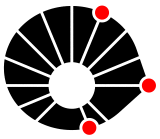

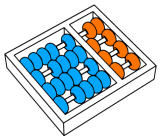

Estratégias de testes de imparcialidade para sistemas de recomendação

I. G. Fagioli *E. Martins*

Relatório Técnico - IC-PFG-24-34
Projeto Final de Graduação
2024 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Estratégias de testes de imparcialidade para sistemas de recomendação

Isabella Garcia Fagioli* Eliane Martins†

Resumo

Este trabalho estuda imparcialidade em sistemas de recomendação, que têm se tornado cada vez mais comuns em plataformas digitais para recomendação de filmes, séries, músicas, livros, produtos, entre outros. Tais sistemas, embora eficientes na personalização da experiência do usuário, podem incorrer em vieses que afetam a parcialidade das recomendações, muitas vezes prejudicando grupos específicos ou favorecendo outros de forma não intencional baseado em atributos sensíveis.

Para avaliar a imparcialidade desses sistemas, são necessários testes e métricas específicas, o que não é tão simples quanto os testes de acurácia, pois é difícil estabelecer o que se espera em termos de imparcialidade e ainda mais desafiador medi-la de forma precisa.

Entretanto, existe uma vasta gama de ferramentas disponíveis para realizar tais testes, e a escolha da mais adequada pode ser um desafio. Este trabalho tem como objetivo entender os principais fatores que devem ser considerados ao selecionar uma ferramenta para testar a imparcialidade em sistemas de recomendação.

1 Introdução

1.1 Contexto e Motivação

Conforme o uso de sistemas de recomendação se torna cada vez mais presente em nossas vidas - seja por meio de sugestões de filmes e séries nas plataformas de streaming, de produtos no e-commerce ou de conteúdo nas redes sociais - surge uma crescente preocupação com a imparcialidade desses sistemas. Eles têm um impacto significativo em nossas escolhas diárias e influenciam diretamente nossas decisões de consumo. Como esses sistemas determinam o que vemos, compramos ou consumimos, falhas em sua construção ou em suas decisões podem não apenas afetar a experiência do usuário, mas também perpetuar estereótipos, preconceitos e discriminação. Se, por exemplo, os algoritmos falharem em promover uma diversidade de opções ou favorecer um determinado grupo em detrimento de outro, isso pode criar um ciclo de exclusão ou de reforço de preconceitos pre-existentes. Essa preocupação com a imparcialidade é crucial para garantir que os sistemas de recomendação sejam justos, equitativos e não prejudiquem grupos minoritários ou marginalizados.

1.2 Objetivos

Este trabalho tem como objetivo estudar os fundamentos dos sistemas de recomendação e dos testes de imparcialidade, com o intuito de compreender suas finalidades e identificar a melhor forma de selecionar a ferramenta mais adequada para realizá-los. Através deste estudo, buscamos explorar os

*Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP, 13083-852
†Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP, 13083-852

requisitos necessários para escolher uma ferramenta eficaz, além de analisar os principais aspectos que podem influenciar essa escolha. Dessa forma, esperamos fornecer uma base sólida para a seleção de ferramentas que possam garantir a justiça e equidade nos sistemas de recomendação, minimizando vieses e promovendo uma experiência mais inclusiva para todos os usuários.

1.3 Metodologia

A metodologia utilizada para levantar os aspectos relevantes para testes de imparcialidade em sistemas de recomendação constou dos seguintes passos:

1. Estudos sobre sistemas de recomendação, o que são, para que servem, quais os tipos e as diferenças entre eles;
2. Levantamento dos desafios existentes para testes de imparcialidade em geral, e em sistemas de recomendação em especial.
3. Estudos sobre testes de imparcialidade em sistemas baseados em Aprendizado de Máquina.
4. Discussão sobre as técnicas estudadas, com vistas a delinear alguns aspectos importantes para testar imparcialidade de sistemas de recomendação.

1.4 Estrutura do texto

O texto está organizado da seguinte forma: na Seção 2 são apresentados os estudos sobre Sistemas de Recomendação. A seção 3 trata dos estudos sobre Testes de Imparcialidade em sistemas baseados em Machine Learning, mostrando as diferentes definições, os componentes a serem testados e os passos a serem realizados. A Seção 4 apresenta algumas ferramentas existentes e propõe alguns aspectos a serem considerados na escolha de ferramentas para testar sistemas de recomendação.

2 Sobre Sistemas de Recomendação

Inicialmente para este trabalho devemos entender o funcionamento de sistemas de recomendação. Para isso nos baseamos na leitura de "A systematic review and research perspective on recommender systems" [2].

