Sistema de Recomendação com *User Reviews*: Caso Americanas

Eduardo Garcia[1], **Juliana Resplande**[1], **Luana Martins**[1], **Werikcyano Lima**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brazil

{eduardogarcia, julianaresplande, luana.martins, werikcyano}@inf.ufg.br

Resumo. *Este trabalho propõe um sistema de recomendação que combina análise de sentimento e filtragem colaborativa para gerar recomendações de produtos mais precisas e relevantes. O sistema leva em consideração as classificações e avaliações dos usuários, criando uma plataforma ideal para que os usuários encontrem produtos que se adaptam às suas necessidades e interesses, utlizando análise de sentimento para melhorar a qualidade das recomendações. Os experimentos realizados no conjunto de dados B2W-Reviews01 demonstram a importância de modelos de análise de sentimento para alcançar melhores recomendações para o usuário.*

Abstract. *This work proposes a recommendation system that combines sentiment analysis and collaborative filtering to generate more accurate and relevant product recommendations. The system takes user ratings and reviews into account, creating an ideal platform for users to find products that suit their needs and interests, using sentiment analysis to improve the quality of recommendations. The experiments carried out on the B2W-Reviews01 dataset demonstrate the importance of sentiment analysis models to reach better recommendations for the user.*

1. Introdução

Sistemas de Recomendação (RS) são métodos usados para analisar e determinar quais conteúdos cada usuário tem maior probabilidade de gostar, otimizando a experiência de consumo. Estas sugestões podem estar relacionadas a vários processos de tomada de decisão, como a compra de itens, escolha de música ou leitura de notícias. Desde o e-commerce até a publicidade online, os sistemas de recomendação são fundamentais para melhorar a experiência do usuário no mundo digital.

Uma recomendação personalizada usa informações do perfil do usuário e outras relacionadas. A qualidade da recomendação depende da quantidade e qualidade dos dados usados. Por exemplo, nos métodos tradicionais de recomendação, quanto mais dados coletados e usados sobre o usuário-alvo e usuários semelhantes, melhores as recomendações [Adomavicius and Tuzhilin 2005].

Muitos sites de compras online pedem aos clientes que avaliem e comentem os itens que vendem. Os clientes analisam essas análises para ver se outros compradores tiveram boas experiências ou deram uma pontuação alta, para decidir se um item é bom para comprar. Ou, se as críticas e avaliações forem ruins, os clientes podem saber se um item não vale a pena comprar. Portanto, usar classificações (números) e avaliações

(palavras) ajuda o sistema de recomendação a fazer melhores sugestões [Dang et al. 2021, Elahi et al. 2023].

Este trabalho propõe um sistema de recomendação que combina análise de sentimento e filtragem colaborativa. Foi desenvolvida uma arquitetura de sistema de recomendação adaptável que inclui técnicas de análise de sentimento, juntamente com algoritmos de recomendação, para gerar recomendações de produtos com base nas classificações e avaliações dos usuários. Os resultados do estudo empírico, realizado com um conjunto de dados de análises de produtos em português, mostram que a combinação de análise de sentimento e métodos de filtragem colaborativa melhoram o desempenho do sistema de recomendação, oferecendo melhores recomendações ao usuário.

No restante deste documento estão definidos os dados utlizados (Seção 2), descrição da metodologia abordada no trabalho (Seção 3), os resultados obtidos (Seção 4) e conclusões finais (Seção 5).

2. *B2W Reviews*

As Reviews utilizadas neste trabalho foram retiradas do dataset "b2w-reviews01", disponibilizado na plataforma Hugging Face [1]. Este conjunto de dados contém avaliações de produtos da B2W (Americanas) feitas pelos próprios clientes, através do site da empresa, onde as principais informações são Texto da avaliação, Id do avaliador, data da submissão, Marca do produto, Nota atribuída, etc. A Tabela 1 abaixo representa um exemplo do conjunto de dados.

submission_date	2018-05-31 23:30:50
product_id	17962233
product_name	Carregador De Pilha Sony + 4 Pilhas Aa 2500mah
product_brand	None
site_category_lv1	Câmeras e Filmadoras
site_category_lv2	Acessórios para Câmeras e Filmadoras
overall_rating	5
recommend_to_a_friend	Yes
review_title	Ótimo produto!
review_text	Vale muito, estou usando no controle do Xbox e me durou uma semana a carga de um par, e isso jogando todos os dias!
reviewer_id	15f20e95ff44163f3175aaf67a5ae4a94d5030b409e521...
reviewer_birth_year	1988.0
reviewer_gender	M
reviewer_state	RS

Tabela 1. Exemplo do conjunto de dados da B2W Reviews. As colunas em negrito representam os dados utilizados pelo sistema de recomendação com a análise de sentimento.

