Previsão de Utilidade de Avaliações de Produtos Online na Língua Portuguesa Brasileira

Larissa F. S. Britto[1,2] **, Luciano D. S. Pacífico** [3] **, Teresa B. Ludermir** [1]

[1]Centro de Informática – CIn
Universidade Federal de Pernambuco – UFPE – Recife, PE – Brasil

[2]Centro de Pesquisa e Desenvolvimento em Telecomunicações –
CPQD – Campinas, SP – Brasil

[3]Departamento de Computação (DC)
Universidade Federal Rural de Pernambuco – UFRPE – Recife, PE – Brasil

lfsb@cin.ufpe.br, luciano.pacifico@ufrpe.br, tbl@cin.ufpe.br

Abstract. *With the growth of e-commerce, online product reviews have become a significant factor in influencing users' purchasing decisions. However, users may be harmed by the information overload on online review platforms. In this study, we evaluate different approaches to identify helpful product reviews. To achieve this, a large dataset of Amazon reviews from various product domains was proposed. The results demonstrate that it is possible to predict the usefulness of online reviews without relying on any handcrafted features.*

Resumo. *Com o crescimento do comércio eletrônico, as avaliações de produtos online se tornaram um fator importante na decisão de compra dos consumidores. No entanto, os usuários podem ser prejudicados pela sobrecarga de informações em plataformas de avaliação online. Neste estudo, avaliamos diferentes abordagens para identificar avaliações de produtos úteis. Para esse propósito, foi proposto um grande conjunto de dados de avaliações da Amazon em diferentes domínios de produtos. Os resultados mostram que é possível prever a utilidade das avaliações online sem depender de recursos personalizados sem depender de quaisquer características feitas manualmente.*

1. Introdução

Com o crescimento da Internet, o comércio eletrônico se tornou um dos métodos de compra mais importantes. O processo de tomada de decisão de compra é único para as compras online. Vários elementos motivacionais, como fatores situacionais, características do produto e experiências anteriores de compras online, podem influenciar as atitudes dos consumidores em relação às compras online [Senecal et al. 2005].

As avaliações de usuários têm um forte impacto na decisão de compra. Dados significativos sobre opiniões de usuários, é uma fonte rica de conhecimento para a Análise de Sentimentos (AS), área focada em detectar sentimentos e opiniões em textos [Henrickson et al. 2019, Tonkin 2016].

Uma tarefa de AS que ganhou popularidade é a predição de utilidade, que visa resolver o problema da sobrecarga de informações em plataformas de avaliação online,

que afeta a capacidade de clientes de avaliar a qualidade de produtos ou empresas ao tomar decisões de compra [Bilal and Almazroi 2022]. As plataformas de avaliação introduziram e implementaram um sistema de votos úteis (onde os usuários votam nas avaliações que consideram úteis), mas essa estratégia depende muito da cooperação do usuário.

Neste artigo, avaliamos experimentalmente diferentes abordagens estabelecidas na literatura de PLN para classificar a utilidade em avaliações de produtos online. Considerando que a grande maioria dessas abordagens se refere a corpora em inglês e que as aplicações em português brasileiro são relativamente escassas, propomos uma grande base de dados de avaliações de produtos da Amazon [Tufchi et al. 2023, ElKafrawy et al. 2023], para predição de utilidade.

O restante do artigo está organizado da seguinte forma. Na Seção 2, apresentamos e descrevemos a base de dados proposto. A Seção 3 descreve nossa configuração experimental. Os resultados experimentais e sua discussão são apresentados na Seção 4. A última seção, Seção 5, conclui o artigo.

2. Base de Dados

Nesta seção, a base de dados utilizado neste trabalho são apresentados e brevemente descritos.

2.1. Data Collection

A base de dados proposto foi extraído do site de vendas Amazon [1], uma das plataformas mais populares para leitura e postagem de avaliações.

Nossa base de dados considera vários domínios de produtos. Para a coleta, selecionamos algumas das categorias mais populares na plataforma. Para cada uma dessas categorias, coletamos todas as avaliações em português do Brasil dos 100 produtos mais vendidos. Além das avaliações, também foram coletadas outras informações que podem ser aplicadas para outras tarefas de análise de sentimentos, como informações sobre o produto e o usuário. A coleta de dados foi realizada entre 28 de junho e 31 de junho de 2022. O Framework Scrapy[2] para extração de dados de websites foi adotado nesta etapa.

