MQD-1222: um Dataset de Análise de Sentimentos em Português Brasileiro com Anotação Pareada por Gênero

Alexander Feitosa[1], **André Fasano**[1] **e Gustavo Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca

alexander.feitosa@aluno.cefet-rj.br, andre.cunha@aluno.cefet-rj.br
gustavo.guedes@cefet-rj.br

Resumo. *Este artigo apresenta o MQD-1222, dataset público de Análise de Sentimentos em português brasileiro com 1.222 textos do Meu Querido Diário, anotados segundo protocolo pareado por gênero. Cada texto foi anotado por quatro participantes masculinos e quatro participantes femininas, que fizeram a atribuição de um entre três rótulos de sentimento: negativa, neutra ou positiva. Além do dataset com os rótulos majoritários por grupo, o estudo disponibiliza os 11.704 julgamentos individuais e seus tempos de resposta. A concordância entre os grupos se apresenta na faixa 'substancial' ($\kappa = 0{,}7664$), com coincidência em 84,5% das instâncias. Nas discordâncias, observou-se padrão assimétrico: em 63,5% dos casos, o grupo feminino atenuou julgamentos polares em direção à classe neutra.*

Abstract. *This paper presents MQD-1222, a publicly available Brazilian Portuguese sentiment analysis dataset composed of 1,222 texts from Meu Querido Diário, annotated under a gender-paired protocol. Each text was annotated by four male participants and four female participants, who assigned one of three sentiment labels: negative, neutral, or positive. In addition to the dataset with majority labels for each group, the study provides all 11,704 individual annotations and their response times. Agreement between the two groups fell within the 'substantial' range ($\kappa = 0.7664$), with matching labels in 84.5% of the instances. In cases of disagreement, an asymmetrical pattern was observed: in 63.5% of them, the female group softened polar judgments toward the neutral class.*

1. Introdução

A Análise de Sentimentos (AS) tem sido amplamente investigada no campo do Processamento de Linguagem Natural (PLN), em razão de sua relevância para a identificação automática de polaridades e avaliações em textos [Saha et al., 2021]. Apesar dos avanços obtidos com métodos supervisionados, a robustez desses sistemas permanece fortemente dependente da qualidade dos dados anotados que sustentam seu treinamento.

Pesquisas mais antigas já indicavam a falta de neutralidade nas anotações humanas: o julgamento de um texto carregado de subjetividade reflete a perspectiva de quem o julga [Aroyo and Welty, 2015]. Características sociodemográficas dos anotadores, como o gênero, influenciam de forma sistemática os rótulos atribuídos a textos ambíguos ou emocionalmente carregados [Biester et al., 2022; Frenda et al., 2024]. Esse fenômeno,

denominado *viés de anotação*, representa um risco latente para modelos treinados em datasets construídos sem transparência sobre as características dos grupos de anotadores.

Apesar da crescente atenção ao papel do anotador em tarefas subjetivas de PLN, muitos *benchmarks* ainda são disponibilizados sem informações sobre o perfil sociodemográfico de quem produziu os rótulos [Biester et al., 2022; Pei and Jurgens, 2023]. Essa lacuna é metodologicamente relevante, pois já se demonstrou que diferenças entre anotadores (incluindo gênero, idade, escolaridade e *background* linguístico-cultural) podem afetar a rotulação dos dados e, por extensão, o comportamento dos modelos treinados sobre eles [Ding et al., 2022; Al Kuwatly et al., 2020; Pei and Jurgens, 2023]. Em AS, a interpretação textual envolve julgamentos avaliativos e contextuais. Desse modo, quando datasets públicos não registram essas características, a investigação de vieses sistemáticos entre grupos comparáveis torna-se limitada, assim como a reprodutibilidade de estudos dedicados à análise desses efeitos [Prabhakaran et al., 2021].

Neste contexto, este artigo apresenta o MQD-1222, um *dataset* público de AS em português brasileiro derivado de textos descritivos e anotado com um protocolo pareado por gênero. O diferencial desse *dataset* se estabelece na disponibilização, de forma pioneira, de um corpus público de AS em português do Brasil (PB) com anotação estratificada por gênero. Diante disso, o trabalho oferece três contribuições principais: (i) a disponibilização de um corpus com 1.222 instâncias e rótulos consolidados separadamente para os grupos masculino e feminino; (ii) a liberação dos 11.704 julgamentos individuais, incluindo o tempo de resposta; e (iii) uma análise descritiva do desacordo intergrupos, mostrando que ele não se distribui de forma simétrica.

