

- **Segurança Social:** A IA deve preservar valores democráticos, garantindo informações precisas e imparciais aos cidadãos e evitando interferência em decisões políticas, a partir de um desenvolvimento que priorize esses princípios [31].
- **Coesão:** As organizações de IA devem assegurar uma distribuição justa dos seus benefícios e contribuir para a justiça global e coesão social, evitando prejudicar sistemas democráticos. Ademais, é essencial desenvolver estratégias proativas para impulsionar a coesão e colaboração, compartilhando conhecimentos com entidades acadêmicas, sociais e industriais [31].

2.1.2 Engenharia de Requisitos e Requisitos Éticos para IA

A Engenharia de Requisitos (ER) é uma área da Engenharia de Software (ES) que se concentra na obtenção e análise dos requisitos de um sistema em desenvolvimento [65], incluindo sistemas baseados em IA. Na ER, os requisitos desempenham um papel crucial como o canal de comunicação entre desenvolvedores e partes interessadas, capturando e documentando as necessidades de um projeto desde o início para garantir uma compreensão completa e precisa [66].

Para Sculley et al. [67], é desafiador estabelecer limites claros em sistemas de IA ao definir um comportamento específico desejado, como acontece na prática da ES convencional. Isso ocorre porque a IA é usada precisamente quando esse comportamento não pode ser expresso de forma eficaz apenas com lógica de software, necessitando de dados externos, o que traz complexidades adicionais. Como consequência, a taxonomia Requirements Engineering for Artificial Intelligence (RE4AI) foi criada para orientar o desenvolvimento de IA, alinhando elementos-chave da IA com atividades de ER. Seu propósito é ajudar praticantes e pesquisadores a focarem em desafios específicos desde o início e a adaptarem suas abordagens de pesquisa e desenvolvimento da IA [68].

A ES tradicional se concentra principalmente na codificação após coletar requisitos, realizar análises e criar projetos detalhados. No entanto, a ES que incorpora IA inclui etapas adicionais, como a coleta de dados, a escolha de algoritmos adequados, como de PLN, e o treinamento do modelo para alcançar resultados desejados. Nessa abordagem, há menos ênfase na escrita de código-fonte [69].

Requisitos Éticos (RE) para sistemas de IA são requisitos derivados de princípios ou códigos éticos, semelhantes aos requisitos legais que resultam de leis e regulamentos. Definir RE desde o início no desenvolvimento do projeto, como por exemplo de IA, permite considerar questões éticas ao longo do processo [65]. Para Kuwajima et al. [70] e Pekka et al. [71], a qualidade dos sistemas baseados em IA é impactada pela especificação de requisitos. Paech e Schneider [72] defendem que RE precisam ter uma relação direta com

os requisitos funcionais e de qualidade do software, pois o valor que um usuário atribui ao software será influenciado pelo seu entendimento e apreciação desses aspectos.

Segundo Guizzardi et al. [73], os RE podem ser interpretados por meio de valor e risco. O valor é determinado pela relação entre as características de um objeto e os objetivos da entidade ou indivíduo que atribui valor a ele. Já o risco envolve a possibilidade de um objeto ser prejudicial, considerando potenciais ameaças e fatores que possibilitam esse risco. É fundamental entender que tanto as noções de valor quanto de risco são moldadas pelo contexto em que se encontram. Pode-se tomar como exemplo os princípios da Beneficência, que se refere à ideia de proporcionar valor e promover o bem, e da Não Maleficência, que se concentra em prevenir danos. Em essência, enquanto a Beneficência visa promover impactos benéficos, a Não Maleficência se esforça para evitar consequências adversas. Tais impactos podem ser categorizados como "Eventos de Ganho", para os resultados positivos, e "Eventos de Perda", para os negativos.

Agbese et al. [74, 75] enfatizam a importância de integrar valores éticos em todos os níveis organizacionais, particularmente no contexto de projetos de IA. Com esse intuito, propuseram um *framework* ético estruturado como uma pilha multinível e interligada, projetada para a implementação de princípios éticos de forma integrada na hierarquia organizacional. Este modelo está em harmonia com abordagens ágeis de gerenciamento de portfólio e é organizado em camadas: estratégica, onde são definidos os requisitos éticos principais; de épicos, que traduzem e adaptam esses requisitos para projetos maiores; operacional, que transforma a estratégia em ações práticas; e finalmente, a camada de histórias, que individualiza os requisitos para atividades específicas de equipes ou indivíduos.

2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é um campo interdisciplinar da ciência da computação e da IA, focado em criar técnicas e sistemas que habilitem máquinas a compreender, interpretar e interagir com a linguagem humana. Integrando princípios de computação, linguística e matemática, sua missão é transformar a linguagem natural em comandos que os computadores possam processar. O PLN possui duas vertentes: Compreensão da Linguagem Natural (NLU), do inglês, *Natural Language Understanding* e Geração de Linguagem Natural (NLG), do inglês, *Natural Language Generation* [76].

NLU se concentra na compreensão computacional e interpretação da linguagem humana, capacitando máquinas a interpretar a linguagem natural, reconhecendo entidades, conceitos, emoções, palavras-chave, entre outros elementos [77]. Esta tecnologia desempenha um papel crucial em várias aplicações, como Análise de Sentimentos [78] para

determinar se textos expressam sentimentos positivos, negativos ou neutros; Extração de Entidade Nomeada para identificar pessoas, locais, organizações e outras entidades em textos [79]; Reconhecimento de Intenções em sistemas conversacionais, identificando o objetivo ou ação desejada do usuário com base em sua entrada [80].

A NLG permite que máquinas gerem texto ou fala de maneira autônoma, que se assemelhe ao estilo natural humano. Na prática, a NLG propõe métodos para criar textos em contextos do mundo real [81]. Este campo engloba várias aplicações, incluindo: tradução automática entre idiomas [82]; geração de respostas contextualizadas para perguntas [83]; redução da complexidade linguística de um texto para torná-lo mais compreensível [84]; resumo de informações para leituras mais diretas [85]; e formulação de respostas coerentes para agentes de diálogo como *chatbots*, onde a máquina precisa interagir com os usuários de forma natural [86].

2.2.1 Pré-Processamento de Textos

O pré-processamento de texto é a etapa inicial em muitas tarefas de PLN, que envolve técnicas para preparar e limpar dados textuais destinados à análise ou modelagem. Essa fase é crucial, e não pode ser ignorada, pois potencializa a qualidade do conjuntos de dados, especialmente para classificação de textos. Métodos como remoção de palavras desnecessárias (*stopwords*) e pontuações; conversão de letras para minúsculas; remoção de *tags HTML*; simplificação de *emoticons* e substituição de links por caracteres curingas podem ser benéficos para esta tarefa de PLN [87].

2.2.2 AM e Classificação de Textos Automatizada

Aprendizado de máquina (AM) é um subconjunto da IA que se concentra no desenvolvimento de algoritmos que permitem aos computadores aprender e fazer previsões com base em dados. Dependendo da abordagem utilizada, o aprendizado pode ser supervisionado, onde o modelo é treinado com textos já rotulados para aprender a associar características do texto às categorias desejadas [88, 89]. Aprendizado não supervisionado, que lida com dados não rotulados, buscando padrões e agrupamentos sem conhecimento prévio dos resultados. Aprendizado semi-supervisionado, onde combina aprendizado supervisionado e aprendizado não supervisionado, utilizando poucos dados rotulados e muitos não rotulados para melhorar a performance do modelo. Aprendizado por reforço, que baseia-se em agentes que interagem com um ambiente e aprendem por recompensa e penalização, sendo aplicado em jogos, robótica e otimização de decisões [90].

Na classificação de textos, o aprendizado supervisionado é amplamente utilizado, pois permite que modelos sejam treinados para categorizar automaticamente novos textos com

base em padrões aprendidos a partir de um conjunto rotulado. Essa classificação pode ser realizada de diferentes formas, dependendo do número de categorias atribuídas a cada texto. As mais comuns são: Classificação Binária, Multiclasse e Multi-Rótulo. A Classificação Binária envolve categorizar textos em uma de duas classes distintas, com o classificador decidindo em qual categoria o texto se enquadra com base em características específicas [91]. Já a Classificação Multiclasse designa cada instância a uma única categoria dentro de um conjunto que possui mais de duas opções [92]. Em contraste, a Classificação Multi-Rótulo permite que um texto seja vinculado a múltiplos rótulos de um conjunto disponível [93].

Entre as técnicas mais avançadas de classificação de texto, destacam-se os modelos baseados em *Transformers*, uma arquitetura de rede neural que utiliza mecanismos de atenção. Essa abordagem elimina a necessidade de camadas recorrentes ou convolucionais, permitindo o processamento de sequências de dados de forma mais eficiente. O mecanismo de atenção dos *Transformers* foca em palavras relevantes independentemente de sua posição na frase, proporcionando melhor compreensão e geração de linguagem. A arquitetura *Transformer* revolucionou o campo da PLN, definindo novos padrões em tarefas como classificação de texto, tradução automática e geração de linguagem [94].

Um dos modelos mais influentes com base na arquitetura Transformer é o BERT *(Bidirectional Encoder Representations from Transformers)*, introduzido por Devlin et al. [5]. O BERT aproveita o poder do *Transformer* para pré-treinar representações bidirecionais profundas, permitindo que o modelo entenda o contexto completo de uma palavra analisando simultaneamente as palavras a esquerda e direita em uma frase. O pré-treinamento BERT é realizado por meio de duas tarefas principais:

- **Modelagem de Linguagem Mascarada (MLM):** Nesta tarefa, o modelo mascara aleatoriamente 15% das palavras em uma frase e é treinado para prever essas palavras com base no contexto bidirecional. Essa abordagem permite uma compreensão mais precisa do significado das palavras em diferentes contextos. Para evitar dependência excessiva do token de máscara ([MASK]), palavras mascaradas são substituídas por outras palavras ou deixadas inalteradas em uma pequena porcentagem de casos durante o treinamento [5].

- **Predição da próxima frase (NSP):** Esta tarefa visa permitir que o modelo entenda a relação entre frases consecutivas, o que é essencial para aplicações como resposta a perguntas e inferência de linguagem natural. Durante o treinamento, o modelo recebe pares de frases, 50% das quais são sequências reais e 50% frases desconectadas, e aprende a identificar se as frases são sequenciais ou não [5].

Após o pré-treinamento, o BERT passa por um processo de ajuste fino para tarefas específicas, como classificação de texto, análise de sentimentos ou sistemas de resposta a perguntas. Nesta fase, o modelo é treinado adaptando-o a diferentes aplicações [5].

XLM-RoBERTa [47] é um modelo de linguagem mascarada multilíngue projetado para lidar com texto em uma ampla gama de idiomas. Ele estende os recursos do BERT pré-treinando 100 idiomas e alavancando um corpus CommonCrawl, de 2 TB+. Essa abordagem permite o aprendizado de representações multilíngues de forma auto-supervisionada, permitindo desempenho superior em tarefas multilíngues, especialmente em idiomas de poucos recursos[95].

DistilBERT [48] é uma versão otimizada e compacta do BERT, projetada para manter a maior parte de seu desempenho enquanto reduz significativamente o tamanho do modelo e melhora sua eficiência computacional. Ele usa um processo chamado destilação de conhecimento, no qual o modelo reduzido (modelo do aluno) aprende a imitar o comportamento do modelo completo (modelo do professor). Essa técnica transfere conhecimento do modelo maior para o menor, preservando sua capacidade de generalização, mas com menor custo computacional. O DistilBERT é especialmente útil em aplicações práticas que exigem alta eficiência, como dispositivos móveis e sistemas de tempo real.

Modelos baseados em transformadores, como BERT, XLM-RoBERTa e DistilBERT revolucionaram o campo de PLN, oferecendo soluções relevantes para uma ampla variedade de tarefas.

2.2.3 Métricas de Avaliação de Desempenho

Para avaliar modelos de classificação, métricas como Precisão, *Recall*, F1 Score, Weighted AVG e Macro AVG são amplamente utilizadas em problemas binários e multirótulos [96, 97]. Essas métricas, derivadas da matriz de confusão, fornecem uma análise detalhada do desempenho do modelo em termos de previsões corretas e incorretas [2].

Matriz de Confusão

A Matriz de Confusão compara as classificações previstas do modelo com os valores reais. A Tabela 2.1, adaptada do trabalho de Grandini et al. [2], ilustra essa abordagem. Os elementos da matriz incluem verdadeiros positivos (TP), falsos negativos (FN), falsos positivos (FP) e verdadeiros negativos (TN), que representam as instâncias classificadas corretamente, instâncias negativas classificadas incorretamente, instâncias positivas classificadas incorretamente e instâncias negativas classificadas corretamente, respectivamente.

No contexto multirrótulo, o cálculo de métricas pode ser realizado de forma semelhante. Cada classe é tratada de forma independente, e uma matriz de confusão é gerada

para cada classe. As métricas podem ser calculadas individualmente para cada rótulo e, em seguida, agregadas usando médias (macro ou ponderadas (Weighted)), dependendo do objetivo da análise. Isso nos permite avaliar o desempenho geral do modelo e sua eficácia na identificação de rótulos específicos em cenários com vários rótulos por instância [97]. A tabela 2.2 demonstra essa abordagem.

Tabela 2.1: Matriz de confusão adaptada de Grandini et al. [2]

		Predicted Class	
		Class 1	Class 0
Actual Class	Class 1	TP	FN
	Class 0	FP	TN

Tabela 2.2: Matriz de confusão para classificação multi-rótulo

		Predicted Label for Class k	
		Label 1 (present)	Label 0 (absent)
Actual Label for Class k	Label 1 (present)	TP_k	FN_k
	Label 0 (absent)	FP_k	TN_k

Precisão

É calculada pela proporção de Verdadeiros Positivos *(TP)* em relação ao total de elementos classificados como positivos pelo modelo. Esta métrica mostra o grau de confiabilidade do modelo ao identificar um elemento como Positivo [2].

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2.1}$$

Recall

Representa a proporção de Verdadeiros Positivos *(TP)* em relação ao conjunto de instâncias que são de fato positivas. Essa métrica avalia a habilidade do modelo em detectar corretamente todas as ocorrências da classe positiva no conjunto de dados [2] .

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

F1 Score

É a média harmônica entre Precisão e *Recall*, oferecendo um equilíbrio entre ambas as métricas. Ela proporciona uma avaliação unificada do desempenho do modelo, sendo particularmente útil em cenários onde as classes estão desequilibradas [2].

$$F1\ Score = \frac{2 \times Precisão \times Recall}{Precisão + Recall} \tag{2.3}$$

Weighted Average (AVG)

A média ponderada das métricas de classificação (como Precisão, Recall e F1 Score) atribui pesos diferentes aos resultados de cada classe. Esta métrica é especialmente útil em conjuntos de dados desbalanceados, pois ao ponderar as classes de acordo com sua importância relativa, reflete mais precisamente cada elemento no cálculo final [45].

$$Weighted\ Avg = \frac{\sum_{i=1}^{N}(Measure \times weight)_i}{Total\ number\ of\ sample} \tag{2.4}$$

Macro Average (AVG)

Calcula métricas de avaliação como Precisão, Recall e F1 Score, computando-as individualmente para cada classe e, em seguida, calculando a média entre elas. Neste método, todas as classes recebem o mesmo peso na média, independentemente de sua frequência ou tamanho no conjunto de dados [2, 46].

$$Macro\ AVG = \frac{\sum_{k=1}^{K} \text{Measure}_k}{K} \tag{2.5}$$

2.3 Trabalhos Correlatos

As diferenças entre as práticas de engenharia de requisitos para softwares tradicionais, que são em grande parte determinísticos e para softwares baseados em IA, caracterizados frequentemente por sua natureza imprevisível e operações de "caixa-preta", demandam uma revisão e adaptação das ferramentas e métodos de ER existentes [68].

Vakkuri et al. [40] apresentaram o ECCOLA, um método interativo e flexível projetado para integrar-se às práticas ágeis no desenvolvimento de sistemas de IA e autônomos. O objetivo do ECCOLA é promover a reflexão ética ao longo de todo o processo de design e desenvolvimento do sistema. Compatível com abordagens convencionais, o ECCOLA é materializado através de um baralho de 21 cartas, agrupadas em 8 categorias distintas que contemplam diferentes aspectos da ética em IA. Este método possui três etapas que são repetidas em cada iteração: *Preparação*, onde cartas apropriadas são escolhidas para a *sprint* corrente; *Revisão*, as cartas são consultadas frequentemente para registrar atividades e diálogos éticos; *Avaliação*, verifica-se a execução integral das ações planejadas, garantindo a continuidade ética do processo. Essa abordagem é projetada para facilitar a incorporação prática da ética em IA para desenvolvedores e organizações. Contudo, ela não oferece diretrizes específicas para a fase de engenharia de requisitos, a qual poderia maximizar o método, fazendo com que as necessidades e restrições éticas fossem identificadas e formalizadas desde as etapas iniciais do desenvolvimento dos sistemas.

Cerqueira et al. [42] desenvolveram o *RE4AI Ethical Guide*, um guia baseado no ECCOLA [40], que desempenha o papel de auxiliar na eliciação de requisitos éticos em sistemas que utilizam IA. O objetivo principal é permitir que os usuários elicitem requisitos e os integrem em seus *Sprint Backlogs* na forma de histórias de usuário, abrangendo requisitos funcionais, não funcionais e éticos. Isso auxilia na criação de histórias de usuário éticas para o *backlog* de produtos em um contexto de desenvolvimento ágil. Em essência, os requisitos do sistema correspondem às histórias de usuário no *backlog*, que são trabalhadas em iterações de 1 a 4 semanas, chamadas *sprints*. O guia é caracterizado por sua flexibilidade, abordagem iterativa e orientação, além de oferecer documentação para facilitar a familiarização do usuário com o sistema. Ele promove a participação de diversas partes interessadas, incluindo os próprios usuários na elicitação de requisitos éticos para sistemas de IA. No trabalho em questão, foi possível avaliar apenas a utilidade e usabilidade do *framework* proposto, não sendo possível examinar se os requisitos elicitados atendiam ao que foi proposto no próprio guia.