2.1.1. Processamento de Dados

Após a coleta de dados, realizamos uma etapa de processamento para normalizar os dados numéricos. Os dados textuais estão disponíveis sem nenhum pré-processamento, para que os pesquisadores que desejam usar esta base de dados possam escolher os métodos mais adequados de acordo com sua pesquisa. As estatísticas da base de dados estão listadas na Tabela 1.

2.2. Anotação dos Dados

A anotação de utilidade foi feita com base nas informações de votos úteis. Como a plataforma da Amazon não fornece informações sobre votos não úteis, neste trabalho, assumimos que qualquer avaliação sem pelo menos um voto útil é considerada não útil.

[1] https://www.amazon.com.br/
[2] https://scrapy.org/

Table 1. Estatísticas da base de dados por classe. Símbolos como emojis não são considerados, resultando em algumas avaliações com comprimento 0.

Medida	Útil	Não Útil	Total
Tamanho Máximo	250	221	250
Tamanho Mínimo	0	0	0
Tamanho Médio	27,47	14,44	15,49
Tokens Únicos	28783	55308	61605
Número de Documentos	30516	348883	379399

2.3. Balanceamento dos Dados

Nossa base de dados apresenta um alto grau de desbalanceamento, o que poderia afetar o desempenho dos classificadores. Para balancear nossa base de dados, reduzimos o tamanho da classe que é abundante, removendo aleatoriamente documentos dessa classe.

2.4. Disponibilidade dos Dados

A base de dados final e o script para seu desenvolvimento serão disponibilizados no seguinte repositório público: `www.github.com/larifeliciana/Helpful-Amazon-PT`.

3. Configuração Experimental

A avaliação experimental deste trabalho tem como objetivo comparar o desempenho de modelos de classificação na predição de utilidade. Foi selecionado um método popular da literatura de classificação de texto para extração de características: TF-IDF. Essas características foram utilizadas como entrada para classificadores tradicionais de aprendizado de máquina: k-vizinhos mais próximos (k-Nearest Neighbors), regressão logística (Logistic Regression), Naive Bayes, Floresta Aleatória (Random Forest) e Máquinas de Vetores de Suporte (Support Vector Machines). Além desses, dois modelos BERT [ElKafrawy et al. 2023] foram ajustados (*fine-tuned*) e utilizados como modelos de classificação: BERTimbau [Souza et al. 2020] (modelos BERT pré-treinados para o português brasileiro) e BERT Multilíngue [Devlin et al. 2018] (pré-treinado em 104 idiomas, incluindo o português).

Na nossa avaliação, foi utilizada a validação cruzada com cinco *folds*, na qual a base de dados proposta foi dividida aleatoriamente em cinco partes balanceadas para formar o conjunto de treinamento e o conjunto de teste. Quatro partes são usadas cada vez para formar o conjunto de treinamento, e a parte restante é usada como conjunto de teste. O processo de reamostragem foi realizado para evitar resultados obtidos por acaso. Foram adotadas métricas de classificação bem conhecidas: Precisão Macro e Micro, Revocação (Recall) e Medida-F (F-measure).

4. Resultados e Discussão

Nesta seção, os resultados experimentais são apresentados e discutidos. A Tabela 2 mostra os resultados de todos os modelos na predição de utilidade.

Conforme mostrado nos resultados da Tabela 2, alguns classificadores tradicionais de aprendizado de máquina tiveram bom desempenho, como a Regressão Logística e a Floresta Aleatória. O SVM obteve o melhor desempenho, alcançando 84,3% de precisão. Os modelos BERT tiveram desempenho muito semelhante ao SVM, alcançando uma precisão de 84,1%. Apesar de terem alcançado um bom desempenho, esses os classificadores

Table 2. Resultados experimentais para predição de utilidade. Melhores resultados para cada métrica aparecem em negrito.