2.1 O que são Sistemas de Recomendação

Sistemas de recomendação são sistemas que utilizam técnicas de Machine Learning para sugerir itens, como produtos, serviços ou conteúdos, a usuários com base em suas preferências, comportamentos anteriores ou características semelhantes a outros usuários.

Os sistemas de recomendação possuem dois componentes principais: usuários e itens. Os usuários são aqueles que interagem com o sistema, e suas escolhas e preferências são levadas em consideração para gerar as recomendações. Já os itens são os conteúdos sugeridos ao usuário, como filmes, músicas ou produtos. A interação entre usuários e itens, assim como as características de ambos, é utilizada para personalizar as recomendações e tornar a experiência mais relevante para cada usuário.

Eles são amplamente utilizados em plataformas de e-commerce, streaming de música e vídeo, redes sociais e muitos outros serviços online.

2.2 Desafios

Os sistemas de recomendação apresentam alguns desafios nas sua implementação e utilização, entre eles:

- Cold start: O sistema não pode prever propriamente para novos usuários por falta de dados
- Ataques Shilling: usuários mal-intencionados (ou bots) manipulam deliberadamente o sistema para influenciar as recomendações a seu favor
- Latência: Sistema não consegue recomendar itens novos
- Esparsidade: a maioria dos usuários interage com apenas uma pequena fração dos itens disponíveis
- Problema "gray sheep": Um usuário tem gostos e preferências que não se alinham claramente com nenhum grupo de usuários ou comunidades específicas
- Escalabilidade: Muitos dados entre usuários e itens tornam escalabilidade um desafio em sistemas de recomendação
- Usuário X Item: O sistema de recomendação deve ser imparcial tanto para os usuários quanto para os itens

2.3 Tipos de Sistemas de Recomendação

Sistemas de recomendação podem ser classificados em três tipos principais: baseado em conteúdo, colaborativo e híbrido, como mostra a Figura 1. Cada um desses tipos utiliza abordagens distintas para sugerir itens aos usuários, variando desde a análise das características dos próprios itens até a consideração das interações e preferências de outros usuários.

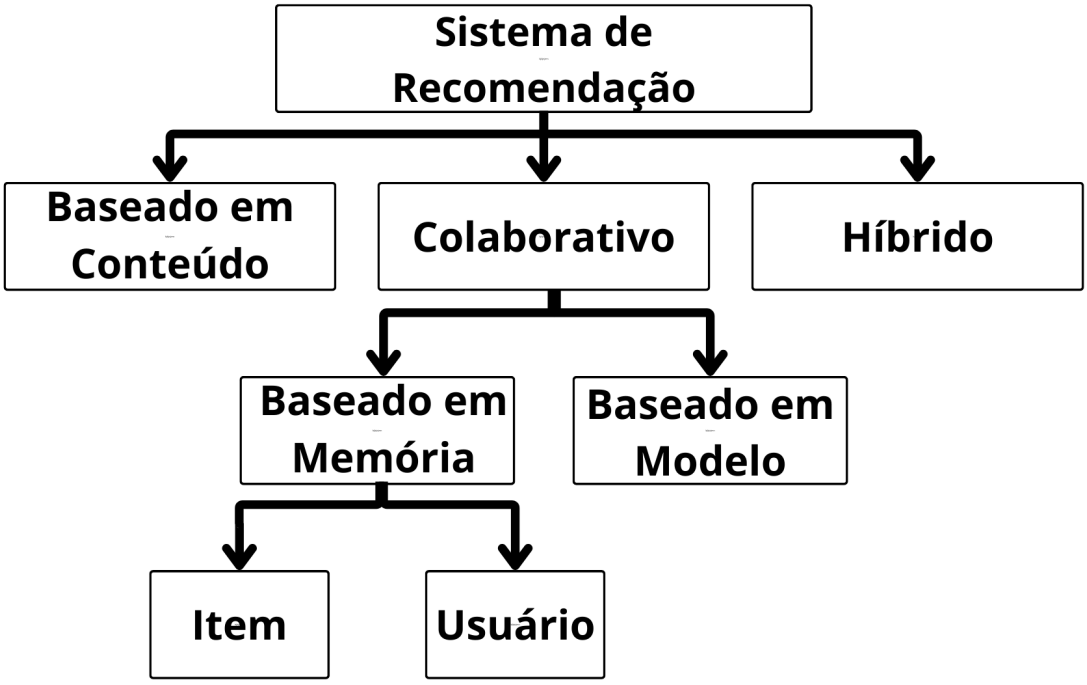

Figura 1: Tipos de sistemas de recomendação [2]

2.3.1 Baseado em conteúdo

As recomendações são geradas com base nas características dos itens e no histórico de preferências do usuário. Por exemplo, se um usuário demonstra interesse por filmes de ação, o sistema sugerirá outros filmes desse gênero, mesmo que esses filmes ainda não tenham sido preferidos por outros usuários. A Figura 2 mostra uma visão esquemática desta técnica. O objetivo é oferecer sugestões personalizadas, adaptadas ao comportamento individual, sem depender exclusivamente das interações de outros usuários.