Como este conjunto de dados é retirado do site da empresa e as avaliações podem

[1] https://huggingface.co/datasets/ruanchaves/b2w-reviews01

ter sido feitas por qualquer tipo de cliente que tenha comprado o produto, existe uma grande chance de conter informações incompletas, erradas ou que não reflitam aquilo que o usuário tinha a intenção de expressar, como nos exemplos da Tabela 3.

Após uma análise mais detalhada e direcionada para o problema abordado neste trabalho, realizamos uma limpeza nos dados utilizando os seguintes critérios:

1. Retiradas de colunas com muitos dados faltantes
2. Exclusão de colunas que não agregariam ao problema
3. Exclusão de colunas que poderiam vazar informações para o modelo
4. Exclusão de linhas (avaliações) que continham dados faltantes
5. Exclusão de linhas em que as avaliações ultrapassem 80 palavras
6. Criação da coluna que é a junção do Título com o Corpo da avaliação
7. Criação de um faixa de classificação sobre as notas atribuídas pelos usuários:

   (a) 1 e 2 estrelas -> Ruim (-1)

   (b) 3 estrelas -> Neutro (0)

   (c) 4 e 5 estrelas -> Bom (1)

Uma vez que essa sequência de critérios foi atendida, nesta ordem, as colunas resultantes foram: *Review, Review-id, product-id, product-name, overall-rating, rating.* Desta forma, na Tabela 2 é mostrado como o dataset pôde ser dividido entre os interesses do problema, de forma balanceada e que minimize os problemas supracitados, sobre a qualidade dos dados.

	Treino		Teste	Total
	SA	RS	Ambos	
%	40	40	20	100
#	51513	51513	25757	128783

Tabela 2. Divisão dos dados para treinamento e testes dos modelos avaliados.

Item ID	Avaliação	Classificação	Sentimento
120587985	livro muito bom. livro muito bom para se estudar e complementar os estudos.	3	Positivo
10561638	Não gostei do produto. Produto faz muito barulho e não está gelando como esperava, além de ter chegado avariado e eu não ter como trocar pois era para estar no lacre só que não tenho como analisar o produto sem abrir. Não recomendo.	3	Negativo
127795015	Perfeito ! Produto bem embalado e chegou antes do prazo, totalmente satisfeita com esse produto da Samsung.	1	Positivo
26550711	não compre com a FARMA DELIVERY. Eles têm estoque mas mandam produto vazando. A Farma Delivery não cumpre prazo para retirada do produto quando tem cancelamento.	5	Negativo

Tabela 3. Exemplos de inconsistências entre as avaliações fornecidas pelos usuários e do sentimento extraído dessas avaliações. Nota: o intervalo da classificação é [1,5].

3. Metodologia

Nesta seção será discutido do princípios do funcionamento de métodos de Análise de Sentimento e de Sistema de Recomendação, e como ambas as técnicas foram aplicadas para resolver o problema de recomendação para e-commerce.

3.1. Análise de sentimentos

A Análise de Sentimentos (SA) é um processo que usa ferramentas de Inteligência Artificial e Processamento de Linguagem Natural (PLN) para analisar documentos textuais e identificar sentimentos e opiniões [Chen et al. 2022]. Abordagens de SA podem ser usados para descobrir pistas ocultas e potenciais pontos fracos, a fim de obter a satisfação do cliente e expandir o escopo do mercado [Tang et al. 2019]. De forma objetiva, a tarefa de SA classifica trechos documentos e sentenças como polaridade positiva ou negativa [Burke 2002, Birjali et al. 2021].

Uma forma básica de tratar o problema seria utilizar classificadores estatísticos como regressão logística para categorizar cada sentença com o conjunto de palavras ba-

lanceadas por importância dado pelo cálculo do *Term Frequency–Inverse Document Frequency* (TF-IDF) [Imamah and Rachman 2020, Birjali et al. 2021].