No estudo conduzido por Han et al. [98], foi apresentada a técnica que utiliza a IA explicável para aprimorar os modelos de classificação de requisitos. Inicialmente, foi realizado treinamento do modelo para identificar quais palavras eram decisivas para a classificação dos textos e, posteriormente, o examinaram com uma estrutura explicativa para compreender seu funcionamento interno. Isso permitiu que eles identificassem formas de realizar melhorias no modelo. Em seguida, depuraram os dados originais, removendo informações irrelevantes para garantir um treinamento de alta qualidade. Por fim, refinaram o modelo original com dados mais consistentes, resultando em maior precisão. Para avaliar a eficácia da técnica, realizaram experimentos utilizando o modelo BERT

na classificação dos requisitos. Os resultados obtidos destacaram uma melhoria notável no desempenho do modelo. No entanto, é importante salientar que o trabalho em questão não abordou a inclusão de requisitos éticos em sistemas de IA, apenas requisitos de *softwares* tradicionais.

Com um foco semelhante ao estudo anterior, Jain et al. [99] desenvolveram uma abordagem automatizada voltada para a detecção de requisitos de segurança e privacidade em contratos de ES. Para validar esta abordagem, conduziram experimentos utilizando o modelo *Transformer T5*[5] *(Text-to-Text Transfer Transformer)* para a identificação automatizada desses requisitos. Os resultados foram significativos, alcançando uma métrica *F1 Score* de 91%. No entanto, é importante notar que, embora esse trabalho tenha obtido resultados notáveis, ele é bastante restrito em relação ao escopo do problema a ser resolvido e ao tipo de aplicação de software abordada.

Rossini et al. [100] propuseram um *framework* para a integração de Princípios Éticos por Design, com ênfase na aprendizagem de comportamento ético por meio de redes neurais profundas. Esta abordagem visa criar máquinas que façam escolhas assertivas, levando em conta critérios éticos. A ideia fundamental é usar a ética como um tipo de raciocínio automático, com base em descrições formais do sistema de IA, como ontologias, e aplicá-la desde o início do processo de aprendizado. Ao contrário de abordagens que geram regras éticas com base nas ações da IA, esta abordagem parte de princípios éticos claramente definidos e se concentra em aprender ética a partir de fontes de conhecimento ético. A rede neural profunda proposta modela tanto as condições operacionais quanto as questões éticas relacionadas à tomada de decisões. O trabalho de Rossini et al. embora distinto em sua natureza da proposição da atual pesquisa, pode servir de complemento, pois juntas, podem englobar todo o ciclo de vida de um sistema baseado IA.

Canedo e Mendes [1] compararam diferentes técnicas de extração de recursos de texto, como Bag of Words (BoW), TF-IDF e CHI^2, juntamente com algoritmos de aprendizado de máquina (Regressão Logística (LR), Máquinas de Vetores de Suporte (SVM), Multinomial Naive Bayes (MNB) e k-Nearest Neighbors (kNN)) para classificar RFs e RNFs usando o conjunto de dados PROMISE_exp. Para abordar o desequilíbrio do conjunto de dados, os autores aplicaram validação cruzada de 10 folds, onde a combinação de TF-IDF e Regressão Logística (LR) produziu o melhor desempenho, alcançando uma F-measure de 91% na classificação binária.

Seguindo uma abordagem semi-supervisionada, Casamayor et al. [24] propuseram um método para identificar e classificar RNFs, usando um pequeno conjunto de requisitos pré-categorizados para treinar um classificador inicial. Por meio de um processo iterativo, o classificador identificou novos RNFs. Integrado a um sistema de recomendação, o método

[5] `https://huggingface.co/docs/transformers/model_doc/t5`

dos autores oferece suporte a analistas de requisitos e designers de software em design arquitetônico. A abordagem proposta atingiu uma precisão de mais de 70%, superando os métodos supervisionados tradicionais e exigindo menos esforço humano para rotulagem. Além disso, o sistema tem potencial para melhoria contínua por meio do feedback de analistas de requisitos.

Como parte de uma avaliação de diferentes algoritmos de AM, Yucalar [25] testou vinte algoritmos, incluindo Naive Bayes, Rotation Forests, Convolutional Neural Networks e BERT, em um conjunto de dados em idioma turco. BERTurk se destacou, alcançando uma F-measure de 95%, destacando que modelos baseados em Transformer, como BERT, DistilBERT e RoBERTa, são significativamente superiores na distinção entre RFs e RNFs. Isso demonstra a importância de modelos de linguagem avançados na classificação de requisitos.

Cleland-Huang [101] teve como objetivo automatizar a detecção e classificação de RNFs, como requisitos de segurança, desempenho e usabilidade. O trabalho foi dividido em três fases principais: treinamento, classificação e aplicação. Na fase de treinamento, palavras-chave foram extraídas de requisitos rotulados manualmente e pesos probabilísticos atribuídos indicando sua relevância para tipos específicos de RNFs. Na fase de classificação, esses termos ponderados foram usados para determinar a probabilidade de um novo requisito pertencer a um tipo específico de RNFs. Na fase de aplicação, os requisitos classificados deram suporte a atividades de engenharia de software, como requisitos de negócios e design de arquitetura de software. No geral, o classificador atingiu taxas de recall entre 60% e 90% para diferentes classes. No entanto, os autores observaram que melhorias adicionais eram necessárias para aperfeiçoamento das métricas de desempenho.

Por fim, Hey et al. [29] introduziram o NoRBERT, uma adaptação do modelo BERT, usando o conjunto de dados PROMISE RNFs. O NoRBERT não apenas superou as abordagens anteriores, alcançando uma pontuação F1 Score de até 94% na classificação de RFs e RNFs, mas também demonstrou desempenho significativo em cenários não vistos, com uma melhoria de 15 pontos percentuais. Isso destaca a relevância dos Transformers para a classificação de requisitos em diversos projetos.

Embora grande parte das abordagens mencionadas tenham sido eficazes na classificação de RFs e RNFs, elas não foram projetadas para lidar com a complexidade e as nuances dos requisitos éticos para sistemas de IA. Contudo, os estudos que exploram princípios e requisitos éticos, bem como ferramentas para a sua aplicação, como Vakkuri et al. [40], Cerqueira et al. [42] e Rossini et al. [100] fornecem um arcabouço fundamental para a compreensão da dimensão ética na IA, complementando a presente pesquisa. Este trabalho visa abordar a lacuna na classificação de requisitos éticos introduzindo um modelo projetado especificamente para identificar tais requisitos em projetos de IA. Aproveitando

técnicas avançadas como Transformers, o modelo proposto visa capturar a complexidade e o contexto ético necessário para uma análise precisa. Espera-se que esta abordagem não apenas preencha uma lacuna existente, mas também estabeleça um novo padrão para a classificação de requisitos éticos em IA.

A Tabela 2.3 exibe a comparação dos artigos correlatos e a proposição desta pesquisa, o modelo *Ethical Requirements Classifier for AI (ERC4AI)*. A comparação é fundamentada em características pertinentes a este estudo, destacadas como: "Practical", que indica se o estudo apresenta uma aplicação prática do modelo, como experimentos, implementações ou testes empíricos utilizando técnicas de AM; "NLP", que refere-se à utilização de técnicas de PLN, incluindo a classificação de texto; e "Ethical Requirements for AI", que avalia se o estudo aborda explicitamente requisitos éticos em IA, seja na modelagem de requisitos ou desenvolvimento de sistemas alinhados a princípios éticos.

Tabela 2.3: Comparação entre os Trabalhos Correlatos vs. ERC4AI

Research	**Characteristics**		
	Practical	NLP	Ethical Requirements for AI
Vakkuri et al.[40]			✓
Cerqueira et al.[42]			✓
Han et al.[98]	✓	✓	
Jain et al.[99]	✓	✓	
Rossini et al.[100]	✓		✓
Canedo e Mendes[42]	✓	✓	
Casamayor et al.[24]	✓	✓	
Yucalar [25]	✓	✓	
Cleland-Huang [101]	✓	✓	
Hey et al. [29]	✓	✓	
ERC4AI	✓	✓	✓

2.4 Síntese do Capítulo

Este capítulo abordou a ética em IA, explorando 11 princípios éticos centrais. Cada princípio foi desmembrado para oferecer uma visão geral das questões éticas que surgem ao integrar a IA em diversas aplicações. A ligação entre a Engenharia de Requisitos tradicional e os Requisitos Éticos para IA foi apresentada, destacando sua importância desde a concepção de sistemas de IA. Além disso, o capítulo adentrou no domínio de PLN,

explanando os fundamentos e técnicas como Pré-Processamento de Textos, que prepara e transforma dados para análises subsequentes, e a Classificação de Textos Automatizada, para compreender e categorizar textos, além disso, foram apresentadas as Métricas de Avaliação de Desempenho, importantes para medir a eficácia de modelos em tarefas de PLN. Para concluir, o capítulo posicionou esta proposta de trabalho em relação a outros estudos na área, estabelecendo sua singularidade e relevância no panorama atual da ética em IA e PLN.

Capítulo 3

Classificador de Textos ERC4AI

Este capítulo apresenta o processo de desenvolvimento do *Ethical Requirements Classifier for AI (ERC4AI)*, desde a etapa inicial, que é a Revisão Bibliográfica, passando pela Obtenção, Análise e Preparação dos Dados, e por fim, a etapa de modelagem, utilizando algorítimos AM.

3.1 Fase 1 - Revisão Bibliográfica

Esta fase da pesquisa teve como principal objetivo a compreensão dos requisitos éticos em IA e a exploração de técnicas para a sua classificação, com base nos princípios éticos considerados por Ryan e Stahl [31]. Portanto, foi realizada uma revisão bibliográfica, analisando publicações acadêmicas para compreender as questões centrais relacionadas à esta temática, bem como as abordagens existentes que são utilizadas. Esse esforço permitiu identificar as práticas correntes e as principais lacunas na classificação automática de requisitos éticos em IA. Nas Seções 2.1.1 e 2.3 é possível verificar o detalhamento desta fase. Essa investigação revelou uma lacuna significativa: a ausência de propostas para automatizar a classificação de requisitos éticos em IA.

3.2 Fase 2 - Obtenção dos Dados

A obtenção dos dados no contexto de desenvolvimento de soluções éticas em IA assume uma complexidade adicional se comparada aos softwares convencionais, pois os dados são menos acessíveis. A maioria dos conjuntos de dados disponíveis publicamente tendem a se concentrar em Requisitos Funcionais (RFs) [102] e Requisitos Não Funcionais (RNFs) [102]. No entanto, conforme identificado por Carleton et al. [43], existe uma interseção significativa entre os sistemas convencionais e baseados em IA. Isso decorre do fato de que sistemas de IA são em sua essência sistemas de software, e consequentemente aderem aos

mesmos princípios fundamentais de design e implantação. Esta sobreposição sugere que as práticas bem estabelecidas e os requisitos de qualidade em softwares tradicionais podem ser relevantes e aplicáveis aos sistemas de IA. Nesta pesquisa foi considerada a base de dados *PROMISE* [1], que possui requisitos RFs e RNFs. A hipótese era que, ao analisar os requisitos de software convencionais, seria possível obter requisitos que também poderiam ser aplicáveis em sistemas baseados em IA.

Uma revisão da literatura possibilitou a descoberta da base de dados *Passager Flow* [44], uma fonte importante de requisitos éticos em IA, oriunda de um estudo acadêmico aplicado em um ambiente industrial. Esta base utiliza a metodologia de Design Alinhado à Ética [103] e Histórias de Usuários Éticas [104]. Seu principal enfoque é a indústria marítima, com ênfase no gerenciamento do fluxo de passageiros. A *Passager Flow* [44] inclui uma série de requisitos éticos de IA que são categorizados de acordo com um dos princípios éticos: Non-Maleficence, Privacy, Responsibility e Transparency.

Esta etapa da pesquisa não apenas enfatizou a necessidade de diligência na seleção de dados apropriados, mas também realçou a compreensão de que o campo da ética em IA ainda está em evolução, com várias áreas ainda por explorar. As bases de dados reunidas, abrangendo tanto os requisitos de softwares tradicionais quanto os de aspectos éticos específicos da IA forneceram insumos para a próxima fase do estudo, possibilitando uma análise mais aprofundada e embasada dos requisitos éticos em IA. A Tabela 3.1 apresenta uma análise quantitativa de requisitos por base de dados.

Tabela 3.1: Quantidade Total de Requisitos por Base de Dados

Dataset	**Total Quantity**
PROMISE [1]	970
Passager Flow [44]	122

3.3 Fase 3 - Análise dos Dados

Nesta etapa da pesquisa, a ênfase foi a análise e rotulação dos dados, procedimentos importantes para o desenvolvimento do modelo *Ethical Requirements Classifier for AI (ERC4AI)*. A rotulação consiste na classificação manual de um conjunto de textos destinados ao treinamento do modelo [105].

3.3.1 Rotulação dos textos

Para o processo de rotulação dos textos, inicialmente realizou-se uma análise do conteúdo da base de dados *Passager Flow* [44], com o intuito de compreender os textos ali presentes. Essa análise revelou que, embora cada requisito ético em IA fosse associado a um princípio ético específico, alguns destes também refletiam aspectos de outros princípios éticos em IA. Esta constatação surgiu ao avaliar os requisitos juntamente com princípios e questões éticas abordadas na Seção 2.1.1. Sendo assim, optou-se por reavaliar cada requisito desta base de dados, reclassificando-os, quando necessário, com novas categorias de princípios éticos em IA, baseadas no trabalho de Ryan e Stahl [31]. A Tabela 3.2 apresenta um exemplo de cada princípio ético encontrado na base *Passager Flow* [44], que abrange os princípios Non-Maleficence, Privacy, Responsibility e Transparency, onde cada requisito possui apenas um princípio ético associado a ele.

Tabela 3.2: Princípios Éticos Definidos Originalmente na Base *Passenger Flow*

Ethical Principle	**Ethical Requirement in AI**
Responsibility	"Audit requirements have been written out in system description. Requirements have been discussed among all relevant parties related to systems development."
Non-maleficence	"The system is clear and intuitive to use, and leaves as little as possible for interpretation. It will not offer possibilities for passengers to do a wrong, dangerous or forbidden thing."
Privacy	"The system follows regulations on data privacy and ensures that data access is granted only to relevant personnel."
Transparency	"Ensure the quality of the data by testing. Testing at least with race, gender, age should be done and variance between groups should not be significant. Realtime testing in real context needs to be done before launch."

A Tabela 3.3 complementa a Tabela 3.2, expandindo os princípios éticos em IA associados aos requisitos previamente listados. Esses acréscimos foram realizados pela pesquisadora deste trabalho, a partir de um processo de reavaliação, onde foi verificada a presença de padrões correspondentes a outros requisitos éticos em IA.

Tabela 3.3: Inclusão de Novos Princípios Éticos aos Textos

Ethical Principle	**Ethical Requirement in AI**
Responsibility Transparency	"Audit requirements have been written out in system description. Requirements have been discussed among all relevant parties related to systems development."
Non-Maleficence Transparency	"The system is clear and intuitive to use, and leaves as little as possible for interpretation. It will not offer possibilities for passengers to do a wrong, dangerous or forbidden thing."
Privacy Trust Responsibility	"The system follows regulations on data privacy and ensures that data access is granted only to relevant personnel."
Transparency Justice and Fairness Non-Maleficence	"Ensure the quality of the data by testing. Testing at least with race, gender, age should be done and variance between gropes should not be significant. Realtime testing in real context needs to be done before launch."

A Figura 3.1 ilustra uma perspectiva quantitativa da prevalência inicial de cada princípio ético na base de dados *Passager Flow* [44], onde o princípio ético Não-Maleficência (Non-Maleficence) possui maior quantidade de registros (47), seguido em ordem decrescente por Transparência (Transparency)(30), Privacidade (Privacy)(27) e Responsabilidade (Responsibility)(18).

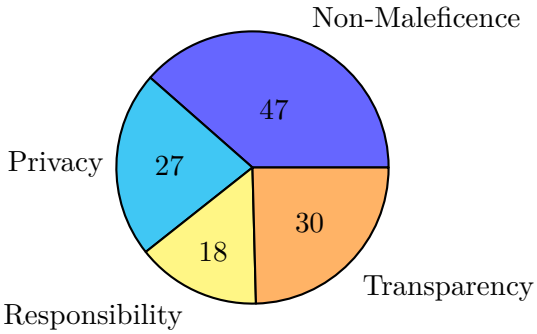

Figura 3.1: Distribuição dos Princípios Éticos na base de dados *Passager Flow*

A Tabela 3.4 mostra a nova classificação dos princípios éticos do conjunto de dados *Passager Flow*. As colunas R, NM, P e T da tabela representam os 4 princípios éticos, *Responsibility* (R), *Non-Maleficence* (NM), *Privacy* (P) e *Transparency* (T), inicialmente identificados no *Passager Flow* [44], enquanto as linhas da tabela representam os princípios éticos adicionais identificados nesta pesquisa para cada um dos 4 princípios iniciais R, NM, P e T após a reavaliação dos textos. A coluna ER representa o número total de requisitos para cada um dos princípios éticos identificados após a reclassificação. A reclassificação revelou que dos 47 requisitos éticos originalmente classificados como *Non-Maleficence*, 18 também poderiam ser classificados como *Transparency*; 5 como *Justice and Fairness*; 32 como *Responsibility*; 4 como *Privacy*; 12 como *Beneficence*; 2 como *Freedom and Autonomy*; 15 como *Trust*; 1 como *Sustainability*, 2 como *Dignity*.