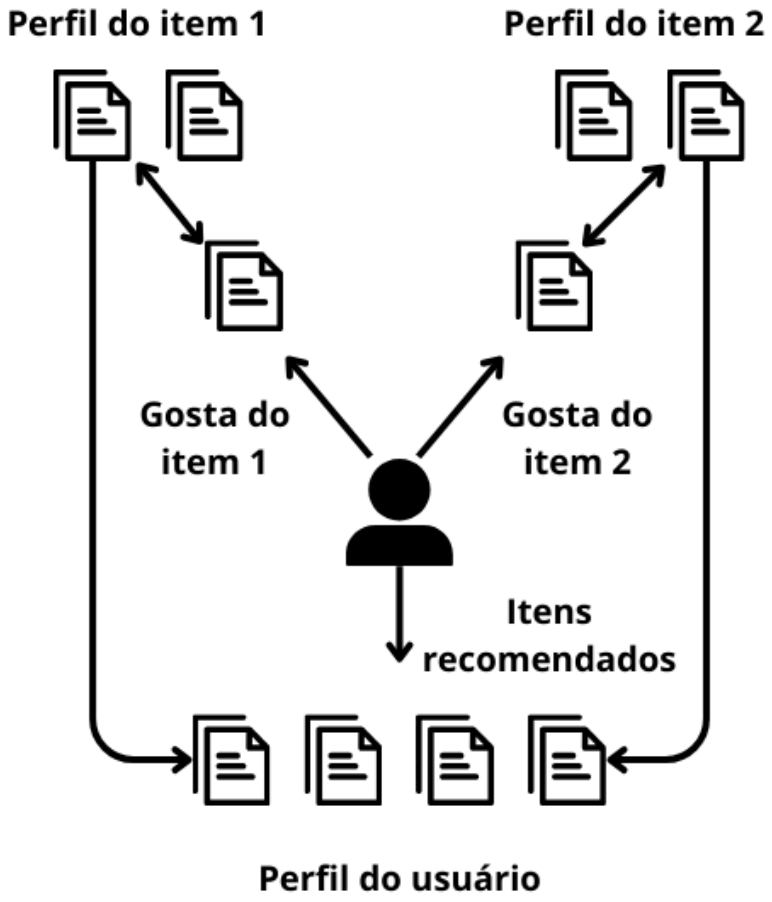

Figura 2: Sistema de recomendação baseado em conteúdo [2]

2.3.2 Colaborativo

Faz recomendações com base no comportamento de usuários semelhantes, conforme ilustra a Figura 3. Por exemplo, se o usuário A gostou de um determinado filme, e o usuário B tem gostos similares, o sistema pode sugerir esse filme ao usuário B. Este pode ser dividido em:

- Baseado em memória: Armazena e utiliza diretamente as interações passadas dos usuários para gerar recomendações. Pode ser baseado no item ou no usuário.

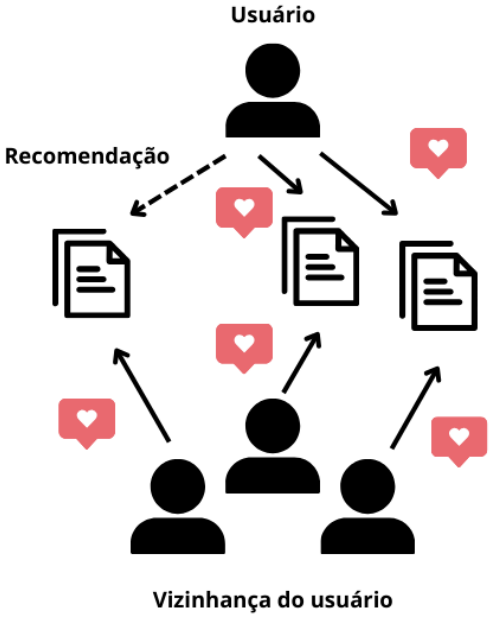

Figura 3: Sistema de recomendação colaborativo [2]

- Baseado no modelo: Usa algoritmos de aprendizado de máquina para aprender padrões a partir dos dados e criar um modelo preditivo.

