Impacto do pré-processamento em *datasets* de sentimento de *e-commerce* em português

Diego D. Bottero[1], Giancarlo Lucca[1], Joelson S. Junior[2], João Pedro S. Moreira[1]
Eduardo N. Borges[2], Rafael A. Berri[2], Bruno L. Dalmazo[2]

[1]Universidade Católica de Pelotas (UCPel), Campus I,
Rua Gonçalves Chaves 373, 96015-560 Pelotas, RS, Brazil

[2]Universidade Federal do Rio Grande (FURG),
Campus Carreiros, Av. Itália km 8, 96203-900 Rio Grande, RS, Brazil

{diego.bottero,giancarlo.lucca,joao.moreira}@ucpel.edu.br

{dalmazo,eduardoborges,rafaelberri}@furg.br, joelsonsartori@gmail.com

Abstract. *Although significant amounts of datasets are available, data cleaning and standardisation usually reduce the number of instances. This study analyses the impact of a preprocessing pipeline text normalisation, deduplication, and downsampling applied to three public datasets. Deduplication removed 6.8% of redundant records, while class balancing through downsampling cut the volume by 73%. The experiments show that cleaned datasets yield more reliable experimental conditions. These findings underscore the effectiveness of systematic preprocessing and highlight the need to continually expand and update open datasets to advance Portuguese e-commerce sentiment-analysis research.*

Resumo. *Embora haja acesso a quantidades significativas de datasets, a limpeza e padronização normalmente reduzem a quantidade de instâncias. Este trabalho analisa o impacto de um fluxo de pré-processamento para normalização textual, deduplicação e downsampling, sobre três datasets públicos. A deduplicação removeu 6,8% de instâncias redundantes e o balanceamento por downsampling reduziu o volume em 73%. Os experimentos demonstram que os datasets limpos fornecem condições experimentais mais confiáveis. Os resultados evidenciam a eficácia do pré-processamento sistemático e reforçam a necessidade de ampliar e atualizar continuamente datasets abertos para impulsionar a pesquisa de sentimentos em e-commerce em língua portuguesa.*

1. Introdução

Com o crescente volume de dados gerados pelos usuários da internet, especialmente por meio de redes sociais, blogs e sites de comércio eletrônico, a quantidade de informação textual produzida como opiniões, resenhas, experiências e conhecimento, tornou-se uma fonte valiosa para diversas áreas de pesquisa [Kaplan and Haenlein 2010]. Esses dados oferecem um potencial imenso para gerar insights significativos, possibilitando, por exemplo, a identificação de padrões de comportamento do consumidor, a avaliação da percepção de produtos ou serviços e o suporte à tomada de decisões empresariais.

Ainda que existam diversos *datasets* públicos sendo amplamente utilizados, muitos deles são limitados quanto ao período de coleta, à abrangência temática e à representatividade da linguagem natural em português. Entre os datasets disponíveis, existem com dados de avaliações de filmes, *twitters*, *e-commerce*, entre outros. No contexto deste trabalho, selecionamos os principais *datasets* de avaliações de compras em plataformas de *e-commerce*, disponibilizados pela Olist [Olist 2018], pelo Buscapé por meio do projeto Opinando [Hartmann et al. 2014] e pelo grupo B2W [B2W Digital 2020].

Nesta pesquisa, propomos uma análise dos impactos da aplicação das principais técnicas de pré-processamento, desde a padronização dos *datasets* até os procedimentos de balanceamento. Inicialmente, a Seção 2 discute-se trabalhos relacionados e já na Seção 3 apresentamos a metodologia adotada, contemplando desde a coleta até o pré-processamento dos dados. Na Seção 4, avaliam-se os resultados obtidos e, por fim, na Seção 5, discutem-se as conclusões alcançadas.

2. Trabalhos Relacionados

Os principais *datasets* públicos de avaliações em *e-commerce* brasileiro, citados na Seção 1, são amplamente utilizados para diversos estudos em processamento de linguagem natural. Por exemplo, o trabalho intitulado "Sentiment Analysis on Brazilian Portuguese User Reviews" utilizou-se diversos datasets para análise de sentimento além do contexto de *e-commerce*, aplicando e avaliando diversos modelos de *machine learning* para classificação [Souza and Filho 2021]. Além desse, também há o trabalho intitulado "Lexicon-Based Sentiment Analysis for Reviews of Products in Brazilian Portuguese" que também aplica técnicas para classificação de análise de sentimento [Avanço and Nunes 2014].