2. Trabalhos Relacionados

Estudos recentes têm mostrado que a anotação humana, sobretudo em tarefas subjetivas, não deve ser tratada como uma etapa neutra da construção de *datasets*. Geva et al. [2019], por exemplo, mostram que características do anotador podem influenciar sistematicamente os rótulos produzidos, levantando a questão de até que ponto os modelos aprendem o fenômeno de interesse ou regularidades associadas ao próprio processo anotativo. No recorte específico de gênero, Biester et al. [2022] investigam os efeitos do gênero do anotador em diferentes tarefas de PLN e mostram que essa variável pode alterar de forma consistente os padrões de rotulação. Em AS, Ding et al. [2022] demonstram que diferenças demográficas entre anotadores afetam a atribuição de rótulos e podem repercutir inclusive na avaliação de modelos treinados sobre esses dados. Em complemento, Prabhakaran et al. [2021] defendem a preservação de rótulos ao nível do anotador como requisito de transparência metodológica, justamente para permitir a análise de divergências sistemáticas entre grupos. Apesar da relevância desses estudos, sua formulação empírica permanece concentrada, em grande medida, em dados e tarefas em língua inglesa.

Com relação aos *datasets* de Análise de Sentimentos em PB, estes distribuem-se por domínios distintos, como *tweets* [Brum and Volpe Nunes, 2018], relatos pessoais em ambiente digital [Azevedo et al., 2021; Nascimento et al., 2018], avaliações de produtos [dos Santos Silva et al., 2024] e conteúdo compartilhado em comunidades do Reddit [Piorino et al., 2025]. Embora esses recursos constituam referências importantes para a área, sua documentação pública privilegia, em geral, aspectos como domínio textual, esquema de rotulação e resultados de *baseline*, sem explicitar, como parte estruturante

do recurso, características sociodemográficas, como por exemplo, o gênero. Essa lacuna restringe investigações sistemáticas sobre viés de anotação em tarefas subjetivas e dificulta a reprodutibilidade de análises centradas no papel do perfil do anotador.

O MQD-1222 foi concebido para enfrentar essa dupla limitação, oferecendo um corpus em português brasileiro com protocolo pareado por gênero e preservação da distribuição de votos de cada grupo como atributos nativos do *dataset*.

3. O Dataset MQD-1222

3.1. Corpus de origem

O corpus textual do MQD-1222 deriva do MQD-1465, *dataset* em português brasileiro disponibilizado por Azevedo et al. [2021] e construído a partir de textos extraídos do *Meu Querido Diário*, plataforma digital marcada por relatos pessoais e, consequentemente, por elevada carga subjetiva [Nascimento et al., 2018]. A escrita informal e espontânea dos dados originais o torna particularmente desafiador para tarefas de AS.

Este corpus passou por um *pipeline* para a remoção de registros duplicados, padronização textual e randomização com semente fixa (`seed = 42`) para garantir a reprodutibilidade. O corpus foi, então, particionado em dez blocos sequenciais de até 150 frases enviadas para a etapa de rotulagem.

3.2. Protocolo de coleta

A coleta dos julgamentos foi realizada na plataforma PCIbex Farm [Zehr and Schwarz, 2018], que permite a condução de experimentos psicolinguísticos online com registro automático do tempo de resposta. O grupo de participantes foi formado por adultos entre 40 e 50 anos, moradores da cidade do Rio de Janeiro, com nível superior completo, recrutados por conveniência através de redes sociais e contatos pessoais, que acessaram a tarefa de forma autônoma e assíncrona, sem contato com as avaliações dos demais.

Após a autodeclaração de gênero no início da sessão, os participantes foram anonimizados por meio do *hash* MD5 do endereço IP e passaram à rotulagem dos textos do *dataset* MQD-1465 em três classes de sentimento: *negativa*, *neutra* ou *positiva*. Participaram da anotação 46 indivíduos (26 do grupo masculino e 20 do grupo feminino).