2.3.3 Híbrido

Por último, o híbrido combina múltiplos métodos de recomendação, como filtragem colaborativa e baseada em conteúdo, para oferecer sugestões mais precisas e personalizadas.

2.4 Aplicações

Os sistemas de recomendação têm se tornado parte essencial de nossas vidas, moldando a forma como consumimos conteúdo e fazemos escolhas no dia a dia. Seja na Netflix ou em outras plataformas de streaming, eles sugerem filmes e séries que combinam com nossos gostos. No e-commerce, ajudam a personalizar a experiência de compra, indicando produtos que atendem às nossas preferências ou necessidades. Nas redes sociais, esses sistemas são responsáveis por selcionar e recomendar conteúdos que mantêm nosso interesse e engajamento. Em todas essas áreas, os algoritmos estão cada vez mais precisos, conectando-nos de maneira intuitiva ao que realmente importa ou desperta nossa curiosidade.

3 Testes de Imparcialidade

Nesta seção, abordaremos o conceito de imparcialidade, suas definições e como testá-la, apresentando os passos e os componentes. O conteúdo apresentado nesta seção se baseia majoritariamente em [4], a menos que seja explicitamente indicada outra referência.

3.1 Definições de imparcialidade

Com o aumento do uso e da quantidade de dados de um sistemas de recomendação, não basta garantir sua qualidade em termos de utilidade, algumas outras características também se fazem necessárias de serem analisadas, como diversidade, privacidade, etc. Uma dessas é a imparcialidade, que refere-se à distribuição justa e equitativa de recursos, como a exposição de conteúdos e a qualidade das recomendações, a diferentes indivíduos (imparcialidade individual) ou grupos (imparcialidade de grupo), especialmente quando envolvemos atributos sensíveis como sexo, raça, idade. Sem essa preocupação, temos podemos ter resultados que prejudicam minorias e trazem problemas éticos.

No caso específico dos sistemas de recomendação, temos um agravante quanto a imparcialidade com relação a outros usos de machine learning, pois ela precisa ser aplicada ao mesmo tempo aos usuários e aos itens.

Dentro da imparcialidade individual temos algumas diferentes definições: a imparcialidade pelo desconhecimento, no qual os atributos sensíveis são retirados da tomada de decisão; a imparcialidade pelo conhecimento em que indivíduos similares devem receber saídas similares; imparcialidade contrafactual que implica que o resultado para um indivíduo deve ser o mesmo tanto no mundo real quanto em um mundo contrafactual em que a pessoa pertence a outro grupo demográfico; por último, imparcialidade causal que explora as relações entre os atributos sensíveis e a saída, mudando-os e analisando como isso influencia o resultado.

Para a imparcialidade por grupo, também temos algumas diferentes definições: paridade estatística, em que a probabilidade de um resultado favorável deve ser a mesma para diferentes grupos; igualdade de oportunidades, em que os grupos privilegiados e os desprivilegiados devem ter uma taxa igual de falsos negativos e positivos entre si; oportunidade igual em que os grupos privilegiados e desprivilegiados devem ter a mesma taxa de positivos verdadeiros.

3.2 Definições sobre defeito e teste de imparcialidade

Defeito de imparcialidade diz respeito a uma condição de imparcialidade que está discordante da condição ideal. Essa condição depende da definição de imparcialidade adotada.

Existem vários defeitos de imparcialidade bem conhecidos em sistemas de recomendação, entre os quais podemos destacar o efeito Matthew e a câmara de eco. O efeito Matthew descreve a tendência de itens mais populares ganharem cada vez mais visibilidade, criando um ciclo em que os itens populares se tornam ainda mais destacados, enquanto itens menos conhecidos recebem menos atenção. Isso limita a descoberta de novos conteúdos pelos usuários, reforçando a popularidade daqueles que já têm uma base estabelecida de visibilidade.

De forma semelhante, o conceito de câmara de eco refere-se à exposição dos usuários a conteúdos que apenas reforçam seus gostos e crenças existentes. Em vez de serem desafiados ou expostos a novas ideias, os usuários permanecem em uma "bolha", onde suas preferências são constantemente alimentadas, o que impede a descoberta de novas perspectivas ou conteúdos fora de seu escopo habitual.

Assim, o teste de imparcialidade visa identificar esses defeitos de imparcialidade (não só os mencionados, mas também qualquer um que esteja diferente da condição ideal de imparcialidade definida) a fim de corrigi-los, de forma que o sistema de recomendação seja justo e igualitário para com todos os itens e usuários.