Aém disso, o trabalho "RePro: a benchmark for Opinion Mining for Brazilian Portuguese" utilizou o *dataset* para uma anotação manual, sendo um benchmark de 10 mil *reviews* de *e-commerce* em português, retornando a comunidade um *dataset* mais padronizado em relação a critérios para cada nota [dos Santos Silva et al. 2024]. Deste modo, o diferencial com o nosso trabalho proposto é entender o comportamento em cada etapa do pré-processamento e os impactos em relação a quantidade e qualidade do *dataset* após este processo.

3. Metodologia

A metodologia adotada nesta análise baseia-se nas fases do processo de Descoberta de Conhecimento em Banco de Dados [Fayyad et al. 1996]. Na subseção 3.1 descreve-se a seleção dos *datasets* e, na subseção 3.2, definem-se as etapas de pré-processamento e transformação dos dados. Por fim, na Seção 4 analisam-se os resultados obtidos, condensando as fases de mineração de dados, validação e extração de conhecimento previstas pela metodologia.

3.1. Coleta e Seleção de Datasets

Este estudo fundamenta-se em três *datasets* principais para a análise, sendo o foco em *e-commerce*. Logo, os três *datasets* selecionados são de acesso público e amplamente utilizados em tarefas de análise de sentimentos em português sendo da Olist, Buscapé (do projeto Opinando) e B2W Digital.

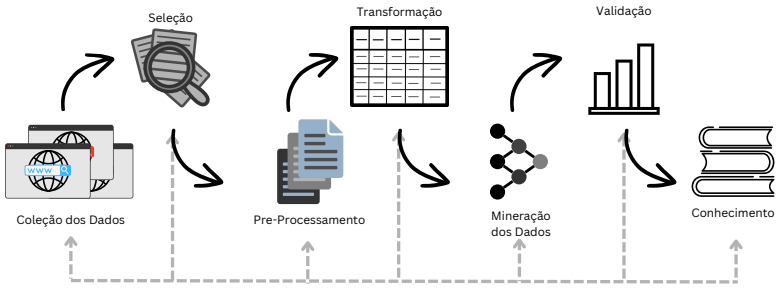


Figura 1. Ilustração do Processo de Descoberta em Conhecimento em Banco de Dados, adaptado de [Fayyad et al. 1996]

Primeiramente, a empresa Olist, uma das principais soluções de comércio eletrônico no Brasil, disponibilizou um *dataset* com aproximadamente 100 mil avaliações de pedidos, abrangendo os anos de 2016 a 2018. Esse *dataset* é proveniente de pequenas lojas parceiras que utilizam a plataforma Olist e inclui informações como status do pedido, preços e outros atributos. No contexto deste estudo, focamo-nos exclusivamente nas avaliações dos clientes.

Ademais, o projeto Opinando[1], proposto pela Universidade de São Paulo (USP), tem como objetivo investigar a análise de sentimentos no contexto brasileiro. Como parte dessa iniciativa, o dataset Buscapé contém aproximadamente 80.000 avaliações [Hartmann et al. 2014].

Por fim, a B2W Digital, um dos maiores grupos varejistas do Brasil, disponibilizou publicamente um *dataset* com aproximadamente 130 mil avaliações de produtos, denominado *B2W-Reviews01*[B2W Digital 2020]. Esse *dataset* contém resenhas submetidas na plataforma de comércio eletrônico Americanas[2], incluindo informações como nome do produto, categoria no site, título da resenha, texto, avaliação (*rating*) e outros atributos relacionados.

3.2. Pré-processamento

Por ser um campo livre de escrita, é possível que avaliações sejam escritas de diversas maneiras, como, por exemplo: todas as palavras maiúsculas, minúsculas, com ou sem acentuação, etc. Dessa forma, a etapa de pré-processamento visa a padronização das sentenças coletadas.

3.2.1. Normalização das sentenças

Em primeiro lugar, no processo de pré-processamento é realizado a transformação de todas as palavras nas sentenças em minúsculas. Essa é uma técnica bastante vantajosa, uma vez que consiste na normalização do texto e redução do vocabulário. Deste modo,

[1]https://sites.google.com/icmc.usp.br/opinando/
[2]https://www.americanas.com.br/

utilizando a linguagem de programação Python[3] e usando a função nativa *lower()*[4], sua aplicação realiza conversão de todas as letras de uma sentença em letras minúsculas.