A base de dados *PROMISE* [1] foi submetida a uma análise para identificar requisitos convencionais que poderiam ser considerados como éticos em IA. Este procedimento envolveu a rotulação dos requisitos em uma ou mais categorias de princípios éticos em IA, seguindo uma abordagem semelhante à aplicada anteriormente na base *Passenger Flow* [44]. A Figura 3.2 exibe a distribuição dos requisitos classificados como éticos em IA e aqueles que não se enquadram nesse padrão. É possível observar na Figura uma prevalência de requisitos que não atendem ao critério ético em IA nessa base de dados.

Tabela 3.4: Nova categorização do Passenger Flow após reclassificação

Additional Ethical Principles	**Passenger Flow**				**ER**
	R	NM	P	T	
Transparency (T)	13	18	16	30	77
Justice and Fairness (JF)	2	5	0	3	10
Non-maleficence (NM)	1	47	5	3	56
Responsibility (R)	18	32	12	11	73
Privacy (P)	7	4	27	8	46
Beneficence (B)	1	12	1	6	20
Freedom and Autonomy (FA)	0	2	4	6	12
Trust (Tr)	3	15	9	8	35
Sustainability (S)	2	1	0	0	3
Dignity (D)	0	2	0	1	3
Solidarity (So)	0	0	0	0	0

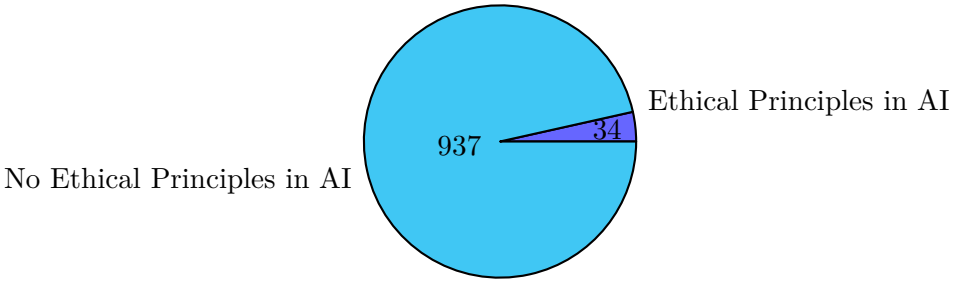

Figura 3.2: Distribuição de Requisitos Éticos e Não Éticos em IA da Base de Dados *PROMISE* [1]

A Tabela 3.5 apresenta exemplos dos requisitos identificados e as categorias que eles receberam. A figura 3.3 mostra a distribuição dos 34 requisitos éticos identificados no conjunto de dados PROMISE [1], classificados em um ou mais princípios éticos de IA, sendo *Privacy* o principio mais associado aos requisitos, e *Freedom and Autonomy* juntamente com *Dignity* os princípios que não nenhuma associação. Os 937 registros restantes foram categorizados como não pertencentes às classes de princípios éticos analisados. No entanto, esses requisitos foram mantidos no conjunto de dados para a etapa de modelagem, pois podem auxiliar o modelo na distinção entre requisitos associados a princípios éticos e aqueles sem essa característica, contribuindo assim para a representação da classe negativa e ampliação da diversidade do conjunto de treinamento.

Ao concluir as análises das bases de dados *Passager Flow* [44] e *PROMISE* [1], ambas foram unidas para formar um único conjunto de dados.

3.3.2 Extensão do Processo de Rotulação de Dados

Para aumentar a confiabilidade dos dados e reduzir o risco de viés, mais participantes foram incluídos no processo de rotulação dos requisitos. Conforme mencionado por

Tabela 3.5: Categorização do conjunto de dados PROMISE

Ethical Principles	Ethical Requirement in AI
Privacy Trust	"The system shall protect private information in accordance with the organization's information policy."
Transparency Trust	"The system shall notify customers of changes to its information policy."
Non Maleficence	"System use shall not cause any harm to human users."
Privacy	"The product shall protect the identity of the players. The product shall provide players no access to information that might reveal the identity of another player."

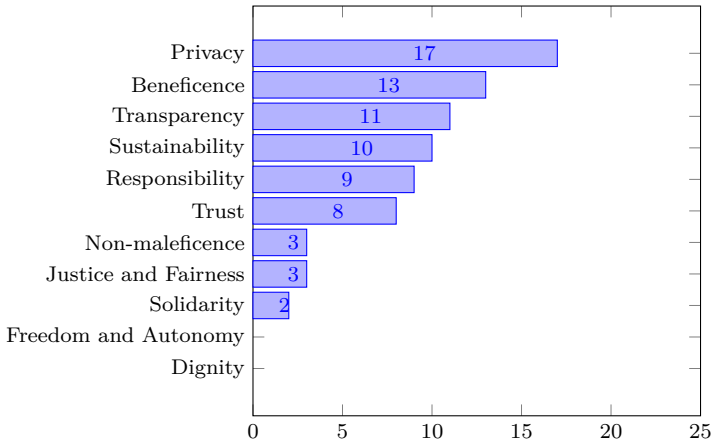

Figura 3.3: Categorização do conjunto de dados PROMISE [1]

Desmond et al. [106], à medida que o número de categorias aumenta, as decisões dos rotuladores tornam-se mais complexas. Isso faz com que a mecânica de rotulagem seja mais trabalhosa e demorada, elevando o risco de erros. Visto que, o conjunto de dados em questão possui tal característica, justificou-se incluir mais pessoas no processo de rotulação, com o propósito de obter maior precisão na identificação e classificação dos requisitos. Esta seção descreve os passos adotados para execução deste processo. A Figura 3.4 ilustra a abordagem adotada para realizar o processo de rotulação.

Passo 1: Participantes

A escolha dos novos rotuladores (Passo *Choice of Participants*, Figura 3.4) levou em consideração a expertise dos profissionais na área de IA e Ética em IA. Os participantes foram convidados devido à colaboração anterior em projetos que envolveram o uso de IA, o que lhes proporcionou uma visão prática das questões éticas associadas. A seleção de participantes com diferentes perfis teve como objetivo realizar uma avaliação diversificada dos requisitos em relação aos princípios éticos identificados na Seção 3.3.1. Para preservar a imparcialidade do processo, os rotuladores participaram da pesquisa sem receber informações sobre as classificações anteriores. A Tabela 3.6 apresenta as informações profissionais dos participantes:

- **ID:** contém a identificação dos rotuladores. Cada rotulador é designado por um código único, como P1, P2, P3 e P4. Esses códigos são utilizados para anonimizar os dados dos participantes. O identificador P0 refere-se à pesquisadora responsável pelo trabalho e é o único não anonimizado.
- **Role:** apresenta o cargo atual ocupado pelo rotulador, indicando a função ou posição profissional que cada rotulador exerce em sua organização.
- **Educational Level:** apresenta o nível educacional dos rotuladores.
- **Experience in AI:** descreve a experiência que cada rotulador possui na área de Ética em IA.
- **Practitioner Experience:** descreve a experiência profissional dos rotuladores.

Tabela 3.6: Perfil dos Rotuladores

ID	Role	Educational Level	Experience in AI	Practitioner Experience
P0	Data Scientist	Master's student	2 years as an AI ethics researcher	More than 5 years of experience in data analysis, machine learning, and AI. Worked on projects in operational efficiency, risk management, compliance, investment fund, and customer experience.
P1	Data Scientist	PhD in Applied Computing	Participated in the Responsible AI and Ethics Working Group at a Financial Institution. Contributed to the analysis and positioning regarding AI-related bills. Assisted in the creation of AI ethical guidelines.	More than 10 years of experience in data analysis, machine learning, and AI. Worked on projects in operational efficiency, risk management, and customer experience. Currently a data science and AI mentor at Banco do Brasil.
P2	Data Scientist	Specialization in Data Science and Big Data, MBA in AI and Data Science for Business	Participated in the Responsible AI and Ethics Working Group at a Financial Institution. Contributed to the analysis and positioning regarding AI-related bills. Assisted in the creation of AI ethical guidelines.	8 years working in a financial institution. 2 years in a business agency. 2 years with web development. 4 years with data science and AI.
P3	IT Advisor	Master's in Computer Science	Works/researches software engineering for chatbots, in terms of AI, since 2019	Has been working since 2019, initially with chatbot services, and later with web programming up to the present.
P4	Student	Master's student	2 years as an AI ethics researcher	4 years as a python backend and ML developer, 2 years as an AI researcher

Passo 2: Processo de Rotulação dos Requisitos

Inicialmente, foram fornecidas informações aos participantes sobre os princípios éticos que deveriam ser considerados durante a rotulagem, mapeados na Seção 3.3.1, bem como as suas definições. Foram utilizadas as definições descritas no trabalho de Ryan e Stahl [31] como base para esta atividade. Além disso, foi enfatizada a importância de realizar uma avaliação independente, livre da influência de classificações anteriores e de outros participantes. Cada rotulador recebeu uma base de dados em formato CSV *(Comma-Separated Values)*, contendo os mesmos textos de requisitos, porém sem nenhuma classificação (Passo *Data Labeling*, Figura 3.4). As bases foram distribuídas em arquivos distintos para que os rotuladores não tivessem acesso às classificações uns dos outros.

Em todas as bases de dados, havia uma coluna dedicada a cada princípio ético em IA considerado. Os rotuladores deveriam indicar se os requisitos pertenciam ou não a esses princípios éticos. Individualmente, cada rotulador analisou os requisitos e atribuiu as classificações que consideravam adequadas.

Embora muitos estudos busquem um consenso entre os anotadores, optamos por preservar as perspectivas individuais dos participantes. Essa escolha buscou refletir melhor a subjetividade na interpretação dos requisitos éticos, considerando as diversas nuances do processo de rotulação.

Passo 3: Compilação e Análise de Dados

Após a rotulagem, os conjuntos de dados dos participantes foram compilados e então analisados para verificar o consenso entre eles - Passo Compilation of Labeled Data and Data Analysis, conforme apresentado na Figura 3.4.

O coeficiente AC1 [53] de Gwet foi usado para medir o grau de concordância na classificação de requisitos éticos. Este coeficiente fornece uma medida mais estável do que métodos tradicionais como o Kappa de Cohen, especialmente em situações com categorias desequilibradas ou alta concordância esperada [53].

No contexto deste trabalho, onde as categorias têm frequências desequilibradas, o AC1 provou ser particularmente adequado. Sendo também útil para cenários em que vários rotuladores estão analisando os mesmos itens [107]. Em geral, as interpretações dos níveis de concordância AC1 [108, 109, 110, 111] são:

- $< 0,00$: Desacordo entre rotuladores maior do que o esperado por acaso, indicando uma falta de consistência na rotulagem.
- **[0,00 − 0,20]**: Baixa concordância, sugerindo que os rotuladores têm pouca consistência em sua classificação.

- **[0,21 – 0,40]**: Concordância razoável, com classificações ainda próximas de chance, mas mostrando algum grau de concordância.
- **[0,41 – 0,60]**: Concordância moderada, representando um grau razoável de concordância entre os rotuladores.
- **[0,61 – 0,80]**: Concordância substancial, com forte concordância entre os rotuladores e classificações bem alinhadas.
- **[0,81 – 1,00]**: Concordância quase perfeita, refletindo um alto grau de alinhamento e consistência nas classificações.

O cálculo foi realizado utilizando o pacote *irrCAC*, [1] da linguagem R, e resultou nas métricas apresentadas na Tabela 3.7, onde: *PA* (porcentagem de concordância observada), *PE* (porcentagem esperada de concordância ao acaso), *AC1* (estimativa do coeficiente de concordância de Gwet), *SE* (erro padrão associado ao AC1), *CI Inferior - Superior* (intervalo de confiança para o coeficiente AC1) e *P-Value* (valor p para testar a significância do coeficiente AC1). Essas métricas avaliam a consistência entre os rotuladores e os princípios éticos analisados.

Ethical Principle	PA	PE	AC1	SE	CI (Lower - Upper)	P-Value
Transparency	0.771	0.499	0.542	0.040	0.463 - 0.621	0
Solidarity	0.955	0.048	0.953	0.012	0.929 - 0.977	0
Dignity	0.940	0.067	0.935	0.014	0.907 - 0.964	0
Sustainability	0.965	0.035	0.964	0.010	0.944 - 0.984	0
Trust	0.692	0.461	0.429	0.040	0.351 - 0.508	0
Freedom & Autonomy	0.927	0.140	0.915	0.019	0.878 - 0.952	0
Beneficence	0.867	0.157	0.842	0.022	0.799 - 0.885	0
Privacy	0.747	0.403	0.577	0.043	0.492 - 0.662	0
Responsibility	0.662	0.454	0.380	0.043	0.295 - 0.465	0
Non-Maleficence	0.715	0.348	0.563	0.041	0.482 - 0.645	0
Justice & Fairness	0.882	0.204	0.852	0.023	0.806 - 0.898	0

+

Tabela 3.7: Resultados da análise de confiabilidade entre avaliadores (coeficiente AC1 de Gwet) para cada princípio ético

Os valores da Tabela 3.7 mostram as métricas de concordância para os diferentes princípios éticos avaliados. A proporção observada de concordância (*PA*) variou entre 66,2% e 96,5%, indicando diferentes níveis de alinhamento entre os rotuladores em relação aos princípios analisados.

Para o princípio **Sustentabilidade**, foi observada a maior proporção de concordância ($PA = 96{,}5\%$) e um índice AC1 de 0,964, com intervalo de confiança entre 0,944 e 0,984. Esses resultados refletem uma concordância quase perfeita entre os avaliadores. Da mesma forma, princípios como **Solidariedade** ($AC1 = 0{,}953$), **Dignidade** ($AC1 =$

[1] `https://cran.r-project.org/web/packages/irrCAC/`

0,935), **Liberdade & Autonomia** ($AC1 = 0{,}915$), **Justiça & Equidade** ($AC1 = 0{,}852$) e **Beneficência** ($AC1 = 0{,}842$) também demonstraram altos níveis de concordância, evidenciando avaliações consistentes e bem alinhadas.

Por outro lado, princípios como **Privacidade** ($AC1 = 0{,}577$), **Transparência** ($AC1 = 0{,}542$) e **Não Maleficência** ($AC1 = 0{,}563$) apresentaram valores de concordância intermediários, indicando níveis entre moderado e substancial. O princípio **Confiança** ($AC1 = 0{,}429$) apresentou concordância moderada, com seu índice AC1 próximo ao limite inferior dessa faixa. O princípio **Responsabilidade** ($AC1 = 0{,}380$) ficou na faixa de concordância razoável, indicando menor alinhamento entre os avaliadores.

De forma geral, os valores de *P-Value* foram iguais a zero para todos os princípios, sugerindo que a concordância observada foi estatisticamente significativa em todos os casos, e os intervalos de confiança indicam baixa incerteza nas estimativas dos índices AC1.

A análise dos resultados obtidos com AC1 se mostrou relevante em cenários com alta concordância e categorias desbalanceadas. A capacidade desse coeficiente de ajustar o acordo observado para levar em conta o acaso foi importante para interpretar com mais precisão os níveis de alinhamento entre os avaliadores.

O conjunto de dados contendo os requisitos éticos rotulados pelos cinco participantes (P0, P1, P2, P3, P4) foi denominado **EthicalRequirements4AI** e está disponível na plataforma Zenodo [2]. Este conjunto de dados, consistindo de 156 requisitos éticos, foi complementado por 937 requisitos adicionais do conjunto de dados PROMISE — que não foram considerados relacionados a princípios éticos em IA — totalizando 1.093 requisitos rotulados que formam a verdade fundamental para este estudo. O objetivo desta compilação é servir como um recurso para o desenvolvimento do modelo **Ethical Requirements Classifier for AI (ERC4AI)**.

[2] `https://zenodo.org/records/14750448`

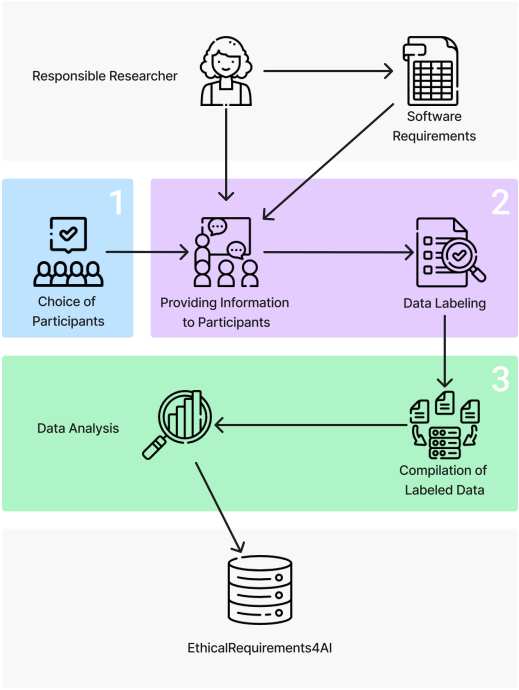

Figura 3.4: Processo de Rotulação

3.4 Fase 4 - Preparação dos Dados

A preparação dos dados textuais inclui a aplicação de técnicas de pré-processamento, com o objetivo de aprimorar o poder preditivo dos modelos e, consequentemente, na otimização de suas métricas [112]. A seleção destas técnicas varia conforme o contexto e as necessidades específicas do projeto [113]. Neste estudo, os algoritmos de pré-processamento foram desenvolvidos na linguagem de programação Python[3]. O código completo desta etapa está disponível para consulta na plataforma Zenodo.

- **Remoção de Múltiplos Espaços:** Eliminação de múltiplos espaços para obter uma formatação consistente do texto.
- **Remoção de Pontuação:** A pontuação foi removida pois, para muitos modelos de aprendizado de máquina, elementos como vírgulas, pontos e exclamações podem não agregar valor significativo, além da possibilidade de tornar o texto ruidoso [113].
- **Remoção de Números:** Números frequentemente não agregam significado relevante para a análise de texto em tarefas como classificação de texto [114], sendo assim, estes foram removidos.