3.3. Consolidação por maioria simples

Concluída a etapa de coleta, cada texto passou a reunir quatro julgamentos de participantes masculinos e quatro de participantes femininas. Para cada grupo, o rótulo atribuído foi definido pela classe mais frequente, contemplando os casos de unanimidade (*e.g.,* 4-0-0; 0-4-0), maioria qualificada (*e.g.,* 3-1-0) e maioria simples de uma classe sobre as demais (*e.g.,* 2-1-1). Quando ocorria empate entre duas classes (*e.g.,* 2-2-0), o texto era descartado naquele conjunto.

O *dataset* final foi composto apenas pelas instâncias em que houve definição de um rótulo majoritário em `ambos` os grupos de resposta, descartando os textos em que houve empate em pelo menos um dos grupos. Desse processo resultaram 1.222 instâncias com rótulo majoritário consolidado nos dois grupos, o que resultou no *dataset* MQD-1222 (`mqd-1222.csv`).

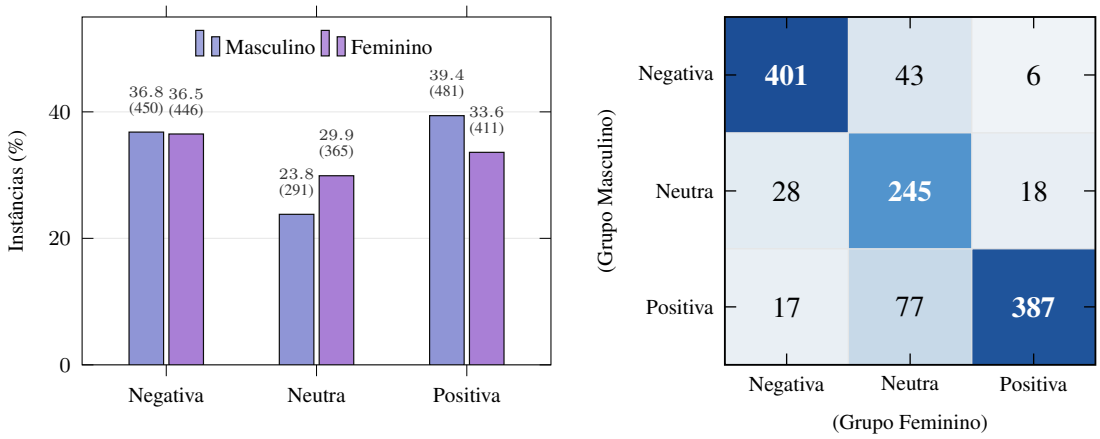
3.4. Estrutura e atributos

O arquivo principal `mqd-1222.csv` é separado por tabulação, codificado em UTF-8, contém 1.222 linhas de dados e um cabeçalho com 16 atributos. Para cada instância, foram registrados o texto original da frase, o rótulo majoritário com a distribuição de votos por classe para cada grupo de gênero (masculino e feminino), a duração média dos julgamentos por grupo e um indicador binário de concordância entre os grupos.

3.5. Análise descritiva

As frases do corpus apresentam um comprimento médio de 18,1 tokens e 97 caracteres, com mediana de 16 tokens, o que reforça o caráter breve e fragmentado da escrita informal. A concentração de 42,0% das instâncias na faixa entre 11 e 20 palavras, somada ao fato de que apenas 3,0% dos textos ultrapassam 40 palavras, indica um conjunto predominantemente composto por enunciados curtos.

Com relação à distribuição dos rótulos, a Figura 1(a) mostra que a diferença mais expressiva entre os grupos se concentra na classe neutra, atribuída pelo grupo feminino com uma frequência 25,4% superior à observada no grupo masculino (365 contra 291 instâncias).



(a) Distribuição por gênero do participante.

(b) Concordância entre rótulos majoritários.

Figura 1. Distribuição das classes e concordância entre grupos no MQD-1222 ($n = 1.222$). Em (a), apresentam-se os percentuais e as frequências absolutas das classes *negativa*, *neutra* e *positiva* nos grupos masculino e feminino. Em (b), mostra-se a matriz de concordância entre os rótulos majoritários, com o grupo feminino no eixo x e o grupo masculino no eixo y.