A fórmula do TF-IDF para uma palavra w está representada na equação 1:

$$\text{TF-IDF}(w) = tf_w \times idf_w = tf_w \times log\left(\frac{N}{N_w}\right) \tag{1}$$

Onde:

- tf_w: a frequência da palavra w no documento
- idf_w: a frequência inversa de documentos que contém a palavra w
- N: número total de documentos
- N_w: número de documentos que contém a palavra w

Apesar do potencial do SA para fornecer *insights* úteis sobre o sentimento do cliente, ainda existem alguns desafios que precisam ser abordados. Os resultados da SA são desafiados pela ambiguidade da linguagem humana e pela complexidade de vários modificadores linguísticos, composição e dilemas associados ao contexto. Esses desafios podem levar a discrepâncias entre a polaridade de sentimento induzida automaticamente e a interpretação humana do sentimento, o que pode fazer com que as máquinas percam as nuances sutis da linguagem humana. Além disso, a falta de anotações precisas e abrangentes para modelos de análise de sentimento pode levar a resultados tendenciosos e previsões incorretas, dificultando a captura precisa do sentimento do cliente de maneira confiável. Para superar esses obstáculos, foram propostas várias abordagens para melhorar a precisão do SA, como a construção de conjuntos de dados com anotações mais abrangentes, o uso de modelos mais sofisticados e a combinação de vários modelos para a mesma tarefa.

Em consoante com a área em PLN, os modelos estados da arte em SA pertencem à categoria Transformers [Vaswani et al. 2017]. Essa arquitetura segue a linha de *Transfer-learning*: o modelo é inicialmente pré-treinado em um conjunto de textos com cerca de bilhões de palavras em tarefas de modelo de linguagem, e posteriormente treinado para a tarefa específica, podendo ser sumarização, classificação de texto, etc. O principal destaque dessa arquitetura é o entendimento contextual dos documentos, o sistema artificial consegue diferenciar polissemia e características complexas da linguagem [Kalyan et al. 2021].

O BERT é um modelo de linguagem contextual em inglês. Essa arquitetura é do tipo *encoder*, ou seja, o modelo possui como saída características densas, as quais são aproveitadas para classificadores. No segundo semestre de 2018, o modelo atingiu o estado em tarefas de análise de sentimento [Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina 2018].
Foi-se desenvolvido modelo BERT em português, o Bertimbau, o qual mostrou resultados satisfatórios em tarefas de análises de sentimentos [Souza et al. 2020, Capellaro 2021, Gumiel et al. 2021].

3.2. Sistema de Recomendação

Um Sistema de Recomendação (RS) é projetado para dar sugestões personalizadas sobre produtos ou serviços para ajudar na tomada de decisão. As empresas de *e-commerce*,

por exemplo, costumam usar sistemas de recomendação para sugerir novos produtos aos clientes. Os métodos usados para sistemas de recomendação podem ser divididos em três categorias: baseados em conteúdo, filtragem colaborativa e sistemas de recomendação híbridos.

Sistemas de recomendação baseados em conteúdo (CB) [Pazzani and Billsus 2007] observam as características dos itens e perfis do usuário. Os perfis são criados com base nos itens acessados e na informação de referência sobre os usuários. Os métodos de CB filtra itens similares aos que os usuários já consumiram ou avaliaram positivamente. Por outro lado, a filtragem colaborativa (CF) [Zhang et al. 2014] encontra itens baseados nas avaliações de usuários semelhantes, que compartilham interesses e preferências comuns. Ao combinar essas duas abordagens, os sistemas híbridos podem superar os pontos fracos deles separadamente [Burke 2002]. Esta abordagem híbrida pode ser implementada de diferentes maneiras, como a fusão direta de recomendações CB e CF, a integração de dados CB e CF, ou a aplicação de pesos diferentes para cada método.

Cada abordagem de recomendação tem vantagens e limitações. A filtragem colaborativa tem alguns problemas, como esparsidade, escalabilidade e cold-start. Esparsidade é quando temos muitos dados. Escalabilidade é quando faltam dados de classificação. Cold-start é quando um usuário ou um item é adicionado ao sistema. Podemos resolver esses problemas usando análise de sentimento juntamente com métodos de recomendação.

3.3. Sistema de Recomendação com Análise de Sentimento

O método proposto nesse trabalho consiste em um método de recomendação que combina filtragem colaborativa e análise de sentimentos. Nosso objetivo é tornar as recomendações ao usuário mais confiáveis, combinando a análise de sentimento das avaliações ou comentários do usuário com os métodos tradicionais de recomendação. O sistema tem duas partes (Figura 1) - uma parte constrói os modelos de sentimento e a outra fornece ao usuário recomendações.