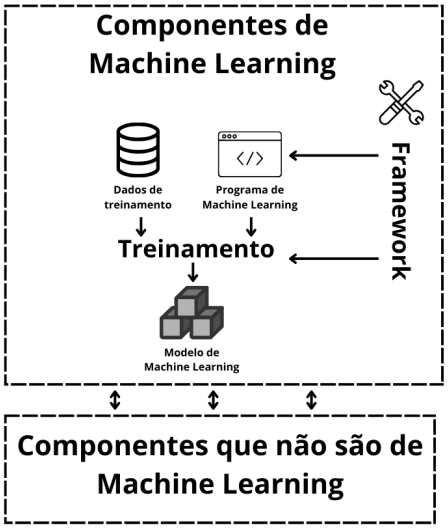

Figura 4: Componentes de teste [4]

3.3 Componentes de testes

Um sistema de recomendação, assim como qualquer software de machine learning, é composto por diversos componentes, como os dados, o modelo, o programa, o framework, entre outros. Para testar a imparcialidade de um sistema, é necessário avaliar diversos desses componentes, pois cada um pode interferir no resultado de maneiras diferentes. Abaixo estão os principais componentes e a forma como podem afetar a imparcialidade:

Dados: O teste de imparcialidade nos dados busca identificar vieses que possam ser replicados pelo conjunto de dados. Esse é um dos maiores causadores de defeitos de imparcialidade em sistemas de machine learning. Existem três tipos principais de viés:

- Viés na feature: Quando a própria característica (feature) é enviesada.
- Viés no rótulo: Quando fatores não relacionados influenciam a criação do rótulo.
- Viés na seleção: Quando o processo de seleção de dados leva à identificação de relações inesperadas entre atributos sensíveis e o rótulo.

Programa: O programa de machine learning pode apresentar defeitos de imparcialidade no processamento incorreto dos dados, na escolha do algoritmo de treinamento ou na definição dos hiperparâmetros. Cada uma dessas etapas pode introduzir viés, afetando a equidade do sistema.

Framework: O teste de imparcialidade no framework busca identificar problemas que possam levar a resultados parciais. No entanto, até agora, a maior parte das falhas identificadas nos frameworks está relacionada a outras métricas, como acurácia, e não especificamente à imparcialidade.

Modelo: A maior parte dos testes de fairness foca diretamente no modelo, visando identificar viéses a partir das entradas e saídas deste. Isso inclui avaliar como o modelo toma decisões com

base nas características dos dados e garantir que essas decisões não sejam influenciadas de forma injusta por fatores sensíveis.

Componentes não diretamente relacionados a machine learning: Alguns componentes do sistema que não estão diretamente ligados ao machine learning também podem introduzir viés. Por exemplo, o armazenamento de dados que resulta na exclusão de informações de grupos desfavorecidos, ou uma interface de usuário (UI) enviesada que favoreça um grupo específico. Testes devem ser realizados para identificar falhas nesses componentes e corrigir potenciais fontes de imparcialidade.

3.4 Como testar a imparcialidade

Para testar a imparcialidade, nós temos duas principais tarefas: geração de entrada do teste e definição de um oráculo.

3.4.1 Geração da entrada de teste

Em um teste de imparcialidade, a geração das entradas tem como objetivo propor dados que possam induzir viés ou discriminação, com a finalidade de revelar defeitos de imparcialidade no sistema. Existem diversas abordagens para gerar essas entradas, incluindo:

- Aleatória: Geração de entradas aleatórias para verificar viés.
- Baseada em busca: Criação de entradas com base em padrões específicos nos dados.
- Baseada em validação: Geração de entradas para testar se o modelo atende a critérios de imparcialidade.
- Geração específica do domínio: Entradas criadas com base nas características do domínio para testar possíveis vieses em contextos específicos.

3.4.2 Identificação do oráculo de teste

O oráculo de teste refere-se à determinação de se os resultados de um sistema estão dentro do esperado em termos de imparcialidade. Ele verifica se os resultados atendem aos requisitos de imparcialidade ou se apresentam defeitos. Até agora, existem duas formas principais de identificar um oráculo:

- Relações Metamórficas: Refere-se às relações entre uma alteração na entrada e a mudança correspondente na saída. No contexto de imparcialidade, podemos modificar um atributo sensível e observar como essa alteração afeta o resultado, ajudando a identificar possíveis vieses.
- Medições Estatísticas: Consiste em utilizar diferentes métricas estatísticas alinhadas com a definição de imparcialidade estabelecida, permitindo uma avaliação quantitativa da imparcialidade. No entanto, é desafiador determinar quais seriam as métricas esperadas e o quanto os resultados podem variar em relação a essas métricas.