Conforme mostra a Figura 1(b), entre as 189 discordâncias observadas, 63,5% correspondem a casos em que o grupo feminino desloca julgamentos polares em direção à classe neutra. Esse comportamento sugere que as divergências entre os grupos não se distribuem aleatoriamente, mas se organizam em torno de um padrão sistemático de atenuação da polaridade. Para verificar se essa direção predomina além do esperado ao acaso, comparam-se os 120 casos em que o grupo feminino neutraliza um julgamento polar do grupo masculino aos 46 casos no sentido inverso. O teste de McNemar confirma que a atenuação não se distribui de forma simétrica ($\chi^2 \approx 33{,}0$; $p < 0{,}001$).

Apesar desse padrão assimétrico de desacordo, a concordância global entre os rótulos majoritários permaneceu elevada, alcançando 84,5%, como mostra a Tabela 1. Quando se calcula a concordância esperada ao acaso por meio do coeficiente κ de Cohen, obtém-se $\kappa = 0{,}7664$ (IC 95%: [0,7351;0,7955]), valor que se insere na faixa de concordância *substancial* segundo a interpretação de Landis and Koch [1977]. Em outras palavras, a convergência entre os grupos permanece alta mesmo após o desconto da parcela de concordância que poderia ocorrer casualmente. O fato de todo o intervalo de confiança permanecer dentro dessa mesma faixa reforça a estabilidade dessa interpretação.

Tabela 1. Perfil descritivo do MQD-1222 por grupo de gênero.

Métrica	
Concordância entre grupos	
Instâncias concordantes	1.033 (84,5%)
Cohen's κ	0,7664 [0,7351 ; 0,7955]
Cramér's V	0,7679 ($\chi^2 = 1441{,}03$, $p \approx 0$)

4. Disponibilidade e Reprodutibilidade

A abordagem proposta segue as recomendações de Prabhakaran et al. [2021] para transparência em recursos de anotação, propiciando as condições necessárias para que estudos de viés possam ser replicados e estendidos por terceiros. O MQD-1222 está publicamente disponível[1] sob a licença Creative Commons Attribution 4.0 International (CC BY 4.0).

O repositório contém dois arquivos de dados. O arquivo principal, `mqd-1222.csv`, reúne as 1.222 instâncias com rótulos majoritários e distribuição de votos por grupo, conforme descrito na Seção 3.4. O arquivo complementar `mqd-11704.csv` disponibiliza os 11.704 julgamentos individuais com o identificador anonimizado do participante, o rótulo atribuído e o tempo de resposta em segundos.

Esse arquivo corresponde ao conjunto balanceado por gênero após a etapa de filtragem, contendo 1.463 textos com quatro julgamentos masculinos e quatro femininos por texto. Desses, 9.776 julgamentos correspondem diretamente às 1.222 instâncias finais do `mqd-1222.csv`; os demais 1.928 julgamentos referem-se a 241 textos descartados da versão final por empate em pelo menos um dos grupos.

Os metadados preservados no nível do julgamento individual constituem, adicionalmente, um recurso para pesquisas sobre o comportamento de anotadores em tarefas de rotulagem, linha de investigação discutida em Mostafazadeh Davani et al. [2022] no contexto de modelos que incorporam a distribuição de votos como sinal de supervisão.

5. Considerações Éticas e Limitações

A coleta de dados foi aprovada pelo Comitê de Ética em Pesquisa (CAAE nº 82267824.8.0000.5289), em conformidade com as diretrizes nacionais para pesquisa envolvendo seres humanos. A participação foi voluntária e condicionada à leitura e aceite de termo de consentimento apresentado no início de cada sessão na plataforma PCIbex

[1] https://zenodo.org/records/19209781

Farm. Cada anotador foi identificado por um *hash* MD5 do IP, sem vínculo a dados pessoais; o gênero autodeclarado serviu exclusivamente como critério de estratificação e não foi combinado com outras variáveis para reidentificação.