Tradicionalmente, os sistemas de recomendação focam nas avaliações dos usuários para construir e avaliar as recomendações. Isso pode não ser suficiente para capturar as preferências dos usuários (veja exemplos na Tabela 3). Espera-se que a utilização das classificações por sentimento aumente a precisão e forneça informações úteis sobre os usuários.

Foram utilizados algoritmos de CF como o Singular Value Decomposition (SVD), Non-Negative Matrix Factorization (NMF), e SVD++ (um derivado do SVD).O objetivo é obter uma melhor precisão preditiva devido à adição de informações de *feedback* implícitas fornecidas pelo sentimento. Os resultados do método de recomendação e da análise de sentimento foram combinados para gerar uma classificação e usados para criar uma lista de recomendações. A estimativa da avaliação do usuário u_a no item i_j no conjunto de teste é definido por:

$$pr_{aj} = \beta * mf_{aj} + (1 - \beta) * sent_j \tag{2}$$

onde:

- mf_{aj}: avaliação do usuário u_a no item i_j predito por métodos de Fatorização de Matrizes (SVD, SVD++, NMF) sem utilizar sentimento.
- $sent_j$: média das avaliação de todos os usuários no item i_j predito por métodos de Análise de Sentimento (TF-IDF, LSTM, BERTimbau)
- β: parâmetro utilizado para ajustar a importância de cada termo da equação

Como citado anteriormente, os modelos de análise de sentimento são utilizados para definir a probabilidade de cada *review* ser um sentimento positivo ($sent_{aj} = 1$) ou um sentimento negativo ($sent_{aj} = 0$). Ao final é aplicado um método de categorização dos valores preditos, convertendo em pontuações de sentimento de 1 a 5.

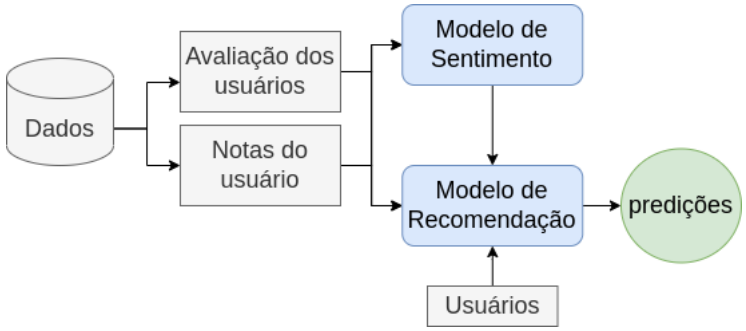

Figura 1. Arquitetura do Sistema de Recomendação.

4. Experimentos e Resultados

Nesta seção será discutido os resultados obtidos dos experimentos realizados para análise de sentimento e para o modelo proposto de recomendação. Os códigos desenvolvidos estão disponíveis no Github [2].

4.1. Análise de sentimentos

Para o problema de Análise de Sentimentos iremos utilizar dois modelos para comparar o desempenho dos modelos, o *TF-IDF* como baseline para o problema e o *BERTimbau* como modelo de linguagem contextual que acreditamos que chegaria mais próximo do resultado estado-da-arte para análise de sentimentos.

Para ambos os modelos realizamos o balanceamento dos dados realizando um *down-sampling* das reviews, obtendo um conjunto de dados com 21708 instancias para treinamento.

Para validação separamos um conjunto de 10000 instancias para busca de hiperparâmetros dos modelos, as bibliotecas utilizadas para treinamento foi o *Huggingface* e o *scikit-learn* respectivamente para os modelos BERT e Regressão logistica, os hiperparâmetros finais utilizados estão descritos na tabela 4.

[2]Código utilizado nos experimentos: `https://github.com/ju-resplande/projeto_nlp`

Hiperparâmetro	BERTimbau	Regressão logística
Learning Rate	5e-6	-
Batch Size	32	-
Num. Epochs	2	-
Weight Decay	0.01	-
Max. Iterations	-	100

Tabela 4. hiperparâmetros utilizados para o treinamento dos modelos de analise de sentimento.

Para o *TF-IDF* fazemos um preprocessamento para a normalização do texto de comentário de usuário, retirando pontuação, caracteres especiais e fazendo o *steamming* dos tokens. Já com o *BERTimbau* apenas utilizamos apenas o preprocessamento do próprio tokenizador do modelo. O resultado de ambos sob os dados de teste está descrito na 5.