[3] `https://www.python.org/`

Conforme destacado por Tabassum e Patil [113], a sequência na qual as técnicas de pré-processamento de textos são aplicadas é crucial em certos contextos. Nesta pesquisa, uma sequência específica foi cuidadosamente estabelecida para maximizar a eficácia do processamento. Inicialmente, o texto é convertido para minúsculas, padronizando a formatação. Em seguida, são removidos números, pontuação e múltiplos espaços em branco.

A Tabela 3.8 apresenta exemplos dos textos dos requisitos originais e após a preparação dos dados.

Tabela 3.8: Exemplos de Textos Antes e Após o Pré-processamento dos Requisitos

Pre-Treatment Text	**Post-Treatment Text**
"An intelligent system connected to the video surveillance in the terminal can be given parameters to search for the current or latest location of a vehicle or person in the terminal by being given the register plate, a photo of a person, or other such identifying data."	"an intelligent system connected to the video surveillance in the terminal can be given parameters to search for the current or latest location of a vehicle or person in the terminal by being given the register plate a photo of a person or other such identifying data"
"Terminal area and booking documents and booking app contains the help information (for different purposes broken down)".	"terminal area and booking documents and booking app contains the help information for different purposes broken down"
"Before a person with a vehicle enters the port, they are given explanation of what information will be handled, for example at the stage of making a reservation when they agree to give the necessary information of their vehicle."	"before a person with a vehicle enters the port they are given explanation of what information will be handled for example at the stage of making a reservation when they agree to give the necessary information of their vehicle"
"A clear advertisement is placed on the booking document about the AI and data privacy."	"a clear advertisement is placed on the booking document about the ai and data privacy"
"For leads that process longer than 25 seconds the system will record the event and duration."	"for leads that process longer than seconds the system will record the event and duration"

- **Estruturação dos Rótulos:**

  Considerando que os requisitos podem ter um ou mais rótulos, como identificado na Seção 3.3.1, foi necessário estruturar esses dados em um formato compreensível pelos algoritmos durante a fase de treinamento dos modelos. Isso envolveu a adoção de uma abordagem de classificação multi-rótulo, descrita em detalhes na Seção 2.2.2.

A Figura 3.5 apresenta a distribuição das classificações atribuídas aos diferentes princípios éticos em IA, categorizados pelo número de rotuladores que os avaliaram (1, 2, 3, 4 ou 5 rotuladores). Pode-se observar que o número de classificações varia entre os princípios, com alguns recebendo avaliação de apenas um rotulador, enquanto outros apresentam maior consenso, com classificações de dois a cinco rotuladores.

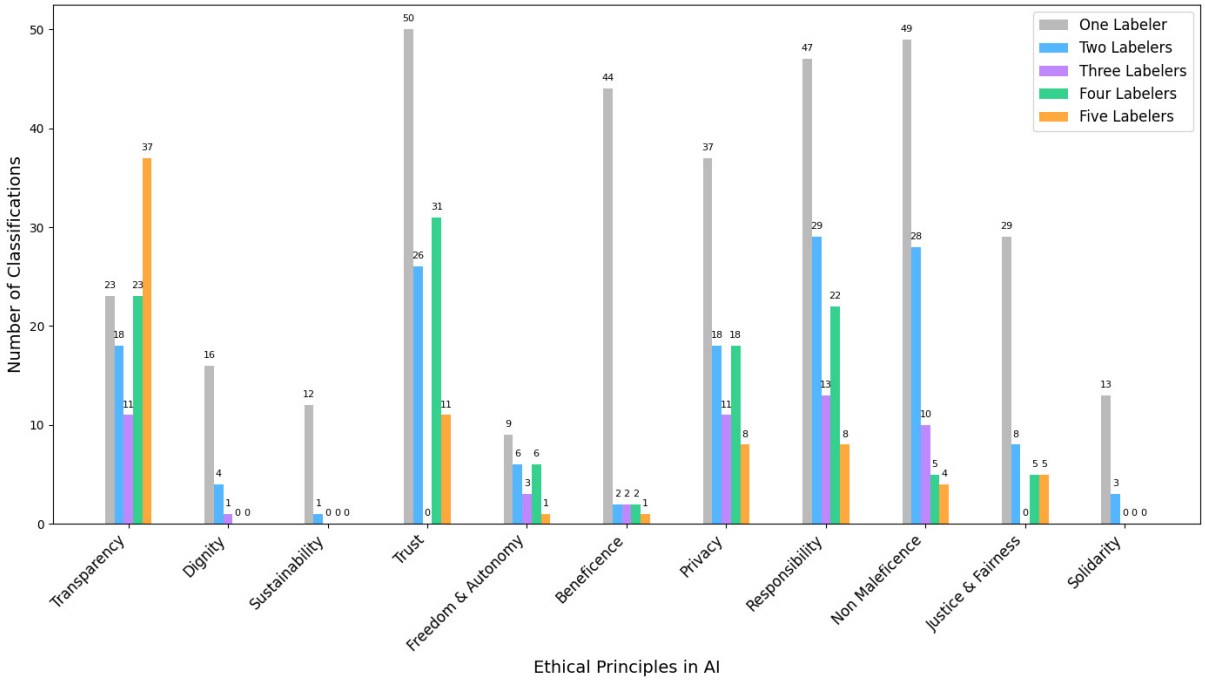

Figura 3.5: Distribuição da classificação por princípio ético e rotuladores

Princípios como *Responsibility*, *Privacy*, *Freedom & Autonomy*, *Trust*, e *Transparency* têm um número maior de classificações feitas por dois ou mais avaliadores em comparação com aquelas feitas por apenas um. Por exemplo, o princípio da *Transparency* tem 89 classificações de dois ou mais avaliadores, enquanto recebeu apenas 23 classificações de um único avaliador. Em contraste, os outros princípios seguem um padrão oposto, com a maioria das classificações sendo feitas por apenas um avaliador, indicando menor consenso entre os avaliadores nesses casos.

Para reduzir o impacto de vieses individuais, foram considerados válidos apenas os rótulos atribuídos por pelo menos dois avaliadores para cada requisito. Além disso, uma classe foi mantida como categoria independente apenas se estivesse associada a mais de 47 requisitos distintos. Princípios com um número inferior de requisitos foram agrupados na categoria Outros Princípios Éticos (*Other Ethical Principles*).

Um requisito foi incluído na categoria *Other Ethical Principles* se tivesse pelo menos um rótulo válido em uma das classes que foram agrupadas. Assim, os princípios

Dignity (5), *Sustainability* (1), *Freedom and Autonomy* (16), *Beneficence* (7), *Justice and Fairness* (18), *Non Maleficence* (47) e *Solidarity* (3) foram incorporados a categoria *Other Ethical Principles*, conforme ilustrado na Figura 3.6.

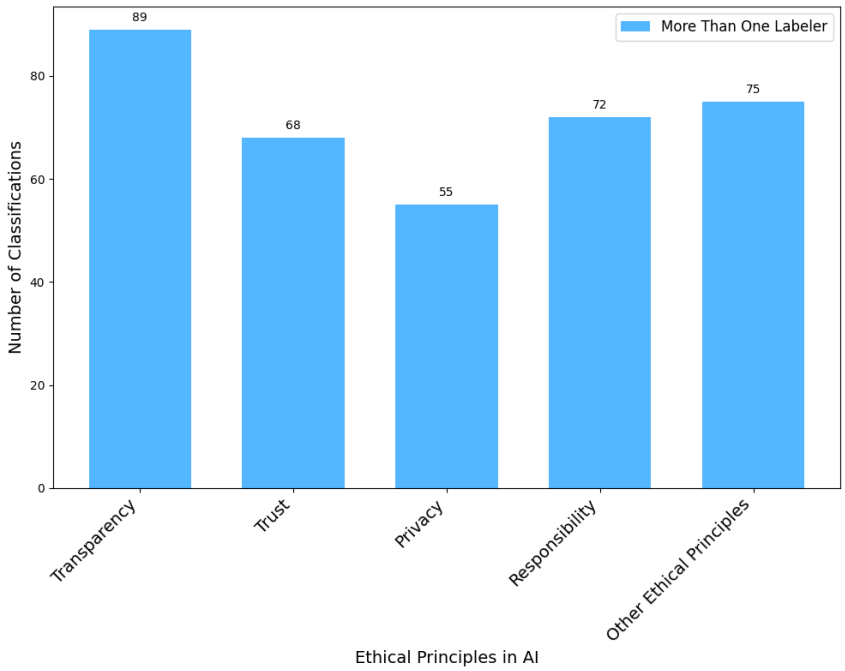

Figura 3.6: Princípios éticos da IA considerados no desenvolvimento do modelo

Posteriormente, a abordagem binária foi implementada para determinar se um requisito pertencia a uma das 11 classes analisadas neste estudo. Para esta categorização, os requisitos que receberam pelo menos duas classificações, independentemente do princípio ético atribuído, foram considerados positivos (valor 1). Aqueles que não atenderam a este critério, não sendo relacionados a princípios éticos foram classificados como negativos (valor 0). Embora a etapa de rotulagem tenha gerado uma classificação inicial, esta escolha teve como objetivo mitigar o impacto de classificações isoladas que poderiam introduzir viés ou inconsistência nos resultados. Por exemplo, pode haver situações em que apenas um rotulador classificou um requisito como relacionado aos princípios éticos em IA, enquanto os outros não concordaram. A implementação de cada tipo de estruturação está disponível na Seção 2.2, em Material Suplementar.

3.5 Fase 5 - Desenvolvimento do Modelo ERC4AI

As seções anteriores forneceram a base necessária para esta fase da pesquisa: a construção do modelo *Ethical Requirements Classifier for AI* (ERC4AI). Foram testados diferentes al-

goritmos de AM para identificar aquele com o melhor desempenho dentre as abordagens binária e multirótulo. Em particular, foram considerados os modelos pré-treinados BERT[4], XLM-RoBERTa[5] e DistilBERT[6]. Esses modelos foram escolhidos com base em seu desempenho em tarefas de PLN. Com arquiteturas adequadas tanto para classificação multirrótulo quanto binária, se destacam por considerar estruturas sintáticas, entender o contexto e a semântica textual com precisão. Além disso, eles têm a capacidade de generalizar bem em conjuntos de dados menores, pois realizam o pré-treinamento em um grande conjunto de dados, sendo eficazes na transferência de conhecimento [115, 116, 117, 118, 95, 29].

O desenvolvimento foi realizado na plataforma Google Colab[7],um ambiente Jupyter Notebook baseado em nuvem, que não requer configuração prévia e fornece acesso a recursos computacionais de alto desempenho. Esta ferramenta é amplamente utilizada para preparação e modelagem de dados, favorecendo a implementação e testes de algoritmos de ML com eficiência e escalabilidade. As etapas adotadas nesta fase serão detalhadas a seguir, e o código utilizado encontra-se disponível na plataformaZenodo.

3.5.1 Configurações e Desenvolvimento do Modelo

O processo de desenvolvimento do modelo foi dividido em duas tarefas. A primeira focou na abordagem multirrótulo, empregando modelos pré-treinados como XLM-RoBERTa [47], BERT [5] e DistilBERT [48]. A segunda tarefa usou os mesmos modelos na abordagem binária, explorando diferentes estratégias de classificação. Ambos os experimentos foram conduzidos usando os modelos pré-treinados disponíveis no repositório Hugging Face [8], que fornece acesso a soluções de última geração.

Para configurar os modelos, um conjunto predefinido de parâmetros foi selecionado levando em consideração descobertas em estudos anteriores e testes preliminares, conforme apresentado na Tabela 3.9, que descreve os parâmetros usados, juntamente com as justificativas para sua seleção. Além disso, os modelos passaram por um processo de ajuste fino, no qual seus pesos pré-treinados foram ajustados para se adaptar às especificidades do conjunto de dados deste estudo.

Neste estudo, a validação cruzada foi usada como uma técnica de reamostragem para as tarefas multirrótulo e binária. O método dividiu iterativamente os dados em subconjuntos de treinamento e validação, permitindo uma avaliação mais robusta da generalização do modelo e reduzindo o risco de overfitting, quando o modelo se ajusta excessivamente aos

[4] `https://huggingface.co/google-bert/bert-large-uncased`
[5] `https://huggingface.co/FacebookAI/xlm-roberta-large`
[6] `distilbert-base-multilingual-cased`
[7] `https://colab.google/`
[8] `https://huggingface.co/models`

Tabela 3.9: Parametrização do modelo

Parameter	Value	Justification
max_seq_length	128	The maximum input sequence length was set to 128. This value was chosen based on the maximum word count per text (86), providing a safe margin to account for all words and any additional tokens generated during tokenization, without causing data truncation. This choice is in line with the recommendations of Devlin et al. [5].
num_train_epochs	[1 - 12]	The number of complete iterations over the dataset during model training was specified. A range of 1 to 12 epochs was selected to balance efficiency and effectiveness, aiming to avoid both underfitting [119]—where the model doesn't learn enough from the data—and overfitting, where the model over-adapts to the training data and loses its ability to generalize [120].
learning_rate	[1e-5, 2e-5]	The learning rate used in training was set to two distinct values, 1e-5 and 2e-5, selected based on their stability and effectiveness in the learning process. The value 1e-5 was chosen for its ability to perform fine and gradual adjustments in the model weights, reducing the risk of overfitting, as evidenced by empirical experiments in similar contexts [121]. The value 2e-5 was also included for its successful application in previous studies, such as the BERT model [5].
manual_seed	42	A seed was set to ensure the reproducibility of results. The value 42 was chosen arbitrarily to maintain consistency across different runs of the model training process.

padrões dos dados de treinamento, comprometendo sua capacidade de generalizar para novos dados. A Figura 3.7 ilustra a estratégia adotada.

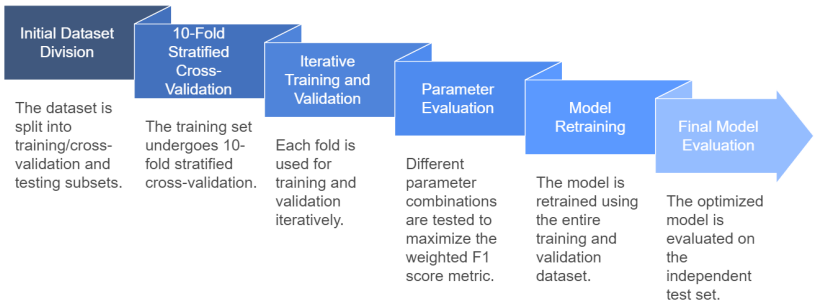

Figura 3.7: Processo de validação cruzada e avaliação de modelos

Inicialmente, o conjunto de dados foi dividido em dois subconjuntos: 80% dos dados foram usados para treinamento e validação cruzada, enquanto os 20% restantes foram reservados exclusivamente para teste do modelo final.

Usamos validação cruzada estratificada de 10 *folds*, uma técnica na qual a amostragem é realizada de uma forma que preserva as proporções de classes nos subconjuntos individuais, refletindo as proporções presentes no conjunto de dados completo. Kohavi [122] recomenda usar 10 *folds* em conjuntos de dados do mundo real, pois esse número equilibra a necessidade de estimativas confiáveis de precisão sem sobreajustar o modelo aos dados de treinamento.

Em cada iteração do processo de treinamento, 80% dos dados foram alocados para treinamento e 20% para validação, dentro dos *folds* de validação cruzada. As Tabelas 3.10 e 3.11 mostram a distribuição de dados por classe e divisão de dados, considerando os cenários de classificação multirrótulo e binária, respectivamente.

A Tabela 3.10 mostra o número total de rótulos por classe no cenário multirrótulo, além de sua divisão entre os conjuntos de treinamento e teste e as partições usadas durante a validação cruzada *(k-fold)*. Por exemplo, a classe *Transparency* tem 89 rótulos no total, dos quais 68 foram alocados para o conjunto de treinamento e 21 para o conjunto de teste, com 54 e 14 rótulos distribuídos, respectivamente, entre as partições de treinamento e validação na validação cruzada. Da mesma forma, a tabela 3.11 apresenta a distribuição de dados no contexto da classificação binária. Neste caso, a classe *Non-Ethical Principle* tem 937 rótulos no total, com 749 alocados para treinamento, 188 para teste e 599 e 150 destinados às partições de treinamento e validação durante a validação cruzada.

Tabela 3.10: Distribuição de dados por classe - Multirrótulo

Class	**Total**	**Training**	**Test**	**Training (k-fold)**	**Validation (k-fold)**
Transparency	89	68	21	54	14
Trust	68	51	17	41	10
Privacy	55	39	16	31	8
Responsibility	72	54	18	43	11
Other Principles	75	60	15	48	12
Non Ethical	937	751	186	601	150

Tabela 3.11: Distribuição de dados por classe - Binário

Class	**Total**	**Training**	**Test**	**Training (k-fold)**	**Validation (k-fold)**
Non-Ethical Principle	937	749	188	599	150
Ethical Principle	156	125	31	100	25

Diferentes combinações de parâmetros foram avaliadas com o objetivo de identificar a configuração que maximizasse a métrica *F1 Score* ponderada (*Weighted*), escolhida como a principal métrica de desempenho. Após selecionar os melhores parâmetros, o modelo foi retreinado usando o conjunto completo de dados de treinamento e validação.

Finalmente, o modelo otimizado foi avaliado no conjunto de teste independente, que representou 20% dos dados totais. Este procedimento permitiu que os resultados finais não fossem afetados pelo vazamento de informações entre os conjuntos de treinamento e teste.

Para uma visão geral detalhada desta implementação, consulte as Seções 2.3 e 2.4 do Material Suplementar. Os conjuntos de dados resultantes deste processo também estão disponíveis no Zenodo.