Quanto ao alcance do estudo, três limitações devem ser registradas. A primeira é o tratamento do gênero como variável binária, que não contempla identidades não-binárias. A segunda diz respeito à composição do grupo de anotadores, homogêneo quanto à faixa etária (40 a 50 anos), à escolaridade (nível superior completo) e à origem geográfica (cidade do Rio de Janeiro); essa homogeneidade limita a validade externa dos padrões observados e recomenda cautela ao transpô-los para anotadores de outros perfis. A terceira é a procedência do corpus de um único domínio (diários pessoais informais), o que restringe a generalização dos achados para outros contextos textuais.

6. Conclusão

Este artigo apresentou o MQD-1222, um *dataset* de Análise de Sentimentos em português brasileiro composto por 1.222 instâncias extraídas do *Meu Querido Diário* e anotadas segundo um protocolo pareado por gênero. Cada texto recebeu quatro julgamentos independentes de participantes masculinos e quatro de participantes femininas. Além dos rótulos majoritários, o recurso disponibiliza os 11.704 julgamentos individuais, acompanhados de metadados de tempo de resposta. Dessa forma, o MQD-1222 não apenas amplia a oferta de corpora públicos para o português brasileiro, mas também incorpora, como parte nativa do *dataset*, informações normalmente suprimidas na agregação final dos rótulos.

Os resultados mostram que, embora a concordância global entre os grupos tenha sido substancial ($\kappa = 0{,}7664$), as discordâncias remanescentes seguem um padrão sistemático. Em 63,5% dos casos de desacordo, o grupo feminino atenuou julgamentos polares em direção à classe neutra, um comportamento consistente com a maior incidência dessa classe em suas anotações.

Esses achados sustentam a importância de tratar a anotação como uma variável analítica relevante em tarefas subjetivas de PLN. As análises aqui reportadas têm caráter descritivo-exploratório; análises inferenciais e de modelagem sobre a distribuição de votos constituem desdobramentos previstos para trabalhos futuros. Ao preservar a distribuição dos votos e os metadados associados ao processo de rotulagem, o MQD-1222 oferece uma base empírica para estudos sobre viés de anotação, comparação de classificadores sob supervisão estratificada por grupo demográfico e investigações sobre concordância interanotador em português. Tais usos devem, contudo, considerar o perfil restrito do grupo de anotadores empregado, descrito na Seção 5.

Referências

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In Akiwowo, S., Vidgen, B., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Azevedo, G. d., Pettine, G., Feder, F., Portugal, G., Schocair Mendes, C. O., Castaneda Ribeiro, R., Mauro, R. C., Paschoal Júnior, F., and Guedes, G. (2021). Nat: Towards an emotional agent. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–4.

Biester, L., Sharma, V., Kazemi, A., Deng, N., Wilson, S., and Mihalcea, R. (2022). Analyzing the effects of annotator gender across 4 NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19. European Language Resources Association.

Brum, H. and Volpe Nunes, M. d. G. (2018). Building a sentiment corpus of tweets in Brazilian Portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ding, Y., You, J., Machulla, T.-K., Jacobs, J., Sen, P., and Höllerer, T. (2022). Impact of annotator demographics on sentiment dataset labeling. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

dos Santos Silva, L. N., Zandavalle, A. C., Rodrigues, C. F. G., da Silva Gama, T., Souza, F. G., Zaidan, P. D. S., da Silva, A. F. S., Soares, K., and Real, L. (2024). RePro: A benchmark dataset for opinion mining in 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.

Frenda, S., Basile, V., Caselli, T., and Patti, V. (2024). Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*.

Geva, M., Goldberg, Y., and Berant, J. (2019). Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mostafazadeh Davani, A., Díaz, M., and Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Nascimento, G., Duarte, F., and Guedes, G. P. (2018). Emoções em português do brasil: um conjunto de dados e resultados de base. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, pages 223–228, Porto Alegre, RS, Brasil. SBC.

Pei, J. and Jurgens, D. (2023). When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In Prange, J. and Friedrich, A., editors, *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Piorino, A. et al. (2025). Sentiment analysis of shared content in Brazilian Reddit communities. *Journal on Interactive Systems*.

Prabhakaran, V., Mostafazadeh Davani, A., and Diaz, M. (2021). On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138. Association for Computational Linguistics.

Saha, K., Yousuf, A., Hickman, L., Gupta, P., Tay, L., and De Choudhury, M. (2021). A social media study on demographic differences in perceived job satisfaction. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):167.

Zehr, J. and Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). OSF.