	Accuracy	Precision	Recall	F1-Score
TF-IDF + Regressão logística	0.95	0.98	0.94	0.96
BERTimbau Base	**0.974**	**0.991**	**0.971**	**0.981**

Tabela 5. Métricas para o dataset de teste de análise de sentimento com os dados da B2W.

4.2. Sistema de Recomendação

Nesta seção é apresentado os experimentos conduzidos para avaliar o desempenho da abordagem proposta para sistemas de recomendação. Para validar nossa abordagem de recomendação, comparamos o desempenho de três métodos de recomendação de CF como baseline e os mesmos métodos aprimorados com nossa proposta envolvendo o uso de análise de sentimento das avaliações. Foi utilizado a implementação dos algoritmos SVD, NMF e SVD++ fornecidos pela biblioteca Surprise [3], e para avaliar o treinamento dos modelos foi utilizado validação cruzada k-fold, com $k = 5$. Para a análise de sentimento, foi utilizado o modelo BERTimbau.

Para avaliar a confiabilidade das previsões de classificação, usamos Raiz do Erro Quadrático Médio (RMSE) e Erro Absoluto Médio (MAE). Ambas as métricas expressam o erro médio de previsão e podem variar de 0 a ∞, onde valores mais baixos são melhores. Contudo, o RMSE dá um peso relativamente alto a erros grandes, o que significa que o RMSE deve ser mais útil quando grandes erros são particularmente indesejáveis.

A Tabela 6 mostra os resultados das métricas MAE e RMSE para previsão de classificação em análises de produtos. Os resultados foram calculados com base no algoritmo SVD, NMF e SVD++ com e sem o uso de análise de sentimento, com $beta = 0.7$. O parâmetro Beta (β) é utilizado para ajustar a importância do resultado da recomendação sem e com sentimento na Equação (2). A Figura 2 ilustra os resultados comparativos obtidos do recomendador com análise de sentimento em diferentes valores do parâmetro β em relação aos obtidos do recomendador sem análise de sentimento.

[3]http://surpriselib.com/

	MAE			RMSE		
	SVD	NMF	SVD++	SVD	NMF	SVD++
sem sentimento	1.07	1.53	1.14	**1.38**	1.48	1.38
com sentimento	**0.96**	**1.07**	**0.95**	**1.38**	**1.39**	**1.37**

Tabela 6. Valores de MAE e RMSE sem e com modelo de análise de sentimento para os dados da B2W.

Os resultados mostram que MAE e RSME produzido pela abordagem que combina CF com análise de sentimento são melhores que as taxas de erro produzidas pelos métodos tradicionais de CF sem sentimento. Na tabela 6 foi fixado o valor de $\beta = 0.7$, que se mostrou consistentemente melhor em todos os cenários avaliados, analisando a métrica de RSME. Entretanto, para o método NMF pode-se observar na Figura 2 melhores resultados para valores menores de β.