3.6 Síntese do Capítulo

Este capítulo abordou aspectos da criação do modelo *Ethical Requirements Classifier for AI (ERC4AI)*, começando pela revisão bibliográfica para identificação de práticas atuais e as principais lacunas na classificação de requisitos éticos em IA. Destacou a complexidade adicional na obtenção de dados para soluções éticas de IA, em comparação com softwares convencionais, devido à menor acessibilidade dos dados. Foi realizada rotulação de textos para construção da base de dados a ser utilizada no desenvolvimento do modelo. O capítulo ainda discute a importância do preparo e pré-processamento de dados textuais para melhoria do poder preditivo dos modelos. Na construção do modelo ERC4AI foram testados algoritmos de AM, como BERT [5], DistilBERT [48] e XLM-RoBERTa [47]. Todo o desenvolvimento foi realizado na plataforma *Google Colab*, aproveitando seus recursos computacionais escaláveis. A avaliação dos resultados e escolha do modelo que foi definido como o ERC4AI será detalhada no Capítulo 4.

Capítulo 4

Validação dos Modelos Classificadores

Este capítulo é dedicado à validação dos modelos classificadores desenvolvidos no Capítulo 3, detalhando os resultados obtidos com os algoritmos XLM-RoBERTa [47], BERT [5] e DistilBERT [48]. Para as abordagens multirótulo e binária são apresentadas as métricas de desempenho obtidas: Precisão [2], *Recall* [2], *F1 Score* [2], *Weighted AVG* [45] e Macro AVG [2, 46].

Os resultados foram obtidos por meio de validação cruzada estratificada com 10 *folds*, permitindo avaliar a capacidade de generalização dos modelos. Adicionalmente, um conjunto de testes independente, correspondente a 20% dos dados da base *EthicalRequirements4AI*, foi utilizado para a avaliação final, possibilitando a análise de desempenho sobre dados não vistos na etapa de treinamento, simulando cenários reais.

4.1 Modelo Multi-Rótulo

Nesta tarefa de modelagem, foram utilizados os dados preparados na estratégia multirótulo (ver Seção 3.4), a qual cada texto relacionado aos requisitos éticos foi associado a um ou mais princípios éticos em IA. A Tabela 4.1 apresenta as métricas médias para Precisão, *Recall* e *F1 Score*, juntamente com seus desvios padrão, calculados para cada rótulo nos diferentes modelos. Em geral, os requisitos apresentaram valores altos para as métricas, com taxas acima de 80%, indicando que os modelos foram capazes de identificar padrões úteis para classificar princípios éticos.

O desvio padrão das métricas fornece informações sobre a consistência dos resultados entre os *folds*. Rótulos como *Non-Ethical Principle* e *Privacy* obtiveram desvios padrão baixos, sugerindo maior estabilidade. Por outro lado, *Transparency* e *Responsibility*

apresentaram desvios-padrão ligeiramente maiores, refletindo variações mais significativas entre os *folds*, embora ainda com desempenho consistente.

Em geral, os modelos demonstraram resultados satisfatórios e estáveis para a maioria dos rótulos, sugerindo uma boa capacidade de generalização.

Tabela 4.1: Médias de desempenho e desvios padrão de modelos multirótulos com validação cruzada de 10 folds

Label	**Precision**		**Recall**		**F1 Score**	
	Mean	±Std Dev	Mean	±Std Dev	Mean	±Std Dev
XLM-RoBERTa						
Transparency	0.9532	0.0865	0.9238	0.1620	0.9302	0.1468
Trust	0.9311	0.1654	0.8665	0.2048	0.8886	0.1990
Privacy	0.9436	0.0939	0.9225	0.2002	0.9166	0.1853
Responsibility	0.9267	0.1149	0.8941	0.1666	0.9010	0.1593
Other Ethical Principles	0.9682	0.1008	0.9046	0.1800	0.9260	0.1650
Non Ethical Principle	0.9910	0.0204	0.9964	0.0051	0.9936	0.0119
BERT						
Transparency	0.9412	0.1097	0.9399	0.1275	0.9341	0.1331
Trust	0.9074	0.2430	0.8450	0.2758	0.8647	0.2716
Privacy	0.9498	0.0920	0.9294	0.1943	0.9218	0.1793
Responsibility	0.9054	0.1525	0.8936	0.1684	0.8921	0.1692
Other Ethical Principles	0.9737	0.0872	0.9050	0.1812	0.9286	0.1632
Non Ethical Principle	0.9922	0.0183	0.9948	0.0081	0.9934	0.0116
DistilBERT						
Transparency	0.9385	0.1186	0.9342	0.1360	0.9301	0.1411
Trust	0.9303	0.1690	0.8595	0.2406	0.8816	0.2260
Privacy	0.9339	0.1333	0.9044	0.2456	0.8932	0.2265
Responsibility	0.9068	0.1552	0.8764	0.2099	0.8831	0.1951
Other Ethical Principles	0.9628	0.1165	0.9125	0.1605	0.9305	0.1527
Non Ethical Principle	0.9919	0.0189	0.9947	0.0082	0.9932	0.0121

A Tabela 4.2 apresenta uma visão geral do desempenho dos modelos, considerando as médias Macro e *Weighted*. Esses valores permitem avaliar a capacidade dos modelos de lidar com diferentes rótulos, considerando tanto o equilíbrio entre as classes (Macro AVG) quanto o impacto das proporções dos dados (Weighted AVG). O XLM-RoBERTa se destacou durante a validação cruzada, obtendo a maior *F1 Score* AVG ponderada (97,2%) e métricas macro ligeiramente superiores aos outros modelos, sugerindo um desempenho equilibrado, mesmo para rótulos menos frequentes.

Tabela 4.2: Comparação de métricas gerais entre modelos multirótulos - validação cruzada

Model	**Metric**	**Precision**	**Recall**	**F1 Score**
XLM-Roberta	Macro Average	0.9523	0.9180	0.9260
	Weighted Average	0.9788	0.9713	0.9722
BERT	Macro Average	0.9449	0.9179	0.9224
	Weighted Average	0.9771	0.9704	0.9710
DistilBERT	Macro Average	0.9440	0.9136	0.9186
	Weighted Average	0.9767	0.9692	0.9700

Após a validação cruzada, os modelos foram avaliados em um conjunto de testes independente (20% dos dados do *EthicalRequirements4AI*) para medir seu desempenho em

dados não vistos. Conforme mostrado na Tabela 4.3, o BERT obteve o melhor desempenho geral, com uma *Weighted AVG F1 Score* de 88%, a mais alta entre os modelos. Ele também atingiu uma Macro AVG F1 Score de 72%, indicando desempenho equilibrado entre as classes.

O BERT se destacou em rótulos como *Transparency* (78%), *Responsibility* (81%) e *Non-Ethical Principle* (98%). Apesar do desempenho inferior em rótulos como *Trust* (56%) e *Other Ethical Principles* (53%), o modelo demonstrou consistência geral. XLM-RoBERTa, líder em validação cruzada, alcançou uma *Weighted AVG F1 Score* de 84%, enquanto DistilBERT alcançou 85%, ficando logo atrás de BERT.

Esses resultados destacam BERT como o modelo mais eficaz para a tarefa de classificação de vários rótulos no conjunto de teste, combinando equilíbrio de classe e desempenho geral.

Tabela 4.3: Métricas de avaliação final para modelos de classificação multirrótulo

Label	**Precision**	**Recall**	**F1 Score**
XLM-RoBERTa			
Transparency	0.56	0.43	0.49
Trust	0.69	0.65	0.67
Privacy	0.79	0.69	0.73
Responsibility	0.71	0.67	0.69
Other Ethical Principles	0.60	0.20	0.30
Non Ethical Principle	0.94	1.00	0.97
Macro AVG	0.71	0.60	0.64
Weighted AVG	0.85	0.85	0.84
BERT			
Transparency	0.72	0.86	0.78
Trust	0.50	0.65	0.56
Privacy	0.67	0.62	0.65
Responsibility	0.93	0.72	0.81
Other Ethical Principles	0.53	0.53	0.53
Non Ethical Principle	0.98	0.97	0.98
Macro AVG	0.72	0.73	0.72
Weighted AVG	0.89	0.88	0.88
DistilBERT			
Transparency	0.78	0.67	0.72
Trust	0.56	0.53	0.55
Privacy	0.58	0.44	0.50
Responsibility	0.82	0.50	0.62
Other Ethical Principles	0.57	0.53	0.55
Non Ethical Principle	0.96	0.99	0.98
Macro AVG	0.71	0.61	0.65
Weighted AVG	0.87	0.85	0.85

4.2 Modelo Binário

Os modelos XLM-RoBERTa, BERT e DistilBERT também foram avaliados no cenário de classificação binária. A Tabela 4.4 apresenta as médias e desvios padrão das métricas Precision, *Recall* e *F1 Score*, obtidas a partir de um processo de validação cruzada estratificada de 10 *folds*. Assim como na abordagem multirrótulo, esta etapa buscou

avaliar a consistência dos modelos em diferentes divisões dos dados e sua capacidade de generalização para novos exemplos.

A classe *Ethical Principle* apresentou variabilidade nos resultados entre os modelos, com BERT obtendo a maior *F1 Score* (96%), seguido pelo DistilBERT (95%) e XLM-RoBERTa (77%). Isso indica que alguns modelos foram mais eficazes em lidar com os desafios desta classe.

Tabela 4.4: Médias de desempenho e desvios padrão de modelos binários com validação cruzada de 10 folds

Label	**Precision**		**Recall**		**F1 Score**	
	Mean	±Std Dev	Mean	±Std Dev	Mean	±Std Dev
XLM-RoBERTa						
Non-Ethical Principle	0.9661	0.0478	0.9835	0.0268	0.9738	0.0265
Ethical Principle	0.9160	0.1152	0.7750	0.3327	0.7705	0.3205
BERT						
Non-Ethical Principle	0.9920	0.0171	0.9972	0.0042	0.9946	0.0096
Ethical Principle	0.9826	0.0267	0.9502	0.1120	0.9616	0.0905
DistilBERT						
Non-Ethical Principle	0.9922	0.0171	0.9947	0.0100	0.9934	0.0126
Ethical Principle	0.9667	0.0639	0.9518	0.1101	0.9560	0.1009

Os valores Macro e *Weighted* AVG, apresentados na Tabela 4.5, fornecem uma visão geral do desempenho geral dos modelos. O BERT se destacou ao atingir a maior *Weighted AVG F1 Score* (99%), sugerindo bom desempenho mesmo quando se considera o impacto das proporções dos dados. O DistilBERT apresentou resultados comparáveis, enquanto o XLM-RoBERTa teve métricas ligeiramente inferiores, possivelmente devido ao menor desempenho em rótulos menos frequentes.

Tabela 4.5: Comparação de métricas gerais entre modelo binários - Validação cruzada

Model	**Metric**	**Precision**	**Recall**	**F1 Score**
XLM-Roberta	Macro Average	0.9410	0.8793	0.8722
	Weighted Average	0.9589	0.9537	0.9448
BERT	Macro Average	0.9873	0.9737	0.9781
	Weighted Average	0.9907	0.9905	0.9899
DistilBERT	Macro Average	0.9794	0.9733	0.9747
	Weighted Average	0.9886	0.9886	0.9881

No conjunto de teste, os resultados reforçam essa tendência, conforme mostrado na Tabela 4.6. O BERT permaneceu como o modelo de melhor desempenho, com uma *F1 Score* de 97% para o rótulo *Non-Ethical Principle* e 81% para *Ethical Principle*. O DistilBERT também apresentou resultados consistentes, com valores próximos aos do BERT, enquanto o XLM-RoBERTa, embora competitivo no rótulo mais representativo, mostrou desempenho mais limitado no rótulo *Ethical Principle*.

Tabela 4.6: Métricas de avaliação final para modelos de classificação binária

Label	Precision	Recall	F1 Score
XLM-RoBERTa			
Non-Ethical Principle	0.93	0.98	0.96
Ethical Principle	0.85	0.55	0.67
Macro AVG	0.89	0.77	0.81
Weighted AVG	0.92	0.92	0.92
BERT			
Non-Ethical Principle	0.96	0.98	0.97
Ethical Principle	0.88	0.74	0.81
Macro AVG	0.92	0.86	0.89
Weighted AVG	0.95	0.95	0.95
DistilBERT			
Non-Ethical Principle	0.97	0.96	0.97
Ethical Principle	0.76	0.84	0.80
Macro AVG	0.87	0.90	0.88
Weighted AVG	0.94	0.94	0.94

Os resultados apresentados destacam a importância das abordagens binária e multirrótulo na tarefa de classificação de requisitos éticos em sistemas de IA. Na abordagem multirrótulo, os modelos demonstraram capacidade de lidar com múltiplas classes simultaneamente, obtendo boas métricas e estabilidade em avaliações de validação cruzada e testes. Destaca-se o BERT, que apresentou o melhor desempenho geral no conjunto de testes, conciliando equilíbrio entre classes e resultados satisfatórios para rótulos menos frequentes.

Na abordagem binária, os modelos demonstraram maior capacidade de distinguir entre os rótulos *Ethical Principle* e *Non-Ethical Principle*, sendo particularmente úteis para a triagem inicial de requisitos em projetos de IA. O BERT novamente se destacou como o modelo mais eficaz, apresentando métricas consistentes e superando ligeiramente o DistilBERT em Precisão e *Recall* para a classe *Ethical Principle*, que é o foco principal do estudo. Esses resultados indicam que a abordagem binária pode facilitar a identificação automatizada de requisitos que possuam aspectos éticos, reduzindo a necessidade de triagem manual e tornando o processo mais eficiente.

Os modelos BERT, tanto na abordagem binária quanto na multirrótulo, foram selecionados como ERC4AI devido ao seu desempenho na classificação de requisitos éticos em IA. Essas versões estão disponíveis como ERC4AI-Binary e ERC4AI-Multilabel, e podem ser acessadas no repositório Zenodo.

4.3 Ferramenta Web para Classificação de Requisitos Éticos em IA

Para facilitar a aplicação prática do modelo ERC4AI, foi desenvolvido um protótipo de ferramenta web que permite a classificação automática de requisitos de software para sistemas de IA. Seu principal objetivo é tornar esse processo acessível e intuitivo, possi-

bilitando que os usuários insiram um requisito textual e obtenham sua classificação de forma automatizada. A ferramenta utiliza os dois modelos de classificação desenvolvidos: binário e multirrótulo.

A Figura 4.1 apresenta a interface do protótipo da ferramenta web, onde o usuário pode:

- Inserir um requisito textual (em inglês) no campo disponível.
- Selecionar o tipo de classificação desejada (Binária ou Multirrótulo).
- Clicar no botão de "Classify Requirements"para iniciar a classificação.

  O requisito inserido passa por um processo de preparação dos dados, conforme descrito na Seção 3.4. Em seguida, o modelo escolhido (binário ou multirrótulo) processa o requisito e retorna a classificação correspondente.
- Visualizar o resultado da classificação.
  - Se a classificação binária for selecionada, o modelo indicará se o requisito está relacionado a princípios éticos em IA ou não.
  - Se a classificação multirrótulo for utilizada, o sistema identificará quais princípios éticos o requisito atende, podendo atribuir múltiplos rótulos ou classificá-lo como não pertencente a nenhuma das categorias de princípios éticos.

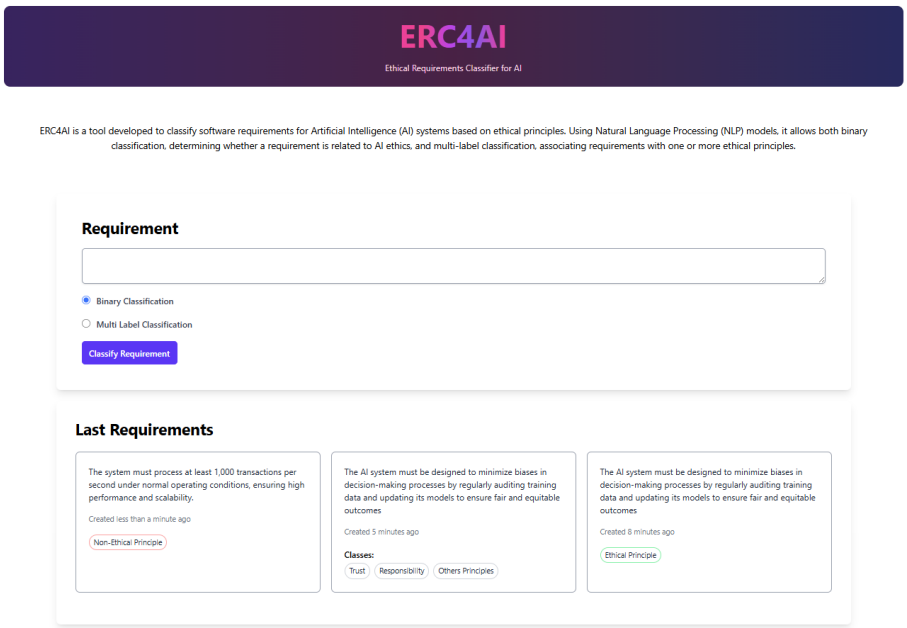

Figura 4.1: ERC4AI Web Tool

A ferramenta foi disponibilizada como software de código aberto, permitindo que pesquisadores e profissionais possam acessá-la, testá-la e contribuir para sua evolução. O acesso pode ser feito através dos links abaixo:

- Ferramenta web – Interface para inserção e classificação de requisitos.
- Repositório no Github – Código-fonte e documentação técnica.

4.4 Síntese do Capítulo

Neste capítulo, foram avaliados os resultados das abordagens multirrótulo e binária para os modelos XLM-RoBERTa [47], BERT [5] e DistilBERT [48]. Durante essa análise, as métricas *Precision*, *Recall*, *F1 Score*, *Weighted AVG* e *Macro AVG* foram consideradas para cada modelo. O BERT [5] apresentou desempenho competitivo em ambas as abordagens, destacando-se entre os modelos avaliados.

