Introdução

Com a finalidade de avaliar o filme Colossus optou-se por utilizar duas réguas de análise. A primeira será uma régua técnica e para isto foi utilizado o artigo Explainable AI: A Review of Machine Learning Interpretability Methods e a palestra

Explicabilidade em Machine Learning: isso existe? Ministrada por Bianca Ximenes.

Principais conceitos

Os termos interpretabilidade e explicabilidade são geralmente usados ​​pelos pesquisadores de forma intercambiável; no entanto, embora esses termos estejam intimamente relacionados, alguns trabalhos identificam

suas diferenças e distinguir esses dois conceitos. Não há uma matemática concreta

definição de interpretabilidade ou explicabilidade, nem foram medidos por alguns

métrica; no entanto, várias tentativas foram feitas [14-16] para esclarecer não

apenas esses dois termos, mas também conceitos relacionados, como compreensibilidade. No entanto, todos

essas definições carecem de formalidade matemática e rigor [17]. Um dos mais

definições populares de interpretabilidade é a de Doshi-Velez e Kim, que, em seu

trabalho [15], definem-no como “a capacidade de explicar ou apresentar em termos compreensíveis a um

humano". Outra definição popular veio de Miller em seu trabalho [18], onde ele define

interpretabilidade como “o grau em que um ser humano pode entender a causa de uma decisão”.

Embora intuitivas, essas definições carecem de formalidade matemática e rigor.

Com base no exposto, a interpretabilidade está principalmente conectada com a intuição por trás das saídas de um modelo [17]; com a ideia de que quanto mais interpretável for um sistema de aprendizado de máquina, mais fácil será identificar relações de causa e efeito nas entradas e saídas do sistema. Por exemplo, em tarefas de reconhecimento de imagem, parte do motivo que levou um sistema a decidir que um objeto específico faz parte de uma imagem (saída) pode ser determinados padrões dominantes na imagem (entrada). A explicabilidade, por outro lado, está associada

com a lógica interna e mecânica que estão dentro de um sistema de aprendizado de máquina. Quanto mais explicável for um modelo, mais profundo será o entendimento que os humanos alcançam em termos dos procedimentos internos que ocorrem enquanto o modelo está treinando ou tomando decisões.

Doshi-Velez and Kim propôs a seguinte classificação de métodos de avaliação

para interpretabilidade: fundamentada em aplicativos, fundamentada em humanos e fundamentada funcionalmente, posteriormente discutindo os possíveis trade-offs entre eles.

A avaliação baseada na aplicação preocupa-se com a forma como os resultados do processo de interpretação afetam o ser humano,

especialista de domínio, usuário final em termos de uma tarefa ou aplicação específica e bem definida. Exemplos concretos sob este tipo de avaliação incluem se um método de interpretabilidade resulta em melhor identificação de erros ou menos discriminação

Avaliação fundamentada no ser humano é semelhante à avaliação baseada em aplicativos; no entanto, existem duas diferenças principais:

primeiro, o testador neste caso não precisa ser um especialista de domínio, mas pode ser qualquer usuário final humano e, segundo, o objetivo final não é avaliar uma interpretação produzida em relação para sua adequação a uma aplicação específica, mas sim para testar a qualidade da interpretação produzida em um ambiente mais geral e medir o quão bem as noções gerais são capturadas.

A avaliação funcionalmente fundamentada não requer experimentos que envolvam seres humanos, mas sim usa definições matemáticas formais e bem definidas de interpretabilidade para avaliar a qualidade de um método de interpretabilidade. Este tipo de avaliação geralmente segue os outros dois tipos de avaliação: uma vez que uma classe de modelos já passou por alguns critérios de interpretabilidade via experimentos fundamentados em humanos ou baseados em aplicativos, então as definições matemáticas podem ser usado para classificar ainda mais a qualidade dos modelos de interpretabilidade. funcionalmente fundamentado avaliação também é apropriada quando experimentos que envolvem seres humanos não podem ser aplicados por algum motivo (por exemplo, considerações éticas) ou quando o método proposto não atingiu um estágio maduro o suficiente para ser avaliado por usuários humanos. Dito isto, determinar o direito critérios de medição e métrica para cada caso é desafiador e continua sendo um problema em aberto.

[Abaixo alguns principios eticos guias que serão preenchidos com comentários acerca do filme utilizando conceitos de interpretabilidade e explicabilidade]

1. PRINCÍPIOS ÉTICOS GERAIS.

Um profissional de computação deve...

1.1 Contribuir para a sociedade e para o bem-estar humano,

reconhecendo que todas as pessoas são partes interessadas

em computação.

1.2 Evite danos.

1.3 Seja honesto e confiável.

1.4 Seja justo e tome medidas para não discriminar.

1.5 Respeitar o trabalho necessário para produzir novas ideias,

invenções, trabalhos criativos e artefatos de computação.

1.6 Respeite a privacidade.

1.7 Honrar a confidencialidade.

2. RESPONSABILIDADES PROFISSIONAIS.

Um profissional de computação deve...

2.1 Esforçar-se para alcançar alta qualidade em ambos os processos

e produtos do trabalho profissional.

2.2 Manter altos padrões de profissionalismo

competência, conduta e prática ética.

2.3 Conheça e respeite as regras existentes relativas à

trabalho profissional.

2.4 Aceitar e fornecer avaliação profissional apropriada.

2.5 Fornecer avaliações abrangentes e completas de

sistemas de computador e seus impactos, incluindo

análise de possíveis riscos.

2.6 Realizar trabalhos apenas nas áreas de competência.

2.7 Promover a conscientização e compreensão do público sobre

computação, tecnologias relacionadas e seus

consequências.

2.8 Acessar recursos de computação e comunicação

somente quando autorizado ou quando compelido pelo

bem público.

2.9 Projetar e implementar sistemas que sejam robustos

e utilizável seguro.

3. PRINCÍPIOS DE LIDERANÇA PROFISSIONAL.

Um profissional de computação, especialmente aquele que atua como líder,

deve...

3.1 Garantir que o bem público seja a preocupação central

durante todo o trabalho de computação profissional.

3.2 Articular, encorajar a aceitação e avaliar

cumprimento das responsabilidades sociais pelos membros da

a organização ou grupo.

3.3 Gerenciar pessoal e recursos para melhorar o

qualidade de vida no trabalho.

3.4 Articular, aplicar e apoiar políticas e

processos que refletem os princípios do Código.

3.5 Criar oportunidades para os membros da

organização ou grupo para crescer como profissionais.

3.6 Tenha cuidado ao modificar ou retirar sistemas.

3.7 Reconhecer e tomar cuidado especial com os sistemas que

tornar-se integrado na infra-estrutura da sociedade

4. CONFORMIDADE COM O CÓDIGO.

Um profissional de computação deve...

4.1 Defender, promover e respeitar os princípios

do Código.

4.2 Trate as violações do Código como incompatíveis com

filiação à ACM.