4 Ferramentas de Teste de Imparcialidade

A automatização é um aspecto importante nos testes de software nos dias de hoje, pois devido aos prazos exíguos e a complexidade cada vez maior dos sistemas, além da frequência com que os testes devem ser realizados, o auxílio de ferramentas é fundamental para a equipe de testes ser capaz de

avaliar adequadamente a qualidade dos sistemas. Algumas ferramentas para Testes de Imparcialidade de sistemas foram propostas, a maioria delas para sistemas baseados em Machine Learning em geral, e não especificamente para Sistemas de Recomendação. Foi feito um levantamento de diferentes ferramentas voltadas para este tipo de testes, discutindo algumas vantagens e limitações das mesmas. Os autores analisam a usabilidade de tais ferramentas pelas equipes de teste e definem alguns fatores importantes de serem considerados na seleção da ferramenta, como: documentação, facilidade de instalação e de uso, versatilidade e escalabilidade [6]. Baseado em nossos estudos e experiências, também elencamos outros fatores como compatibilidade com sistemas de recomendação, suporte e personalização.

4.1 Critérios para escolha de uma ferramenta

4.1.1 Documentação

Para qualquer ferramenta, especialmente no contexto do trabalho diário de uma equipe de testes, é fundamental contar com uma documentação clara e bem estruturada, que inclua instruções simples de instalação e orientações práticas de uso. Isso facilita a adoção da ferramenta, reduz o tempo de treinamento e garante que todos os membros da equipe possam utilizá-la de forma eficiente e sem dificuldades.

4.1.2 Facilidade de instalação

Algumas ferramentas podem apresentar dificuldades de instalação, seja pela falta de instruções claras, suporte limitado a determinadas linguagens ou bibliotecas, conflitos de dependências, entre outros fatores.

A questão pode ir além da simples dificuldade de instalação, pois bibliotecas desatualizadas ou conflitos de dependências podem tornar o problema ainda mais complexo, chegando a ponto de ser impossível resolvê-los sem realizar mudanças significativas no código da ferramenta.

Esses obstáculos podem tornar a adoção da ferramenta inviável ou extremamente dispendiosa, comprometendo sua eficácia e o tempo da equipe.

4.1.3 Facilidade de uso

Além da facilidade de instalação, a facilidade de uso (user-friendliness) é um fator crucial. A ferramenta deve ser intuitiva, fácil de aprender e simples de manusear, com uma interface clara e processos diretos. Esses aspectos influenciam diretamente na curva de aprendizado, determinando quanto tempo a equipe precisará investir para começar a utilizá-la de maneira eficiente. Quanto mais acessível e amigável for a ferramenta, menos tempo será necessário para que a equipe se familiarize com ela e comece a gerar resultados.

Uma consideração importante é que todas as ferramentas de teste de imparcialidade geralmente exigem que os usuários possuam, pelo menos, conhecimentos básicos de Python e conceitos fundamentais de Machine Learning. Esse requisito é essencial para a equipe de testes, pois muitas dessas ferramentas são baseadas em bibliotecas de Python e envolvem a análise de modelos de aprendizado de máquina. Sem esse conhecimento, a equipe pode encontrar dificuldades para entender como usar as ferramentas corretamente, interpretar os resultados ou mesmo adaptar a ferramenta às necessidades específicas do sistema de recomendação em teste.

4.1.4 Versatilidade

A versatilidade de uma ferramenta é um fator importante, especialmente em testes de imparcialidade. Uma ferramenta versátil deve ser capaz de lidar com uma variedade de cenários e tipos de dados, permitindo que seja adaptada a diferentes sistemas de recomendação e abordagens de avaliação. Isso inclui a capacidade de suportar múltiplas métricas de imparcialidade, como paridade demográfica e impacto desigual, e ser eficaz em diferentes tipos de modelos de recomendação, como filtragem colaborativa ou sistemas híbridos. Além disso, a ferramenta deve ser capaz de processar dados tanto estruturados quanto não estruturados, como interações de usuários, textos, imagens ou vídeos. Essa flexibilidade permite que a ferramenta seja aplicada a diversos contextos e ajustada conforme as necessidades do sistema ou do processo de testes, garantindo sua efetividade e evolução ao longo do tempo.