Capítulo 5

Discussão dos Resultados

Este capítulo apresenta os resultados do ERC4AI, respondendo às perguntas de pesquisa e contextualizando os achados em relação aos objetivos do trabalho. Além disso, explora suas contribuições práticas e acadêmicas, analisa as limitações do estudo e aborda as ameaças à validade, bem como as estratégias adotadas para mitigá-las.

5.1 Discussão dos Resultados

Os resultados obtidos com o modelo ERC4AI destacam seu potencial para avançar a ética prática da IA por meio da classificação de requisitos éticos em sistemas de IA. Essa abordagem, até onde sabemos, é nova na literatura. Ao classificar os requisitos em princípios éticos na IA, o ERC4AI fornece uma ferramenta nova e prática para pesquisadores e profissionais.

5.1.1 RQ.1: Como o conjunto de dados de requisitos com anotações estruturadas tem o potencial de contribuir para a operacionalização de princípios éticos no desenvolvimento de sistemas de IA?

O conjunto de dados *EthicalRequirements4AI* foi desenvolvido a partir da base *Passenger Flow* [44], ampliando seu escopo para cobrir 11 princípios éticos de IA, em comparação aos 4 princípios considerados no estudo original. Além disso, ele estende o conjunto de dados PROMISE [1], que classifica requisitos como FR ou NFR, ao adicionar a rotulação de 34 requisitos relacionados a princípios éticos em IA e incorporar 934 requisitos não associados aos princípios éticos. Como resultado, o *EthicalRequirements4AI* contém um total de 156 requisitos éticos, complementados por 937 requisitos não relacionados à ética

em IA provenientes do PROMISE, totalizando 1.093 requisitos rotulados, que serviram como a verdade fundamental para o processo de treinamento dos modelos.

O *EthicalRequirements4AI* pode promover a operacionalização de princípios éticos em IA ao oferecer um recurso estruturado onde os requisitos são anotados com seu princípio ético em IA correspondente. Para trazer maior diversidade de perspectivas ao processo de anotação, a tarefa foi realizada por cinco especialistas com diferentes habilidades relacionadas à ética de IA. Essa abordagem colaborativa buscou reduzir potenciais vieses individuais e fornecer classificações mais confiáveis. Este conjunto de dados fornece aos desenvolvedores um mapeamento de requisitos para 11 princípios éticos de IA apresentados por Ryan e Stahl [31], servindo como uma contribuição valiosa no campo da ética de IA na prática, que é uma área contínua de pesquisa [41].

Ao disponibilizar este conjunto de dados publicamente, o estudo permite que os pesquisadores expandam esta base, aumentando seu tamanho ou refinando suas anotações. Este recurso estruturado ajuda a abordar a atual escassez de ferramentas e orientação prática para implementar a ética de IA, conforme destacado em estudos anteriores [123, 41, 40, 103, 17].

O processo de confiabilidade entre avaliadores revelou que os princípios em que dois ou mais rotuladores concordaram foram *Transparency*, com 89 requisitos, seguido por *Responsibility*, com 72 requisitos, e *Trust*, com 68 requisitos. Por outro lado, os princípios com menor concordância entre os rotuladores foram *Dignity*, com 5 requisitos, *Solidarity*, com 3 requisitos, e *Sustainability*, com apenas 1 requisito.

Isso pode sugerir que certos princípios éticos em IA (por exemplo, *Transparency*, *Responsibility* e *Trust*) são mais facilmente identificados por humanos em comparação a outros (por exemplo, *Dignity*, *Solidarity* e *Sustainability*), que tendem a ser mais abstratos e desafiadores para chegar a um consenso entre os rotuladores. Isso está de acordo com a subjetividade inerente à ética da IA encontrada em estudos anteriores: "A operacionalização de princípios e diretrizes éticas em IA está sujeita à subjetividade daqueles envolvidos no processo de elicitação e, mais importante, dos desenvolvedores." [42].

Agrupamos classes minoritárias em uma nova categoria chamada *Other Ethical Principles*. Isso foi motivado pela baixa frequência das classes, o que poderia afetar o desempenho geral do modelo ao dificultar a identificação de padrões consistentes. O agrupamento permitiu a representatividade dos dados, reduzindo possíveis vieses de sub-representação [101]. Essa abordagem possibilitou que princípios éticos com menor frequência ainda fossem representados no conjunto de dados, evitando que fossem desconsiderados devido à baixa ocorrência.

Portanto, o objetivo desta compilação é servir de base para o desenvolvimento do modelo *Ethical Requirements Classifier for AI* (ERC4AI). Embora o conjunto de dados

EthicalRequirements4AI cubra inicialmente todos os 11 princípios éticos de IA, o processo entre avaliadores resultou em categorias de classificação refinadas: *Transparency*, *Trust*, *Privacy*, *Responsibility*, *Other Ethical Principles* e *Non-Ethical Principles*. Portanto, essas são as categorias nas quais o modelo ERC4AI é treinado na abordagem multirótulo.

5.1.2 RQ.2: Como as técnicas de PLN podem auxiliar na classificação de requisitos éticos em projetos de IA?

Neste estudo, três modelos de linguagem — XLM-RoBERTa, BERT e DistilBERT — foram avaliados em cenários de classificação binária e multirrótulo para fornecer uma análise comparativa e justificar a seleção do modelo mais adequado para nosso contexto. No cenário de classificação binária, os requisitos foram categorizados como *Ethical Principle* (relacionado a algum dos princípios éticos de IA) ou *Non-Ethical Principle* (não relacionado aos princípios éticos de IA). Da análise comparativa dos resultados da classificação binária, as métricas *Weighted AVG F1 Scores* no conjunto de teste foram as seguintes: XLM-RoBERTa 92%, DistilBERT 94% e BERT 95%. Portanto, BERT demonstrou o melhor desempenho para esta tarefa.

Em relação à estratégia de classificação multirrótulo, nossa análise comparativa indica que as *Weighted AVG F1 Scores* são 84% para XLM-RoBERTa, 85% para DistilBERT e 88% para BERT no conjunto de teste. Portanto, BERT demonstra desempenho superior nesta tarefa também. Em termos de desempenho específico de classe dentro de BERT, as classes de melhor desempenho são *Non-Ethical Principle* (98%), *Responsibility* (81%) e *Transparency* (78%). Isso destaca a capacidade do modelo de classificar essas classes com maior precisão. Um fator contribuinte é o maior número de instâncias para essas classes no conjunto de dados. No entanto, vale ressaltar que *Transparency*, apesar de ter menos instâncias do que *Responsibility*, ainda atinge um desempenho relativamente alto. Por outro lado, as classes de menor desempenho dentro do BERT são *Other Principles* (53%), *Trust* (56%) e *Privacy* (65%). Esses resultados sugerem uma oportunidade para trabalho futuro para enriquecer o conjunto de dados, incluindo mais requisitos rotulados sob esses princípios para aumentar a capacidade do modelo de classificá-los com mais assertividade. O ERC4AI multirrótulo , também baseado no BERT, demonstra o potencial em PLN na classificação de requisitos para vários princípios éticos de IA, o que representa grande contribuição neste estudo por sua novidade na literatura.

Ambos os modelos, binário e multirrótulo, demonstram o valor das técnicas de PLN na automação da classificação de requisitos éticos. O uso de modelos avançados de PLN permite que os desenvolvedores classifiquem os requisitos de forma eficiente, reduzindo assim o esforço manual e a experiência tradicionalmente necessária para essa tarefa [20, 1].

5.1.3 Implicações Práticas

Classificar requisitos é essencial para o gerenciamento eficaz de projetos, pois melhora a utilização de recursos, a documentação e minimiza o retrabalho [1], pois classificar requisitos manualmente costuma ser uma tarefa demorada, trabalhosa e custosa [20]. Embora a classificação manual possa funcionar para pequenos conjuntos de dados, ela se torna impraticável e exige muitos recursos para conjuntos de dados maiores, tornando a classificação automatizada uma alternativa valiosa que oferece consistência e precisão [21, 22, 23]– embora seu sucesso dependa da qualidade do modelo e dos dados [124].

Ao automatizar a identificação de princípios éticos no estágio inicial do desenvolvimento de sistemas de IA, o modelo facilita a documentação desses requisitos e auxilia na priorização, contribuindo para a criação de sistemas mais alinhados a diretrizes éticas.

Nossa análise inicial sugere o uso potencial do ERC4AI tanto para a academia quanto para a indústria. Para os desenvolvedores, a capacidade de categorizar os requisitos de acordo com os princípios éticos de IA oferece uma maneira prática de abordar as preocupações éticas de IA desde os primeiros estágios de desenvolvimento. Além disso, o modelo pode ser de grande utilidade para sistemas legados, auxiliando na análise e reclassificação de requisitos já existentes sob a perspectiva de princípios éticos. Para pesquisadores, o conjunto de dados e os modelos disponíveis publicamente fornecem uma base valiosa para o avanço de ferramentas e métodos em ética de IA na prática.

A complexidade inerente dos sistemas de IA representa um desafio significativo para profissionais na compreensão e aplicação de princípios éticos [125]. Isso ressalta que a definição de requisitos éticos implementáveis em IA é uma tarefa desafiadora e ainda em debate [125, 42].

Em suma, o ERC4AI representa um passo à frente na operacionalização da ética de IA. Atualmente, ferramentas para abordar considerações éticas em IA são escassas, muitas vezes carecem de documentação adequada e não há um padrão estabelecido da indústria ou estrutura consistente para orientar seu uso [123, 40, 42, 41]. Muitas ferramentas existentes abordam apenas um número limitado de princípios éticos, frequentemente focando estritamente em aspectos, por exemplo, transparência e explicabilidade [123, 64]. Em contraste, nosso modelo proposto fornece uma abordagem eficaz para classificar requisitos — 4 princípios éticos de IA individualmente, *Non Ethical Principle* e *Other Principles* —, servindo como uma nova ferramenta prática para auxiliar desenvolvedores na criação de sistemas de IA mais alinhados eticamente. Ao facilitar a categorização de requisitos éticos e seu alinhamento com princípios éticos de IA, este trabalho ajuda desenvolvedores a abordar preocupações éticas desde os primeiros estágios do desenvolvimento do sistema baseados em IA.

Apesar dos resultados promissores, este estudo reconhece oportunidades de melhoria. O aprimoramento do conjunto de dados *EthicalRequirements4AI*, com maior diversidade e um número ampliado de requisitos anotados, pode fortalecer a robustez do modelo e melhorar o equilíbrio entre classes. Além disso, a incorporação de estruturas emergentes, como o EU AI Act, e a adoção de técnicas avançadas de IA generativa — incluindo *Large Language Models* (LLMs) e abordagens como *Retrieval-Augmented Generation* (RAG) e ajuste fino — podem aprimorar ainda mais a precisão da classificação e sua aplicabilidade prática.

Tornamos o conjunto de dados e os modelos — EthicalRequirements4AI e ERC4AI, binário e multirrótulo — disponíveis publicamente como recursos de código aberto. Essas contribuições têm como objetivo não apenas ajudar os desenvolvedores a construir sistemas de IA mais éticos, mas também ajudar os pesquisadores a desenvolverem ferramentas práticas para operacionalizar a ética da IA.

5.2 Ameaças a Validade

Este trabalho enfrenta uma ameaça interna relacionada ao viés no processo de rotulagem manual de dados, o que pode comprometer a precisão das classificações. Para mitigar esse risco, incluímos quatro rotuladores adicionais. No entanto, divergências na interpretação entre os rotuladores ainda podem ocorrer, afetando a consistência das classificações. Para minimizar esses efeitos, um requisito foi atribuído a uma classe específica somente quando houve concordância entre pelo menos dois rotuladores. Essa abordagem colaborativa ajudou a reduzir inconsistências e aumentar a confiabilidade do processo de rotulagem.

Quanto à validade externa do estudo, incluímos o conjunto de dados *Passenger Flow*, mesmo sendo proveniente de uma pré-impressão, devido à escassez de pesquisas que abordam a ética da IA na prática durante a fase de engenharia de requisitos. Além disso, sua contribuição foi considerada relevante para os objetivos desta pesquisa.

Uma ameaça adicional está na diversidade dos dados utilizados. Embora o conjunto de dados *EthicalRequirements4AI* seja amplo, observamos um desequilíbrio na representação de alguns princípios éticos em IA. Para mitigar esse problema e permitir maior imparcialidade ao modelo, agrupamos as classes sub-representadas e aplicamos a amostragem estratificada na divisão dos dados de treinamento e teste. Essa estratégia permitiu uma distribuição mais equilibrada entre as categorias nas abordagens binária e multirrótulo, tornando a avaliação do modelo mais consistente.

Além disso, a validação cruzada estratificada foi adotada para minimizar a instabilidade gerada pelo desequilíbrio nos dados de validação, mantendo a proporção relativa das classes nos subconjuntos. Essa abordagem está alinhada com achados da literatura,

onde a validação cruzada estratificada é reconhecida como um método eficaz para lidar com conjuntos de dados desbalanceados, melhorando a generalização dos modelos [126, 127, 128].

A validade de construto também apresenta desafios, uma vez que parte do conjunto de dados de treinamento incluiu requisitos de software tradicionais. Devido à similaridade textual, alguns desses requisitos foram considerados aplicáveis como requisitos éticos em IA. Isso pode gerar questionamentos sobre até que ponto esses requisitos realmente capturam princípios éticos. No entanto, essa questão foi mitigada pelo processo de avaliação dos rotuladores, que analisaram a adequação de cada requisito dentro do contexto ético, permitindo que apenas aqueles considerados pertinentes fossem mantidos.

Para enfrentar esse desafio e aprimorar a capacidade do modelo de distinguir entre requisitos alinhados a princípios éticos e aqueles que não são, mesmo diante de similaridades textuais, adotamos modelos de linguagem pré-treinados, como XLM-RoBERTa, BERT e DistilBERT. Esses modelos foram selecionados devido ao seu bom desempenho em tarefas de PLN, sendo adequados tanto para classificação multirrótulo quanto binária. Além de capturar aspectos sintáticos e semânticos do texto com precisão, esses modelos possuem a vantagem de serem pré-treinados em grandes volumes de dados, o que permite uma melhor generalização, mesmo quando aplicados a conjuntos menores.

5.3 Síntese do Capítulo

Neste capítulo, foram apresentados e analisados os resultados do ERC4AI, juntamente com suas contribuições, incluindo a disponibilização do conjunto de dados *EthicalRequirements4AI* e modelos binário e multirótulo. Apesar das limitações, como o desbalanceamento de classes, os resultados indicam o potencial desses modelos para apoiar a integração de princípios éticos em sistemas de IA.

Capítulo 6

Conclusões

Neste trabalho, apresentamos a base de dados *EthicalRequirements4AI*, que reúne 156 requisitos distribuídos entre 11 princípios éticos de IA [31], além de 937 requisitos não relacionados à ética em IA, totalizando 1.093 requisitos rotulados. Esse conjunto de dados contribui para a operacionalização de princípios éticos, ao fornecer um recurso estruturado que associa cada requisito ao princípio ético correspondente, podendo facilitar sua aplicação prática em contextos de desenvolvimento de sistemas de IA.

Realizou-se uma análise comparativa entre três modelos de linguagem – XLM-RoBERTa, BERT e DistilBERT – tanto na classificação binária quanto na multirrótulo. BERT foi o modelo de melhor desempenho com uma métrica *Weighted AVG F1 Score* de 95% para o modelo binário e 88% para multirrótulo no conjunto de teste, consolidando-se assim como a base para o modelo ERC4AI, disponível nas versões ERC4AI-Binary e ERC4AI-Multirrótulo.

O modelo ERC4AI-Binary classifica requisitos em *Ethical Principle*, quando relacionados à ética da IA, ou *Non-Ethical Principle*, quando não possuem essa relação. Já o modelo ERC4AI-Multilabel categoriza os requisitos em *Transparency* (78%), *Responsibility* (81%), *Trust* (56%), *Privacy* (65%), *Other Principles* (53%) e *Non-Ethical Principle* (98%).

O ERC4AI é uma ferramenta com potencial de auxiliar desenvolvedores na incorporação de princípios éticos em sistemas de IA desde as fases iniciais do desenvolvimento. Além disso, contribui para o avanço da ética aplicada à IA ao fornecer um recurso estruturado para pesquisadores e profissionais da área.

Com esse trabalho, buscou-se impulsionar o desenvolvimento de soluções que favoreçam a implementação dos princípios éticos na prática, promovendo maior alinhamento ético nos sistemas de IA. O conjunto de dados *EthicalRequirements4AI*, bem como as versões ERC4AI-Binary e ERC4AI-Multilabel do modelo, estão disponíveis publicamente no Repositório Zenodo.

Os trabalhos futuros incluem:

- Incorporar novos regulamentos, por exemplo, o EU AI Act, como base para o processo de rotulação.
- Envolver um maior número de anotadores com experiência em ética e regulamentos de IA para enriquecer a diversidade de perspectivas e aprimorar a qualidade das anotações. Além disso, desenvolver um guia formal de anotação, contendo exemplos práticos, para a padronização do processo de rotulagem, minimizando ambiguidades e aumentando a confiabilidade das classificações.
- Expandir o tamanho do conjunto de dados *EthicalRequirements4AI* com mais requisitos anotados.
- Avaliar o modelo utilizando um conjunto de testes distinto, incluindo dados não utilizados no treinamento, com revisão realizada por avaliadores externos à pesquisa, permitindo uma análise independente do desempenho do modelo.
- Comparar o desempenho do ERC4AI com técnicas mais recentes, baseadas em LLM, incluindo modelos como Deepseek[1], GPT-4 [2], Llama 2 [3] e Mistral [4].
- Realizar uma avaliação da ferramenta web com engenheiros de requisitos de sistemas baseados em IA, analisando sua usabilidade e aplicabilidade prática. Esse estudo permitirá identificar melhorias na interface, na experiência do usuário e na adequação dos resultados às necessidades reais desses profissionais.