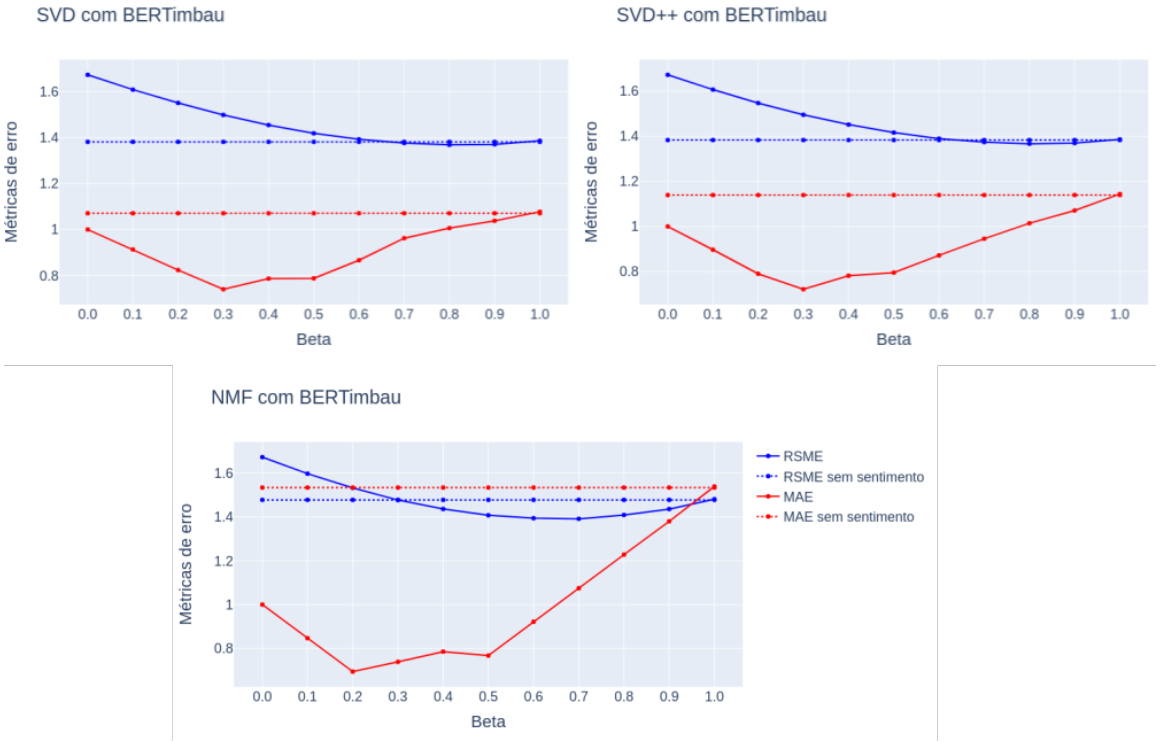

Figura 2. Métricas de erro para os algoritmos SVD, SVD++ e NMF com scores de sentimento do modelo BERTimbau, para diferentes valores de β.

5. Conclusão

Neste trabalho, propomos uma aplicação de análise de sentimentos para sistemas de recomendação. A arquitetura do sistema descrita aqui pode integrar várias técnicas, como modelos de análise de sentimentos e métodos para sistemas de recomendação. Essa arquitetura pode ser usada para desenvolver um sistema de recomendação para *e-commerce* que aproveita a análise de sentimentos de opiniões e avaliações de usuários. Foram realizados experimentos com avaliações de produtos da empresa B2W. Os resultados demonstraram

que o uso conjunto de análise de sentimentos e métodos de filtragem colaborativa melhora significativamente o desempenho do sistema de recomendação. Isso é alcançado explorando informações adicionais dos dados de avaliações/comentários de usuários. A integração dessas informações aos métodos de recomendação tradicionais torna o sistema de recomendação mais confiável e capaz de fornecer melhores recomendações aos usuários.

Referências

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Birjali, M., Kasri, M., and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Capellaro, L. (2021). Análise de polaridade e de tópicos em tweets no domínio da política no brasil.

Chen, J., Song, N., Su, Y., Zhao, S., and Zhang, Y. (2022). Learning user sentiment orientation in social networks for sentiment analysis. *Information Sciences*, 616:526–538.

Dang, C. N., Moreno-García, M. N., and Prieta, F. D. l. (2021). An approach to integrating sentiment analysis into recommender systems. *Sensors*, 21(16).

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elahi, M., Khosh Kholgh, D., Kiarostami, M. S., Oussalah, M., and Saghari, S. (2023). Hybrid recommendation by incorporating the sentiment of product reviews. *Information Sciences*, 625:738–756.

Gumiel, Y. B., Lee, I., Soares, T. A., Ferreira, T. C., and Pagano, A. (2021). Sentiment analysis in portuguese texts from online health community forums: data, model and evaluation. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 64–72. SBC.

Imamah and Rachman, F. H. (2020). Twitter sentiment analysis of covid-19 using term weighting tf-idf and logistic regression. In *2020 6th Information Technology International Seminar (ITIS)*, pages 238–242.

Kalyan, K. S., Rajasekharan, A., and Sangeetha, S. (2021). AMMUS : A survey of transformer-based pretrained models in natural language processing. *CoRR*, abs/2108.05542.

Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In *The Adaptive Web*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Tang, F., Fu, L., Yao, B., and Xu, W. (2019). Aspect based fine-grained sentiment analysis for online reviews. *Information Sciences*, 488:190–204.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Zhang, R., Liu, Q.-d., Chun-Gui, Wei, J.-X., and Huiyi-Ma (2014). Collaborative filtering for recommender systems. In *2014 Second International Conference on Advanced Cloud and Big Data*, pages 301–308.