É importante notar que a grande maioria das ferramentas disponíveis não são voltadas especificamente para sistemas de recomendação, o que torna a versatilidade um fator ainda mais crucial. Como essas ferramentas não são projetadas exclusivamente para esse tipo de sistema, elas precisam ser suficientemente flexíveis para serem adaptadas a diferentes modelos e abordagens de recomendação.

4.1.5 Escalabilidade

A ferramenta deve ser capaz de lidar eficientemente com grandes volumes de dados, múltiplos usuários e itens, permitindo realizar testes de imparcialidade em sistemas de recomendação de grande escala. Ela deve garantir que, mesmo com uma base de dados extensa, a performance não seja comprometida e a precisão dos resultados seja mantida, possibilitando análises rápidas e confiáveis em sistemas complexos e dinâmicos.

4.1.6 Compatibilidade com Sistemas de Recomendação

Conforme mencionado anteriormente, a maioria das ferramentas disponíveis não é específica para sistemas de recomendação. Embora a versatilidade das ferramentas ajude, ela pode não ser suficiente para lidar com os desafios únicos desses modelos. Ter uma ferramenta voltada especificamente para sistemas de recomendação seria altamente benéfico, considerando as particularidades e os desafios desses sistemas, como por exemplo:

- A imparcialidade deve ser avaliada tanto em relação aos usuários (indivíduos ou grupos), assegurando que todos tenham acesso equitativo às recomendações, quanto aos itens (individuais ou em grupo), garantindo que nenhum item seja sistematicamente favorecido ou prejudicado nas sugestões.

- É crucial identificar se um usuário está sendo repetidamente exposto às mesmas recomendações, a chamada câmara de eco, o que pode limitar sua exposição a novos conteúdos e reforçar suas preferências ou visões preexistentes. Esse fenômeno pode levar à falta de diversidade nas recomendações, criando um ambiente onde o usuário não tem a oportunidade de descobrir novos itens ou explorar diferentes perspectivas. Além disso, o efeito da câmara de eco pode contribuir para a polarização, onde o usuário se vê cada vez mais imerso em um "bolha"de conteúdos semelhantes aos seus gostos iniciais, dificultando a descoberta de novas opções e potencialmente exacerbando viéses de comportamento.

- É importante evitar que itens populares sejam recomendados com mais frequência, pois isso pode diminuir a diversidade das recomendações e criar um ciclo de popularidade onde os

itens mais conhecidos são constantemente reforçados. Esse fenômeno está relacionado ao efeito Matthew, onde "os ricos ficam mais ricos", ou seja, os itens que já têm alta popularidade acabam recebendo ainda mais visibilidade, enquanto itens menos populares, mas potencialmente valiosos, ficam em segundo plano.

4.1.7 Personalização de testes

Como vimos, a imparcialidade é um conceito abstrato, o que significa que não há um "certo"ou "errado"definitivo nos testes. Por isso, é essencial que a ferramenta permita personalizar as métricas e se adaptar ao contexto específico do sistema de recomendação, de forma a se manter dentro dos requisitos de imparcialidade estabelecidos. Isso garante que a avaliação seja relevante e alinhada aos objetivos do sistema, ao mesmo tempo em que respeita as definições e as normas de imparcialidade que foram previamente definidas para o projeto.

4.1.8 Suporte

A ferramenta não precisa ser necessariamente nova ou ter atualizações constantes, mas é essencial que ela ainda receba suporte ativo. Isso é importante porque, ao longo do tempo, podem surgir bugs ou outros problemas imprevistos. A falta de suporte torna a ferramenta arriscada, pois ao construir um fluxo de testes em cima dela, a equipe pode se deparar com um problema insolucionável a qualquer momento, o que comprometeria todo o trabalho. Ter suporte assegura que eventuais falhas sejam corrigidas, minimizando riscos e garantindo a continuidade dos testes de forma confiável.

5 Conclusão

5.1 O que foi aprendido

Neste trabalho, estudamos os sistemas de recomendação, explorando seus diferentes tipos: baseado em conteúdo, que baseiam suas recomendações com base em itens similares; colaborativos, que se baseiam em usuários semelhantes; e híbridos, que combinam ambos. Compreender esses modelos é essencial para dar início ao estudo dos testes de imparcialidade, pois nos permite identificar aspectos específicos de cada tipo de sistema e como eles podem influenciar a imparcialidade das recomendações. Por exemplo, em sistemas colaborativos, o viés de popularidade (efeito Matthew) e a formação de câmaras de eco podem limitar a diversidade de itens recomendados e reforçar estereótipos ou padrões existentes.