3.2.2. Unidecode e Pontuações

Nesta etapa, realiza-se o processo de padronização de caracteres, garantindo a representação adequada e simplificada dos caracteres. Sendo assim, este processo consiste em converter caracteres para o formato ASCII, utilizando a biblioteca Unidecode[5] da linguagem Python. Dessa forma, é útil para a normalização do texto, garantindo a padronização de caracteres especiais, acentos ou símbolos. Além da aplicação de Unidecode, também é realizada a remoção das pontuações das sentenças.

3.2.3. Remoção de apenas numéricos ou apenas um carácter

Nessa etapa, são aplicados filtros para remover as sentenças que apresentam apenas números e, consequentemente, não têm valor textual. Deste modo, nativamente a linguagem Python contém o método *isdigit()*[6], validando se só contém numero ou não. Além desse filtro, também é realizado o filtro para eliminar sentenças que possuem apenas um elemento como valor, sendo aplicado o método *len()*[7] e removendo o valor igual a 1.

3.2.4. Processo de Deduplicação

Este procedimento visa manter as sentenças dos *datasets* unificados, conforme os exemplos apresentados na Tabela 1. Dessa forma, para a execução do procedimento de deduplicação dos dados, foi necessária a implementação de uma regra para assegurar a qualidade e unicidade dos dados.

Tabela 1. Exemplo de duplicações das sentenças nos Dataset's

Avaliação	Nota
muito bom	5
muito bom	5
muito bom	4
muito bom	3

Desta forma, para realizar a deduplicação, seguem-se os seguintes procedimentos idealizados para o trabalho:

1. A contagem é realizada para cada sentença e sua nota correspondente.
2. A nota de maior representatividade é selecionada para a sentença.
3. Se houver um empate de representatividade, a nota mais alta é escolhida.

[3]https://www.python.org/
[4]https://docs.python.org/3/library/stdtypes.html
[5]https://pypi.org/project/Unidecode/
[6]https://docs.python.org/3/library/stdtypes.html
[7]https://docs.python.org/3/library/functions.html#len

Conforme os procedimentos descritos anteriormente para o processo de deduplicação, o exemplo da sentença “muito bom” é selecionado para a nota 5, uma vez que é a nota com maior representatividade, segundo a Tabela 2.

Tabela 2. Exemplo de duplicação de sentenças e suas representatividade.

Ranking	Avaliação	Nota	Frequência
1	muito bom	5	2
2	muito bom	4	1
3	muito bom	3	1

3.2.5. Balanceamento dos Dados

A fim de mitigar o desbalanceamento entre classes no *dataset*, é selecionada a técnica de *downsampling*, isto é, subamostragem [Witten et al. 2011]. Essa técnica consiste em reduzir aleatoriamente a quantidade de instâncias das classes majoritárias até igualá-lo ao número de instâncias da classe minoritária, equilibrando assim a distribuição de exemplos entre as classes [He and Garcia 2009, Batista et al. 2004]. Desse modo, sua aplicação evita-se que o algoritmo de aprendizado fique tendencioso em relação às classes com mais exemplos devido ao desbalanceamento, um modelo pode alcançar alta acurácia simplesmente prevendo sempre a classe majoritária, falhando em identificar corretamente instâncias da classe minoritária [He and Garcia 2009].

Ao assegurar quantidades iguais de instâncias para cada classe no treinamento, o modelo é forçado a considerar todas as classes de forma equilibrada, o que tende a melhorar seu desempenho em identificar exemplos da classe menos representada [Batista et al. 2004]. Como ilustrado nas Tabelas 3a e 3b, no *dataset* utilizado a classe B possui a menor quantidade de exemplos; portanto, as classes A e C foram reduzidas por meio da técnica de downsampling para conter o mesmo número de instâncias que a classe B.

Tabela 3. Exemplo de antes e depois da técnica de *downsampling*

Classe	Frequência
A	78
B	39
C	112

(a) Antes

Classe	Frequência
A	39
B	39
C	39

(b) Depois

4. Resultados

O processo de pré-processamento foi dividido em três etapas, sendo a primeira as etapas de remoção de acentos, pontuações, quebras de linhas e normalização das sentenças. Além disso, vale ressaltar que entre as etapas é realizada a remoção de sentenças em branco. Na etapa 2, é realizada a remoção de caracteres únicos, apenas numéricos e normalização da quantidade de espaços entre as palavras. Por fim, na etapa 3 é realizado o processo de deduplicação, conforme mencionado na subseção 3.2.4

O processo de deduplicação resulta em uma redução significativa de sentenças, tendo em vista que do total de sentenças, tivemos uma redução de aproximadamente 7%, conforme ilustrado na Tabela 4. Além disso, vale também ressaltar que a classe que obteve a maior redução entre elas é a da nota 5, tendo em vista uma redução de 10% das sentenças. A classe que contém a menor quantidade de sentenças, isto é, a classe de nota 2 foi a que obteve a menor perda, tendo em vista a redução de 2.56% aproximadamente.