[1] `https://www.deepseek.com/`
[2] `https://chatgpt.com/`
[3] `https://www.llama.com/llama2/`
[4] `https://mistral.ai/`

Referências

[1] Canedo, Edna Dias e Bruno Cordeiro Mendes: *Software requirements classification using machine learning algorithms*. Entropy, 22(9):1057, 2020. ix, 2, 5, 7, 28, 33, 35, 36, 37, 57, 59, 60

[2] Grandini, Margherita, Enrico Bagli e Giorgio Visani: *Metrics for multi-class classification: an overview*. CoRR, abs/2008.05756, 2020. `https://arxiv.org/abs/2008.05756`. x, 6, 8, 24, 25, 26, 50

[3] Huxohl, Tamino: *Deep Learning Based Salient Object Detection for the Detection of Stains and Holes on Patterned Laundry*. Tese de Doutoramento, Bielefeld University, Germany, 2023. `https://pub.uni-bielefeld.de/record/2979082`. 1

[4] Wang, Xiaotian, Zhongjie Pan, Hang Gao, Ningxin He e Tiegang Gao: *An efficient model for real-time wildfire detection in complex scenarios based on multi-head attention mechanism*. J. Real Time Image Process., 20(4):66, 2023. `https://doi.org/10.1007/s11554-023-01321-8`. 1

[5] Devlin, Jacob, Ming-Wei Chang, Kenton Lee e Kristina Toutanova: *BERT: pre-training of deep bidirectional transformers for language understanding*. Em Burstein, Jill, Christy Doran e Thamar Solorio (editores): *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, páginas 4171–4186. Association for Computational Linguistics, 2019. `https://doi.org/10.18653/v1/n19-1423`. 1, 6, 8, 23, 24, 46, 47, 49, 50, 56

[6] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever e Dario Amodei: *Language models are few-shot learners*. Em Larochelle, Hugo, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan e Hsuan-Tien Lin (editores): *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. `https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html`. 1

[7] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever *et al.*: *Language models are unsupervised multitask learners*. OpenAI blog, 1(8):9, 2019. 1

[8] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville e Yoshua Bengio: *Generative adversarial networks*. Commun. ACM, 63(11):139–144, 2020. https://doi.org/10.1145/3422622. 1

[9] Qurashi, Abdul Wahab, Zohaib A. Farhat, Violeta Holmes e Anju P. Johnson: *New avenues for automated railway safety information processing in enterprise architecture: An NLP approach*. IEEE Access, 11:44413–44424, 2023. https://doi.org/10.1109/ACCESS.2023.3272610. 1

[10] Taecharungroj, Viriya: *"what can chatgpt do?" analyzing early reactions to the innovative AI chatbot on twitter*. Big Data Cogn. Comput., 7(1):35, 2023. https://doi.org/10.3390/bdcc7010035. 1

[11] Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu e Mark Chen: *Hierarchical text-conditional image generation with CLIP latents*. CoRR, abs/2204.06125, 2022. https://doi.org/10.48550/arXiv.2204.06125. 1

[12] Acer, Irem, Firat Orhan Bulucu, Semra Içer e Fatma Latifoglu: *Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations*. Turkish J. Electr. Eng. Comput. Sci., 31(1):112–125, 2023. https://doi.org/10.55730/1300-0632.3974. 1

[13] Bhutoria, Aditi: *Personalized education and artificial intelligence in the united states, china, and india: A systematic review using a human-in-the-loop model*. Comput. Educ. Artif. Intell., 3:100068, 2022. https://doi.org/10.1016/j.caeai.2022.100068. 1

[14] Awotunde, Joseph Bamidele, Sanjay Misra, Foluso Ayeni, Rytis Maskeliunas e Robertas Damasevicius: *Artificial intelligence based system for bank loan fraud prediction*. Em Abraham, Ajith, Patrick Siarry, Vincenzo Piuri, Niketa Gandhi, Gabriella Casalino, Oscar Castillo e Patrick Hung (editores): *Hybrid Intelligent Systems - 21st International Conference on Hybrid Intelligent Systems (HIS 2021), December 14-16, 2021*, volume 420 de *Lecture Notes in Networks and Systems*, páginas 463–472. Springer, 2021. https://doi.org/10.1007/978-3-030-96305-7_43. 1

[15] Strümke, Inga, Marija Slavkovik e Vince Istvan Madai: *The social dilemma in artificial intelligence development and why we have to solve it*. AI Ethics, 2(4):655–665, 2022. https://doi.org/10.1007/s43681-021-00120-w. 1

[16] O'neil, Cathy: *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017. 1

[17] Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke e Effy Vayena: *Ai4people - an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations.* Minds Mach., 28(4):689–707, 2018. https://doi.org/10.1007/s11023-018-9482-5. 2, 58

[18] Li, Bo, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi e Bowen Zhou: *Trustworthy AI: from principles to practices.* ACM Comput. Surv., 55(9):177:1–177:46, 2023. https://doi.org/10.1145/3555803. 2

[19] Lim, Sachiko, Aron Henriksson e Jelena Zdravkovic: *Data-driven requirements elicitation: A systematic literature review.* SN Comput. Sci., 2(1):16, 2021. https://doi.org/10.1007/s42979-020-00416-4. 2

[20] Quba, Gaith Y., Hadeel Al Qaisi, Ahmad Althunibat e Shadi AlZu'bi: *Software requirements classification using machine learning algorithm's.* Em *International Conference on Information Technology, ICIT 2021, Amman, Jordan, July 14-15, 2021*, páginas 685–690. IEEE, 2021. https://doi.org/10.1109/ICIT52682.2021.9491688. 2, 59, 60

[21] Arora, Chetan, Mehrdad Sabetzadeh, Lionel C. Briand e Frank Zimmer: *Automated checking of conformance to requirements templates using natural language processing.* IEEE Trans. Software Eng., 41(10):944–968, 2015. https://doi.org/10.1109/TSE.2015.2428709. 2, 60

[22] Rodrigues, Andre e Alberto Rodrigues da Silva: *Validation of rigorous requirements specifications and document automation with the itlingo RSL language.* CoRR, abs/2312.10822, 2023. https://doi.org/10.48550/arXiv.2312.10822. 2, 60

[23] Maalem, Sourour e Nacereddine Zarour: *Challenge of validation in requirements engineering.* J. Innov. Digit. Ecosyst., 3(1):15–21, 2016. https://doi.org/10.1016/j.jides.2016.05.001. 2, 60

[24] Casamayor, Agustin, Daniela Godoy e Marcelo R. Campo: *Identification of non-functional requirements in textual specifications: A semi-supervised learning approach.* Inf. Softw. Technol., 52(4):436–445, 2010. https://doi.org/10.1016/j.infsof.2009.10.010. 2, 28, 30

[25] Yucalar, Fatih: *Developing an advanced software requirements classification model using bert: An empirical evaluation study on newly generated turkish data.* Applied Sciences, 13(20):11127, 2023. 2, 29, 30

[26] Kurtanović, Zijad e Walid Maalej: *Automatically classifying functional and non-functional requirements using supervised machine learning.* Em *2017 IEEE 25th International Requirements Engineering Conference (RE)*, páginas 490–495. Ieee, 2017. 2

[27] Abad, Zahra Shakeri Hossein, Oliver Karras, Parisa Ghazi, Martin Glinz, Guenther Ruhe e Kurt Schneider: *What works better? a study of classifying requirements*. Em *2017 IEEE 25th International Requirements Engineering Conference (RE)*, páginas 496–501. IEEE, 2017. 2

[28] Dalpiaz, Fabiano, Davide Dell'Anna, Fatma Basak Aydemir e Sercan Çevikol: *Requirements classification with interpretable machine learning and dependency parsing*. Em *2019 IEEE 27th International Requirements Engineering Conference (RE)*, páginas 142–152. IEEE, 2019. 2

[29] Hey, Tobias, Jan Keim, Anne Koziolek e Walter F. Tichy: *Norbert: Transfer learning for requirements classification*. Em Breaux, Travis D., Andrea Zisman, Samuel Fricker e Martin Glinz (editores): *28th IEEE International Requirements Engineering Conference, RE 2020, Zurich, Switzerland, August 31 - September 4, 2020*, páginas 169–179. IEEE, 2020. https://doi.org/10.1109/RE48521.2020.00028. 2, 29, 30, 46

[30] Dell'Anna, Davide, Fatma Basak Aydemir e Fabiano Dalpiaz: *Evaluating classifiers in SE research: the ECSER pipeline and two replication studies*. Empir. Softw. Eng., 28(1):3, 2023. https://doi.org/10.1007/s10664-022-10243-1. 2

[31] Ryan, Mark e Bernd Carsten Stahl: *Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications*. J. Inf. Commun. Ethics Soc., 19(1):61–86, 2021. https://doi.org/10.1108/JICES-12-2019-0138. 2, 7, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 32, 34, 39, 58, 63

[32] Wagner, Matthias, Markus Borg e Per Runeson: *Navigating the upcoming european union AI act*. IEEE Softw., 41(1):19–24, 2024. https://doi.org/10.1109/MS.2023.3322913. 3

[33] Cerqueira, José Antonio Siqueira de, Heloise Acco Tives Leão e Edna Dias Canedo: *Ethical guidelines and principles in the context of artificial intelligence*. Em Araújo, Rafael D., Fabiano A. Dorça, Renata Mendes de Araujo, Sean W. M. Siqueira e Awdren L. Fontão (editores): *SBSI 2021: XVII Brazilian Symposium on Information Systems, Uberlândia, Brazil, June 7 - 10, 2021*, páginas 36:1–36:8. ACM, 2021. https://doi.org/10.1145/3466933.3466969. 3, 11

[34] Rousi, Rebekah, Ville Vakkuri, Paulius Daubaris, Simo Linkola, Hooman Samani, Niko Mäkitalo, Erika Halme, Mamia Agbese, Rahul Mohanani, Tommi Mikkonen e Pekka Abrahamsson: *Beyond 100 ethical concerns in the development of robot-to-robot cooperation*. Em *2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), Espoo, Finland, August 22-25, 2022*, páginas 420–426. IEEE, 2022. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics55523.2022.00092. 3

[35] Jobin, Anna, Marcello Ienca e Effy Vayena: *The global landscape of AI ethics guidelines*. Nat. Mach. Intell., 1(9):389–399, 2019. https://doi.org/10.1038/s42256-019-0088-2. 3, 11

[36] Widder, David Gray, Derrick Zhen, Laura Dabbish e James D. Herbsleb: *It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?* Em *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, páginas 467–479. ACM, 2023. https://doi.org/10.1145/3593013.3594012. 3

[37] Shneiderman, Ben: *Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems*. ACM Trans. Interact. Intell. Syst., 10(4):26:1–26:31, 2020. https://doi.org/10.1145/3419764. 3

[38] Lords, House Of *et al.*: *Ai in the uk: ready, willing and able?* Retrieved August, 13:2021, 2018. 4

[39] Chatila, Raja, Kay Firth-Butterfield e John C Havens: *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems version 2*. University of southern California Los Angeles, 2018. 4

[40] Vakkuri, Ville, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme e Pekka Abrahamsson: *ECCOLA - A method for implementing ethically aligned AI systems*. J. Syst. Softw., 182:111067, 2021. https://doi.org/10.1016/j.jss.2021.111067. 5, 11, 27, 29, 30, 58, 60

[41] Floridi, Luciano e Josh Cowls: *A unified framework of five principles for ai in society*. Machine learning and the city: Applications in architecture and urban design, páginas 535–545, 2022. 5, 11, 58, 60

[42] Cerqueira, José Antonio Siqueira de, Anayran Pinheiro De Azevedo, Heloise Acco Tives Leão e Edna Dias Canedo: *Guide for artificial intelligence ethical requirements elicitation - RE4AI ethical guide*. Em *55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022*, páginas 1–10. ScholarSpace, 2022. http://hdl.handle.net/10125/80015. 5, 27, 29, 30, 58, 60

[43] Carleton, Anita D, Erin Harper, Tim Menzies, Tao Xie, Sigrid Eldh e Michael R Lyu: *The ai effect: Working at the intersection of ai and se*. IEEE Software, 37(4):26–35, 2020. 5, 7, 32

[44] Halme, Erika, Mamia Agbese, Hanna-Kaisa Alanen, Jani Antikainen, Marianna Jantunen, Arif Ali Khan, Kai-Kristian Kemell, Ville Vakkuri e Pekka Abrahamsson: *Implementation of ethically aligned design with ethical user stories in SMART terminal digitalization project: Use case passenger flow*. CoRR, abs/2111.06116, 2021. https://arxiv.org/abs/2111.06116. 5, 7, 33, 34, 35, 36, 57

[45] Talukder, Md Simul Hasan e Ajay Krishno Sarkar: *Nutrients deficiency diagnosis of rice crop by weighted average ensemble learning*. Smart Agricultural Technology, 4:100155, 2023. 6, 8, 26, 50

[46] Xu, Baoxun, Xiufeng Guo, Yunming Ye e Jiefeng Cheng: *An improved random forest classifier for text categorization.* J. Comput., 7(12):2913–2920, 2012. 6, 8, 26, 50

[47] Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer e Veselin Stoyanov: *Unsupervised cross-lingual representation learning at scale.* CoRR, abs/1911.02116, 2019. `http://arxiv.org/abs/1911.02116`. 6, 8, 24, 46, 49, 50, 56

[48] Sanh, Victor, Lysandre Debut, Julien Chaumond e Thomas Wolf: *Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter.* CoRR, abs/1910.01108, 2019. `http://arxiv.org/abs/1910.01108`. 6, 8, 24, 46, 49, 50, 56

[49] Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth *et al.*: *Crisp-dm 1.0: Step-by-step data mining guide.* SPSS inc, 9(13):1–73, 2000. 6

[50] Gill, Milapji Singh, Tom Westermann, Marvin Schieseck e Alexander Fay: *Integration of domain expert-centric ontology design into the CRISP-DM for cyber-physical production systems.* CoRR, abs/2307.11637, 2023. `https://doi.org/10.48550/arXiv.2307.11637`. 6

[51] Acosta-Solano, Jairo, Diana Janeth Lancheros Cuesta, Samir F. Umaña Ibáñez e Jairo R. Coronado-Hernández: *Predictive models assessment based on CRISP-DM methodology for students performance in colombia - saber 11 test.* Em Varandas, Nuno, Ansar-Ul-Haque Yasar, Haroon Malik e Stéphane Galland (editores): *The 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2021) / The 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2021), Leuven, Belgium, November 1-4, 2021*, volume 198 de *Procedia Computer Science*, páginas 512–517. Elsevier, 2021. `https://doi.org/10.1016/j.procs.2021.12.278`. 7

[52] Pilgrim, Charlie, Adam Sanborn, Eugene Malthouse e Thomas T Hills: *Confirmation bias emerges from an approximation to bayesian reasoning.* Cognition, 245:105693, 2024. 7

[53] Gwet, Kilem Li: *Computing inter-rater reliability and its variance in the presence of high agreement.* British Journal of Mathematical and Statistical Psychology, 61(1):29–48, 2008. 7, 39

[54] Souza, Fábio, Rodrigo Frassetto Nogueira e Roberto de Alencar Lotufo: *Bertimbau: Pretrained BERT models for brazilian portuguese.* Em Cerri, Ricardo e Ronaldo C. Prati (editores): *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I*, volume 12319 de *Lecture Notes in Computer Science*, páginas 403–417. Springer, 2020. `https://doi.org/10.1007/978-3-030-61377-8_28`. 8

[55] Tegmark, Max: *Life 3.0: Being human in the age of artificial intelligence.* Vintage, 2018. 10

[56] Weizenbaum, Joseph: *ELIZA - a computer program for the study of natural language communication between man and machine*. Commun. ACM, 9(1):36–45, 1966. https://doi.org/10.1145/365153.365168. 10

[57] Weizenbaum, Joseph: *Computer power and human reason: From judgment to calculation.* 1976. 10

[58] Norbert, Wiener: *Human Use of Human Beings*. Doubleday & Company, 1954. 10

[59] Stahl, Bernd Carsten: *From computer ethics and the ethics of AI towards an ethics of digital ecosystems*. AI Ethics, 2(1):65–77, 2022. https://doi.org/10.1007/s43681-021-00080-1. 11

[60] Stahl, Bernd Carsten: *Concepts of Ethics and Their Application to AI*, páginas 19–33. Springer International Publishing, Cham, 2021, ISBN 978-3-030-69978-9. https://doi.org/10.1007/978-3-030-69978-9_3. 11

[61] Siau, Keng e Weiyu Wang: *Artificial intelligence (AI) ethics: Ethics of AI and ethical AI*. J. Database Manag., 31(2):74–87, 2020. https://doi.org/10.4018/JDM.2020040105. 11

[62] Hagendorff, Thilo: *The ethics of AI ethics: An evaluation of guidelines*. Minds Mach., 30(1):99–120, 2020. https://doi.org/10.1007/s11023-020-09517-8. 11

[63] Smit, Koen, Martijn Zoet e John van Meerten: *A review of AI principles in practice*. Em Vogel, Doug, Kathy Ning Shen, Pan Shan Ling, Carol Hsu, James Y. L. Thong, Marco De Marco, Moez Limayem e Sean Xin Xu (editores): *24th Pacific Asia Conference on Information Systems, PACIS 2020, Dubai, UAE, June 22-24, 2020*, página 198, 2020. https://aisel.aisnet.org/pacis2020/198. 11

[64] Agbese, Mamia, Hanna Kaisa Alanen, Jani Antikainen, Halme Erika, Hannakaisa Isomaki, Marianna Jantunen, Kai Kristian Kemell, Rebekah Rousi, Heidi Vainio-Pekka e Ville Vakkuri: *Governance in ethical and trustworthy ai systems: Extension of the eccola method for ai ethics governance using garp*. e-Informatica Software Engineering Journal, 17(1):230101, 2023. 11, 60