Além disso, ao avaliar a imparcialidade, é crucial que ela seja observada tanto para os usuários quanto para os itens. A imparcialidade para os usuários envolve garantir que grupos diversos de usuários não sejam prejudicados por viéses algorítmicos, enquanto a imparcialidade para os itens busca evitar que itens populares dominem as recomendações em detrimento de itens menos conhecidos ou novos.

Assim, pudemos iniciar nossos estudos sobre os testes de imparcialidade em si. Entendemos que imparcialidade é um conceito difícil de definir e aplicar na prática, pois depende dos requisitos específicos de cada contexto. Portanto, existem diversas definições de imparcialidade, e é crucial compreender qual delas se adapta melhor ao nosso sistema de recomendação.

Também aprendemos vários aspectos sobre testes e como podemos identificar defeitos de imparcialidade em diferentes componentes do sistema, até mesmo naqueles que não estão diretamente relacionados ao aprendizado de máquina. Além disso, entendemos os processos envolvidos nesses testes, que são divididos em duas etapas principais: a geração das entradas e a identificação do

oráculo. A definição do oráculo, por sua vez, pode ser um desafio, pois é difícil estabelecer critérios claros e universais de imparcialidade, o que torna esse processo particularmente complexo.

Por último, e como principal foco deste trabalho, pudemos entender que os testes de imparcialidade podem ser realizados por meio de ferramentas, mas que não é uma tarefa simples encontrar uma ferramenta ideal, especialmente quando se trata de uma ferramenta específica para sistemas de recomendação. Analisamos os principais critérios que devem ser considerados na escolha da ferramenta mais adequada, como a facilidade de uso, a capacidade de personalização, a escalabilidade e a versatilidade. Esses fatores são fundamentais para tomar uma decisão informada sobre qual ferramenta melhor se adapta ao escopo desejado e às particularidades do sistema de recomendação em questão.

Em resumo, ao longo deste trabalho, foi possível perceber que, embora existam desafios na definição e avaliação da imparcialidade em sistemas de recomendação, é possível avançar na identificação de ferramentas eficazes para esse fim. A análise dos critérios de escolha, como personalização, escalabilidade e versatilidade, mostrou que a seleção da ferramenta adequada é essencial para garantir sistemas justos e equilibrados. A partir desse estudo, podemos concluir que, embora o campo de testes de imparcialidade ainda apresente complexidades, há caminhos claros para a melhoria contínua e para a criação de recomendações mais justas e inclusivas.

5.2 Sugestões de trabalhos futuros

Para trabalhos futuros, com base nesta análise de como escolher a melhor ferramenta, podemos avançar para a aplicação prática, buscando ou até mesmo desenvolvendo uma ferramenta que esteja alinhada com os requisitos necessários. Isso permitiria não apenas testar sistemas de recomendação existentes, mas também criar um ambiente de avaliação que possibilite a implementação e teste de diferentes abordagens de imparcialidade.

Com uma ferramenta adequada em mãos, seria possível realizar testes de imparcialidade de forma mais robusta, analisando não apenas os resultados, mas também como os diferentes testes impactam a performance e a justiça do sistema de recomendação.

Além disso, essa abordagem prática ajudaria a identificar de forma mais clara como os defeitos de imparcialidade podem ser corrigidos e quais ajustes são necessários para melhorar a equidade dos resultados. Isso também poderia gerar insights valiosos sobre como os diferentes sistemas de recomendação respondem a medidas de imparcialidade e como podemos otimizar essas ferramentas para garantir uma experiência mais justa e inclusiva para os usuários.

Dessa forma, esse trabalho poderia ser um ponto de partida para uma aplicação mais prática e concreta dos conceitos discutidos.

Referências

[1] Jayalakshmi S, Ganesh N, Čep R, Senthil Murugan J., *Movie Recommender Systems: Concepts, Methods, Challenges, and Future Directions.,* (2022).

[2] Roy, D., Dutta, M., *A systematic review and research perspective on recommender systems,* (2022).

[3] Wang, Yifan and Ma, Weizhi and Zhang, Min and Liu, Yiqun and Ma, Shaoping, *A Survey on the Fairness of Recommender Systems* (2023).

[4] Chen, Zhenpeng and Zhang, Jie M. and Hort, Max and Harman, Mark and Sarro, Federica, *Fairness Testing: A Comprehensive Survey and Analysis of Trends* (2024).

[5] Medeiros, *Estudo sobre Sistemas de Recomendação Colaborativos* (2013).

[6] Nguyen, Baldassarre, Santos, de Lima, *From Literature to Practice: Exploring Fairness Testing Tools for the Software Industry Adoption* (2024).