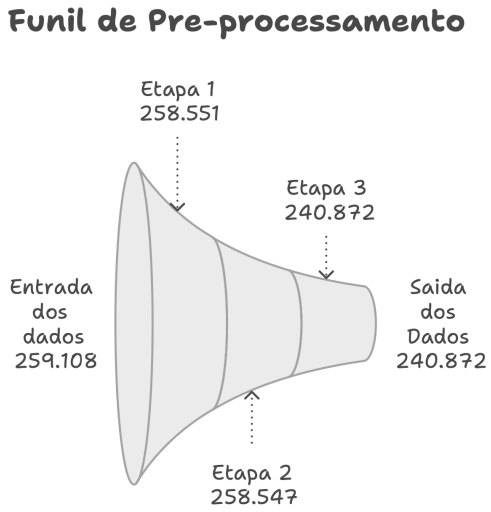


Figura 2. Funil de pré-processamento

Tabela 4. Instâncias antes e depois da deduplicação

Nota	Antes	Depois	Redução
1	39.663	37.653	5.06%
2	14.275	13.910	2.56%
3	31.268	30.285	3.14%
4	71.783	68.014	5.25%
5	101.556	91.010	10.38%
Total	258.545	240.872	6.84%

Além disso, uma das técnicas para o balanceamento de dados utilizadas quando é atuado com o processamento de linguagem natural, é a técnica de *downsample*, mencionada na subseção 3.2.5. Com isto, ao observar na Tabela 4 que a menor classe é a 2, com 13.910 instâncias, todas as outras classes devem ser limitadas a esta quantidade também. Sendo assim, de acordo com a Figura 3, foi obtida uma redução de 73% no total da quantidade de instâncias. Ademais, entre os impactos obtidos é a redução de aproximadamente 85% na classe 5, tendo em vista que a classe menos impactada é a classe 3, sendo 54%.

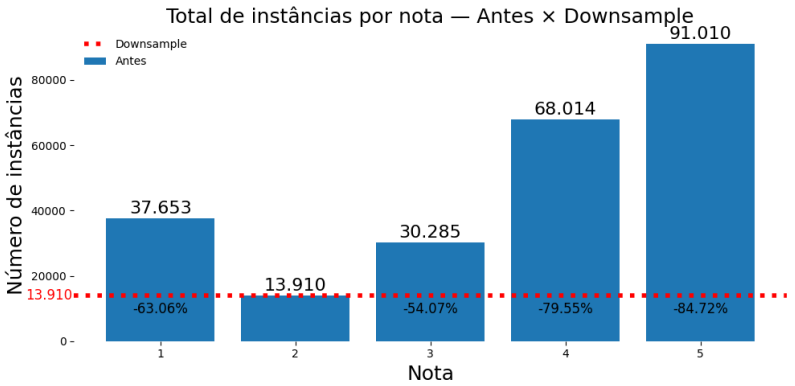


Figura 3. Total de instância por nota antes e depois da técnica de *downsample*

Após a aplicação do fluxo de pré-processamento proposto, analisou-se também o comportamento das três classes do *dataset*: classe 1 (sentimento mais negativo), classe 3 (sentimento neutro) e classe 5 (sentimento de maior satisfação) no contexto de avaliações em plataformas de e-commerce. A árvore de palavras da Figura 4 evidencia o padrão de insatisfação associado à classe 1, especialmente nas ramificações que têm "produto", sendo este o termo mais frequente do corpus, como nó de origem, permitindo observar as construções que ocorrem antes e depois desse termo.