Modelo	Acurácia	Precisão	Revocação	F-Measure	Treinamento (s)	Teste (s)
TF-IDF + KNN	0.7975	0.8947	0.6746	0.7691	2.01	25.46
TF-IDF + LR	0.835	0.8521	0.8108	0.8309	4.45	0.49
TF-IDF + NB	0.8372	**0.9027**	0.7558	0.8227	2.02	0.49
TF-IDF + RF	0.8306	0.8637	0.7852	0.8225	103.68	1.33
TF-IDF + SVM	**0.8427**	0.8751	0.7995	0.8356	1067.68	111.92
BERT (PT-BR)	0.8409	0.8641	0.8091	0.8357	769.74	18.47
BERT (Multilingual)	0.8353	0.8790	0.7781	0.8253	869.06	18.70

apresentaram um longo tempo de execução médio para treinamento e teste, ambos levando mais de 13 minutos. Em contrapartida, os classificadores com desempenho médio, como a Regressão Logística, tiveram um tempo de execução baixo, levando apenas alguns segundos.

Os dois modelos BERT tiveram um desempenho semelhante entre si, o que poderia indicar que ambos os modelos podem ser uma boa opção para essa tarefa.

5. Conclusões

A quantidade de opiniões fornecidas pelos usuários na Internet todos os dias tem feito com que tarefas de análise de sentimentos sejam altamente requisitadas pelas empresas. Dois desses desafios são abordados neste artigo: predição de utilidade. Um conjunto de dados extenso e rico em informações foi desenvolvido por meio da coleta de avaliações de produtos da Amazon em português do Brasil.

Diferentes abordagens da literatura para análise de sentimentos foram adotadas para extração de características e classificação. O BERT, uma técnica de ponta para várias tarefas de Processamento de Linguagem Natural, é usado neste estudo juntamente com uma técnica tradicional de extração de características, TF-IDF, e classificadores de aprendizado de máquina (k-Nearest Neighbors, Logistic Regression, Naive Bayes, Random Forest e Support Vector Machines).

Os modelos BERT selecionados (BERTimbal e Multilingual), apesar do longo tempo de treinamento, demonstraram ser ótimos modelos para ambas as tarefas, alcançando alta precisão. O classificador SVM também obteve ótimos resultados, porém enfrentou problemas de escalabilidade devido ao longo tempo de treinamento e teste.

Existem algumas limitações associadas à nossa metodologia em termos de rotulação de ajuda. Rotular avaliações sem classificação de ajuda como *não útil* pode não ser a melhor abordagem, pois várias razões podem levar a uma avaliação não ser votada, como baixa demanda pelo produto e avaliações publicadas recentemente (com pouca visualização). Apesar desse problema, os resultados garantem a qualidade da base de dados e mostram como essa tarefa pode ser realizada sem o uso de características criadas manualmente.

Em trabalhos futuros, pretendemos estender a avaliação comparativa com outros modelos estado-da-arte da literatura de classificação de texto, como CNN e BiLSTM. Também pretendemos analisar como diferentes etapas de pré-processamento (como remoção de stopwords e stemming) podem afetar os resultados.

References

Bilal, M. and Almazroi, A. A. (2022). Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews. *Electronic Commerce Research*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

ElKafrawy, P., Mahgoub, A., Atef, H., Nasser, A., Yasser, M., Medhat, W. M., and Darweesh, M. S. (2023). Sentiment analysis: Amazon electronics reviews using bert and textblob.

Henrickson, K., Rodrigues, F., and Pereira, F. C. (2019). Chapter 5 - Data Preparation. In Antoniou, C., Dimitriou, L., and Pereira, F., editors, *Mobility Patterns, Big Data and Transport Analytics*, pages 73–106. Elsevier.

Senecal, S., Kalczynski, P., and Nantel, J. (2005). Consumers' decision-making process and their online shopping behavior: A clickstream analysis. *Journal of Business Research*, pages 1599–1608.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.

Tonkin, E. L. (2016). Chapter 2 - A Day at Work (with Text): A Brief Introduction. In Tonkin, E. L. and Tourte, G. J. L., editors, *Working with Text*, Chandos Information Professional Series, pages 23–60. Chandos Publishing.

Tufchi, S., Yadav, A., Rai, V. K., and Banerjee, A. (2023). Sentiment analysis on amazon product review: A comparative study. In Khanna, A., Polkowski, Z., and Castillo, O., editors, *Proceedings of Data Analytics and Management*, pages 139–149, Singapore. Springer Nature Singapore.