[65] Guizzardi, Renata S. S., Glenda Carla Moura Amaral, Giancarlo Guizzardi e John Mylopoulos: *Ethical requirements for AI systems*. Em Goutte, Cyril e Xiaodan Zhu (editores): *Advances in Artificial Intelligence - 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13-15, 2020, Proceedings*, volume 12109 de *Lecture Notes in Computer Science*, páginas 251–256. Springer, 2020. https://doi.org/10.1007/978-3-030-47358-7_24. 20

[66] Ahmad, Khlood, Muneera Bano, Mohamed Abdelrazek, Chetan Arora e John C. Grundy: *What's up with requirements engineering for artificial intelligence systems?* Em *29th IEEE International Requirements Engineering Conference, RE 2021, Notre Dame, IN, USA, September 20-24, 2021*, páginas 1–12. IEEE, 2021. https://doi.org/10.1109/RE51729.2021.00008. 20

[67] Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo e Dan Dennison: *Hidden technical debt in machine learning systems*. Em Cortes, Corinna, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama e Roman Garnett (editores): *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, páginas 2503–2511, 2015. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html. 20

[68] Belani, Hrvoje, Marin Vukovic e Zeljka Car: *Requirements engineering challenges in building ai-based complex systems*. Em *27th IEEE International Requirements Engineering Conference Workshops, RE 2019 Workshops, Jeju Island, Korea (South), September 23-27, 2019*, páginas 252–255. IEEE, 2019. https://doi.org/10.1109/REW.2019.00051. 20, 26

[69] Ahmad, Khlood, Mohamed Almorsy Abdelrazek, Chetan Arora, Muneera Bano e John C. Grundy: *Requirements engineering for artificial intelligence systems: A systematic mapping study*. Inf. Softw. Technol., 158:107176, 2023. https://doi.org/10.1016/j.infsof.2023.107176. 20

[70] Kuwajima, Hiroshi, Hirotoshi Yasuoka e Toshihiro Nakae: *Engineering problems in machine learning systems*. Mach. Learn., 109(5):1103–1126, 2020. https://doi.org/10.1007/s10994-020-05872-w. 20

[71] Vainio-Pekka, Heidi, Mamia Ori otse Agbese, Marianna Jantunen, Ville Vakkuri, Tommi Mikkonen, Rebekah Rousi e Pekka Abrahamsson: *The role of explainable ai in the research field of ai ethics*. ACM Transactions on Interactive Intelligent Systems, 2023. 20

[72] Paech, Barbara e Kurt Schneider: *How do users talk about software? searching for common ground*. Em *1st Workshop on Ethics in Requirements Engineering Research and Practice, REthics@RE 2020, Zurich, Switzerland, August 31, 2020*, páginas 11–14. IEEE, 2020. https://doi.org/10.1109/REthics51204.2020.00008. 20

[73] Guizzardi, Renata, Glenda Amaral, Giancarlo Guizzardi e John Mylopoulos: *An ontology-based approach to engineering ethicality requirements*. Software and Systems Modeling, páginas 1–27, 2023. 21

[74] Agbese, Mamia, Rahul Mohanani, Arif Ali Khan e Pekka Abrahamsson: *Ethical requirements stack: A framework for implementing ethical requirements of AI in software engineering practices*. Em *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, EASE 2023, Oulu, Finland, June 14-16, 2023*, páginas 326–328. ACM, 2023. https://doi.org/10.1145/3593434.3593489. 21

[75] Agbese, Mamia, Rahul Mohanani, Arif Ali Khan e Pekka Abrahamsson: *Implementing AI ethics: Making sense of the ethical requirements*. Em *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, EASE 2023, Oulu, Finland, June 14-16, 2023*, páginas 62–71. ACM, 2023. https://doi.org/10.1145/3593434.3593453. 21

[76] Kang, Yue, Zhao Cai, Chee Wee Tan, Qian Huang e Hefu Liu: *Natural language processing (nlp) in management research: A literature review.* Journal of Management Analytics, 7(2):139–172, 2020. 21

[77] Khurana, Diksha, Aditya Koli, Kiran Khatter e Sukhdev Singh: *Natural language processing: state of the art, current trends and challenges.* Multim. Tools Appl., 82(3):3713–3744, 2023. https://doi.org/10.1007/s11042-022-13428-4. 21

[78] Aljebreen, Mohammed, Bayan Ibrahimm Alabdullah, Mashael M. Asiri, Ahmed S. Salama, Mohammed Assiri e Sara Saadeldeen Ibrahim: *Moth flame optimization with hybrid deep learning based sentiment classification toward chatgpt on twitter.* IEEE Access, 11:104984–104991, 2023. https://doi.org/10.1109/ACCESS.2023.3315609. 21

[79] Li, Jing, Aixin Sun, Jianglei Han e Chenliang Li: *A survey on deep learning for named entity recognition : Extended abstract.* Em *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, páginas 3817–3818. IEEE, 2023. https://doi.org/10.1109/ICDE55515.2023.00335. 22

[80] Wu, Chien-Sheng, Steven C. H. Hoi, Richard Socher e Caiming Xiong: *TOD-BERT: pre-trained natural language understanding for task-oriented dialogue.* Em Webber, Bonnie, Trevor Cohn, Yulan He e Yang Liu (editores): *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, páginas 917–929. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.66. 22

[81] Ramos-Soto, Alejandro, Alberto Bugarín e Senén Barro: *On the role of linguistic descriptions of data in the building of natural language generation systems.* Fuzzy Sets Syst., 285:31–51, 2016. https://doi.org/10.1016/j.fss.2015.06.019. 22

[82] Song, Jongyoon, Sungwon Kim e Sungroh Yoon: *Alignart: Non-autoregressive neural machine translation by jointly learning to estimate alignment and translate.* Em Moens, Marie-Francine, Xuanjing Huang, Lucia Specia e Scott Wen-tau Yih (editores): *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, páginas 1–14. Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.1. 22

[83] Rajpurkar, Pranav, Robin Jia e Percy Liang: *Know what you don't know: Unanswerable questions for squad.* Em Gurevych, Iryna e Yusuke Miyao (editores): *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, páginas 784–789. Association for Computational Linguistics, 2018. https://aclanthology.org/P18-2124/. 22

[84] Ondov, Brian D., Kush Attal e Dina Demner-Fushman: *A survey of automated methods for biomedical text simplification.* J. Am. Medical Informatics Assoc., 29(11):1976–1988, 2022. https://doi.org/10.1093/jamia/ocac149. 22

[85] El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea e Hoda K. Mohamed: *Automatic text summarization: A comprehensive survey*. Expert Syst. Appl., 165:113679, 2021. https://doi.org/10.1016/j.eswa.2020.113679. 22

[86] Xu, Yan, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su e Pascale Fung: *Retrieval-free knowledge-grounded dialogue response generation with adapters*. Em Feng, Song, Hui Wan, Caixia Yuan e Han Yu (editores): *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, DialDoc@ACL 2022, Dublin, Ireland, May 26, 2022*, páginas 93–107. Association for Computational Linguistics, 2022. https://doi.org/10.18653/v1/2022.dialdoc-1.10. 22

[87] HaCohen-Kerner, Yaakov, Daniel Miller e Yair Yigal: *The influence of preprocessing on text classification using a bag-of-words representation*. PloS one, 15(5):e0232525, 2020. 22

[88] Gasparetto, Andrea, Matteo Marcuzzo, Alessandro Zangari e Andrea Albarelli: *A survey on text classification algorithms: From text to predictions*. Inf., 13(2):83, 2022. https://doi.org/10.3390/info13020083. 22

[89] Dalal, Mita K e Mukesh A Zaveri: *Automatic text classification: a technical review*. International Journal of Computer Applications, 28(2):37–40, 2011. 22

[90] Muhamedyev, Ravil: *Machine learning methods: An overview*. Computer modelling & new technologies, 19(6):14–29, 2015. 22

[91] López-Díaz, María Concepción, Miguel López-Díaz e Sergio Martínez-Fernández: *On the optimal binary classifier with an application*. Comput. Stat. Data Anal., 181:107683, 2023. https://doi.org/10.1016/j.csda.2022.107683. 23

[92] Raza, Muhammad, Farookh Khadeer Hussain, Omar Khadeer Hussain, Ming Zhao e Zia ur Rehman: *A comparative analysis of machine learning models for quality pillar assessment of saas services by multi-class text classification of users' reviews*. Future generation computer systems, 101:341–371, 2019. 23

[93] Zhang, Ximing, Qian-Wen Zhang, Zhao Yan, Ruifang Liu e Yunbo Cao: *Enhancing label correlation feedback in multi-label text classification via multi-task learning*. Em Zong, Chengqing, Fei Xia, Wenjie Li e Roberto Navigli (editores): *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 de *Findings of ACL*, páginas 1190–1200. Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.findings-acl.101. 23

[94] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser e Illia Polosukhin: *Attention is all you need*. Em Guyon, Isabelle, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan e Roman Garnett (editores): *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pá-

ginas 5998–6008, 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. 23

[95] Zhao, Yingjia e Xin Tao: *Zyj123@ dravidianlangtech-eacl2021: Offensive language identification based on xlm-roberta with dpcnn*. Em *Proceedings of the first workshop on speech and language technologies for dravidian languages*, páginas 216–221, 2021. 24, 46

[96] Kadhim, Ammar Ismael: *Survey on supervised machine learning techniques for automatic text classification*. Artif. Intell. Rev., 52(1):273–292, 2019. https://doi.org/10.1007/s10462-018-09677-1. 24

[97] Heydarian, Mohammadreza, Thomas E Doyle e Reza Samavi: *Mlcm: Multi-label confusion matrix*. IEEE Access, 10:19083–19095, 2022. 24, 25

[98] Han, Lu, QiXiang Zhou e Tong Li: *Improving requirements classification models based on explainable requirements concerns*. Em *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, páginas 95–101, 2023. 27, 30

[99] Jain, Chirag, Preethu Rose Anish e Smita Ghaisas: *Automated identification of security and privacy requirements from software engineering contracts*. Em Schneider, Kurt, Fabiano Dalpiaz e Jennifer Horkoff (editores): *31st IEEE International Requirements Engineering Conference, RE 2023 - Workshops, Hannover, Germany, September 4-5, 2023*, páginas 234–238. IEEE, 2023. https://doi.org/10.1109/REW57809.2023.00047. 28, 30

[100] Rossini, Daniele, Danilo Croce, Sara Mancini, Massimo Pellegrino e Roberto Basili: *Actionable ethics through neural learning*. Em *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, páginas 5537–5544. AAAI Press, 2020. https://doi.org/10.1609/aaai.v34i04.6005. 28, 29, 30

[101] Cleland-Huang, Jane, Raffaella Settimi, Xuchang Zou e Peter Solc: *Automated classification of non-functional requirements*. Requir. Eng., 12(2):103–120, 2007. https://doi.org/10.1007/s00766-007-0045-1. 29, 30, 58

[102] Glinz, Martin: *On non-functional requirements*. Em *15th IEEE International Requirements Engineering Conference, RE 2007, October 15-19th, 2007, New Delhi, India*, páginas 21–26. IEEE Computer Society, 2007. https://doi.org/10.1109/RE.2007.45. 32

[103] Vakkuri, Ville, Kai-Kristian Kemell e Pekka Abrahamsson: *Ethically aligned design: An empirical evaluation of the resolvedd-strategy in software and systems development context*. Em Staron, Miroslaw, Rafael Capilla e Amund Skavhaug (editores): *45th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2019, Kallithea-Chalkidiki, Greece, August 28-30, 2019*, páginas 46–50. IEEE, 2019. https://doi.org/10.1109/SEAA.2019.00015. 33, 58

[104] Halme, Erika, Marianna Jantunen, Ville Vakkuri, Kai Kristian Kemell e Pekka Abrahamsson: *Making ethics practical: User stories as a way of implementing ethical consideration in software engineering*. Information and Software Technology, 167:107379, 2024, ISSN 0950-5849. https://www.sciencedirect.com/science/article/pii/S0950584923002343. 33

[105] Liu, Bing, Xiaoli Li, Wee Sun Lee e Philip S. Yu: *Text classification by labeling words*. Em McGuinness, Deborah L. e George Ferguson (editores): *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*, páginas 425–430. AAAI Press / The MIT Press, 2004. http://www.aaai.org/Library/AAAI/2004/aaai04-068.php. 33

[106] Desmond, Michael, Michael J. Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi e Qian Pan: *Increasing the speed and accuracy of data labeling through an AI assisted interface*. Em Hammond, Tracy, Katrien Verbert, Dennis Parra, Bart P. Knijnenburg, John O'Donovan e Paul Teale (editores): *IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021*, páginas 392–401. ACM, 2021. https://doi.org/10.1145/3397481.3450698. 37

[107] Ohyama, Tetsuji: *Statistical inference of gwet's ac1 coefficient for multiple raters and binary outcomes*. Communications in Statistics-Theory and Methods, 50(15):3564–3572, 2021. 39

[108] Landis, JR: *The measurement of observer agreement for categorical data*. Biometrics, 1977. 39

[109] Vach, Werner e Oke Gerke: *Gwet's ac1 is not a substitute for cohen's kappa–a comparison of basic properties*. MethodsX, 10:102212, 2023. 39

[110] Wongpakaran, Nahathai, Tinakon Wongpakaran, Danny Wedding e Kilem L Gwet: *A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples*. BMC medical research methodology, 13:1–7, 2013. 39

[111] Cole, Rosanna: *Inter-rater reliability methods in qualitative case study research*. Sociological Methods & Research, 53(4):1944–1975, 2024. 39

[112] Hickman, Louis, Stuti Thapa, Louis Tay, Mengyang Cao e Padmini Srinivasan: *Text preprocessing for text mining in organizational research: Review and recommendations*. Organizational Research Methods, 25(1):114–146, 2022. 42

[113] Tabassum, Ayisha e Rajendra R Patil: *A survey on text pre-processing & feature extraction techniques in natural language processing*. International Research Journal of Engineering and Technology (IRJET), 7(06):4864–4867, 2020. 42, 43

[114] Zhao, Jianqiang e Xiaolin Gui: *Comparison research on text pre-processing methods on twitter sentiment analysis*. IEEE Access, 5:2870–2879, 2017. https://doi.org/10.1109/ACCESS.2017.2672677. 42

[115] Adhikari, Ashutosh, Achyudh Ram, Raphael Tang e Jimmy Lin: *Docbert: BERT for document classification*. CoRR, abs/1904.08398, 2019. http://arxiv.org/abs/1904.08398. 46

[116] Büyüköz, Berfu, Ali Hürriyetoğlu e Arzucan Özgür: *Analyzing elmo and distilbert on socio-political news classification*. Em *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, páginas 9–18, 2020. 46

[117] Rajapaksha, Praboda, Reza Farahbakhsh e Noël Crespi: *Bert, xlnet or roberta: The best transfer learning model to detect clickbaits*. IEEE Access, 9:154704–154716, 2021. https://doi.org/10.1109/ACCESS.2021.3128742. 46

[118] Jiang, Xiaoze, Yaobo Liang, Weizhu Chen e Nan Duan: *XLM-K: improving cross-lingual language model pre-training with multilingual knowledge*. CoRR, abs/2109.12573, 2021. https://arxiv.org/abs/2109.12573. 46

[119] Jabbar, H e Rafiqul Zaman Khan: *Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)*. Computer Science, Communication and Instrumentation Devices, 70(10.3850):978–981, 2015. 47

[120] Ying, Xue: *An overview of overfitting and its solutions*. Em *Journal of physics: Conference series*, volume 1168, página 022022. IOP Publishing, 2019. 47

[121] Faraj, Dalya e Malak Abdullah: *Sarcasmdet at semeval-2021 task 7: Detect humor and offensive based on demographic factors using roberta pre-trained model*. Em Palmer, Alexis, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurélie Herbelot e Xiaodan Zhu (editores): *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, páginas 527–533. Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.semeval-1.64. 47

[122] Kohavi, Ron: *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Em *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, páginas 1137–1145. Morgan Kaufmann, 1995. http://ijcai.org/Proceedings/95-2/Papers/016.pdf. 47

[123] Cerqueira, José Antonio Siqueira de, Lucas Dos Santos Althoff, Paulo Santos De Almeida e Edna Dias Canedo: *Ethical perspectives in AI: A two-folded exploratory study from literature and active development projects*. Em *54th Hawaii International Conference on System Sciences, HICSS 2021, Kauai, Hawaii, USA, January 5, 2021*, páginas 1–10. ScholarSpace, 2021. https://hdl.handle.net/10125/71257. 58, 60

[124] Di Salvo, Cristina: *Improving results of existing groundwater numerical models using machine learning techniques: A review*. Water, 14(15):2307, 2022. 60

[125] Khan, Arif Ali, Muhammad Azeem Akbar, Mahdi Fahmideh, Peng Liang, Muhammad Waseem, Aakash Ahmad, Mahmood Niazi e Pekka Abrahamsson: *AI ethics: An empirical study on the views of practitioners and lawmakers*. IEEE Trans. Comput. Soc. Syst., 10(6):2971–2984, 2023. https://doi.org/10.1109/TCSS.2023.3251729. 60

[126] Takase, Tomoumi, Satoshi Oyama e Masahito Kurihara: *Evaluation of stratified validation in neural network training with imbalanced data*. Em *IEEE International Conference on Big Data and Smart Computing, BigComp 2019, Kyoto, Japan, February 27 - March 2, 2019*, páginas 1–4. IEEE, 2019. https://doi.org/10.1109/BIGCOMP.2019.8678924. 62

[127] Dube, Lindani e Tanja Verster: *Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models*. Data Science in Finance and Economics, 3(4):354–379, 2023. 62

[128] Szeghalmy, Szilvia e Attila Fazekas: *A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning*. Sensors, 23(4):2333, 2023. https://doi.org/10.3390/s23042333. 62