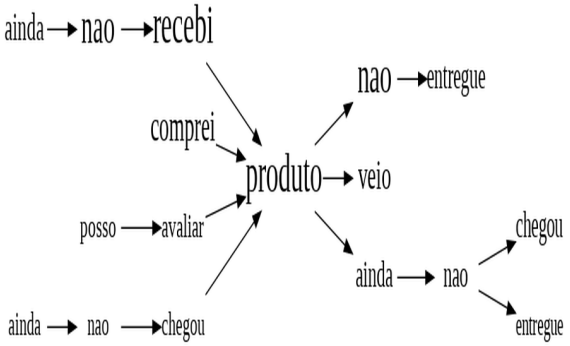
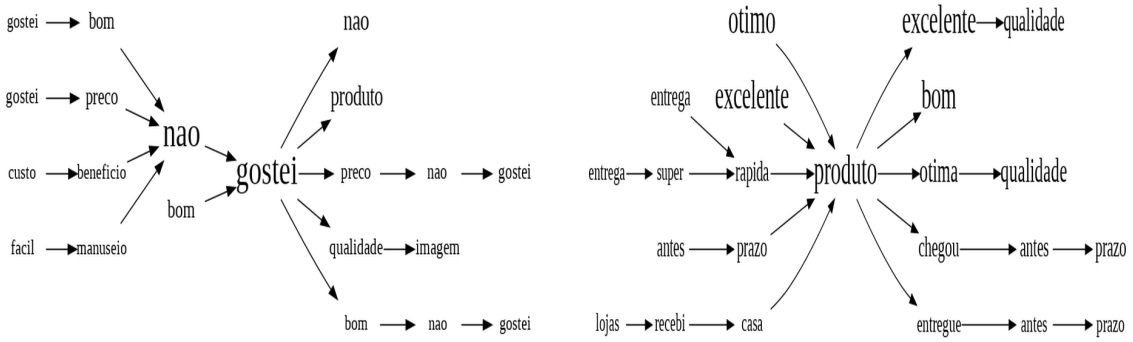


Figura 4. Árvore de Palavras da classe 1

Na classe 3, sendo esta mais neutra, observam-se construções tanto positivas quanto negativas. Na árvore da Figura 5a, o termo "gostei" aparece como central na formação da árvore, formando sequências como "o que não gostei" e "gostei muito do produto". Já a Figura 5b mostra a árvore da classe 5, sendo esta a mais positiva, com a palavra "produto" no centro, surgem composições como "satisfeito", "ótima qualidade" e outras expressões de forte aprovação.



(a) Classe 3 **(b) Classe 5**

Figura 5. Árvores de Palavras das classes 3 e 5.

5. Conclusões

Os resultados obtidos mostram que as escolhas feitas durante o pré-processamento afetam tanto o volume de dados quanto a sua uniformidade. A sequência de normalização, remoção de ruído, deduplicação e balanceamento padronizou o *dataset*. Apesar de esses

tratamentos reduzam o número total de instâncias, elas proporcionam condições experimentais mais imparciais para a avaliação de modelos, sobretudo em contextos originalmente marcados por forte desequilíbrio de rótulos.

Além disso, a expressiva redução de instâncias decorrente do *downsampling* evidencia que, no domínio das avaliações em português de *e-commerce*, os *datasets* ainda apresentam baixa volumetria. Desse modo, ressalta-se a necessidade de ampliar a coleta e o acesso público a dados, tanto no contexto de *e-commerce* quanto em análises de sentimento em geral, a fim de fortalecer a pesquisa nacional e promover o acesso aberto à informação.

6. Agradecimentos

Este trabalho foi financiado pelas seguintes agências de fomento: CAPES, FAPERGS (24/2551-0001396-2) e FAPERGS/CNPq (23/2551-0000126-8).

Referências

Avanço, L. and Nunes, M. (2014). Lexicon-based sentiment analysis for reviews of products in brazilian portuguese. pages 277–281.

B2W Digital (2020). B2w-reviews01: Brazilian e-commerce product reviews dataset.

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29.

dos Santos Silva, L. N., Real, L., Zandavalle, A. C. B., Rodrigues, C. F. G., da Silva Gama, T., Souza, F. G., and Zaidan, P. D. S. (2024). RePro: a benchmark for opinion mining for Brazilian Portuguese. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 432–440, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors (1996). *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, USA.

Hartmann, N., Avanço, L., Balage, P., Duran, M., das Graças Volpe Nunes, M., Pardo, T., and Aluísio, S. (2014). A large corpus of product reviews in Portuguese: Tackling out-of-vocabulary words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3865–3871, Reykjavik, Iceland. European Language Resources Association (ELRA).

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68.

Olist (2018). Brazilian e-commerce public dataset by olist.

Souza, F. and Filho, J. (2021). Sentiment analysis on brazilian portuguese user reviews.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.