ConsumerBR: A Large-Scale Corpus of Consumer Complaints in Brazilian Portuguese

Luis A. Duarte[1], Pedro Giacomin[1], Vitória Bispo[1], Mariana O. Silva[1], Adriano C. M. Pereira[1], Gisele L. Pappa[1]

[1]Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
luisantonio@dcc.ufmg.br, plsg2020@ufmg.br,
{vitoriabispo,mariana.santos,adrianoc,glpappa}@dcc.ufmg.br

Abstract

We present **ConsumerBR**, a large-scale corpus of consumer complaints and company responses in Brazilian Portuguese, compiled from publicly available data on the *Consumidor.gov.br* platform. The corpus comprises over 3.1 million consumer–company interactions collected between 2021 and 2025 and combines anonymized textual content with rich structured metadata, including temporal information, complaint outcomes, and consumer satisfaction indicators. We describe a data collection strategy tailored to the platform's dynamic interface, a preprocessing pipeline that includes response clustering to identify template-based replies, and a hybrid anonymization approach designed to mitigate privacy risks. We also provide a detailed statistical characterization of the corpus, highlighting its scale, coverage, and distributional properties. **ConsumerBR** is publicly available for research purposes and supports a wide range of applications, including complaint analysis, sentiment modeling, dialogue and response generation, and preference-based evaluation.

1 Introduction

Large-scale, publicly documented textual resources are a fundamental requirement for the development, evaluation, and comparison of natural language processing (NLP) methods (Charles et al., 2022; Garcia et al., 2022; Hartmann et al., 2014). For Brazilian Portuguese in particular, the availability of corpora that capture real-world institutional interactions remains limited, especially in domains involving formal and semi-formal written communication between citizens and organizations (Cardoso et al., 2021; Souza et al., 2016). This scarcity constrains reproducible research and hinders the development of language technologies tailored to socially relevant domains.

In Brazil, digital platforms for consumer dispute resolution constitute one of the largest sources of naturally occurring written interactions between individuals and companies (Silva et al., 2024; Oliveira et al., 2025; Pagano et al., 2022). These platforms allow consumers to report problems, describe the circumstances of their complaints, and receive responses from suppliers within a regulated public environment. Despite their scale and relevance, such platforms have not yet resulted in openly documented large-scale corpora suitable for research in Portuguese NLP.

Consumidor.gov.br[1] is a prominent example of this type of platform, enabling direct online dialogue between consumers and companies for the resolution of consumption-related disputes (Oliveira et al., 2025; Pagano et al., 2022). The service is monitored by the *Secretaria Nacional do Consumidor* (National Secretariat for Consumer Affairs – Senacon), the Ministry of Justice, Procons, public defenders, and other civic institutions. According to official statistics, the platform aggregates approximately 9 million finalized complaints, nearly 7 million registered users, and more than 1,600 participating companies, making it one of the largest repositories of consumer–company interactions in Brazilian Portuguese.

However, access to the textual content of these interactions remains largely unavailable in a structured and documented form for research purposes (Silva et al., 2024; Pagano et al., 2022). Existing Portuguese-language resources in related domains typically focus on restricted datasets, synthetic or simulated interactions, or small manually annotated collections, which fail to reflect the linguistic diversity, scale, and variability observed in real consumer disputes (Hartmann et al., 2014; Souza et al., 2016; Zhou and Ganesan, 2016). As a result, there is a clear gap in the availability of large-scale, domain-specific corpora that can support empirical studies, benchmark downstream tasks, and enable

[1]https://www.consumidor.gov.br/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 667–675
April 13-16, 2026 ©2026 Association for Computational Linguistics

the training and evaluation of language models for Brazilian Portuguese.

This work introduces **ConsumerBR**, a large-scale corpus of consumer complaints and company responses in Brazilian Portuguese, constructed from data collected directly from the *Consumidor.gov.br* platform. We detail the corpus creation process and provide a comprehensive characterization of the resulting resource. In addition, we describe a carefully designed data collection strategy tailored to the platform's dynamic content loading, along with a fully documented preprocessing pipeline that includes text extraction, normalization, template-based detection of automated responses, and anonymization of personally identifiable information (PII). The resulting dataset combines anonymized textual content with rich structured metadata, enabling a wide range of downstream research tasks in Portuguese NLP.

Our contributions are threefold: (i) the release of a large-scale, structured corpus of consumer complaints and company responses in Brazilian Portuguese; (ii) a fully documented data collection and preprocessing methodology, including response clustering and PII anonymization; and (iii) a detailed statistical and linguistic characterization of the corpus, illustrating its structure, coverage, and potential use cases for Portuguese NLP research.

2 Related Work

Research on consumer complaints and opinionated user-generated content has been explored in Portuguese and other languages from different perspectives, including sentiment analysis, complaint classification, and sociolinguistic analysis. However, the availability of large-scale, openly documented corpora containing real-world consumer–company interactions in Portuguese remains limited. In this section, we position **ConsumerBR** with respect to existing datasets and studies, highlighting differences in domain coverage and resource availability.

Although several studies have analyzed data from the *Consumidor.gov.br* platform, these efforts have mainly focused on applied analyses rather than on corpus creation. Silva et al. (2024), for instance, investigate sentiment analysis of consumer complaints using transformer-based models, relying on a manually labeled subset for evaluation. Oliveira et al. (2025), in turn, conduct a large-scale quantitative analysis of complaints, focusing primarily on structured metadata and aggregated statistics rather than on the underlying textual content.

Linguistic analyses have also been conducted on subsets of data collected from the platform. Pagano et al. (2022), for example, explored gender-related patterns in the language of consumer complaints using NLP techniques. While these studies demonstrate the analytical value of *Consumidor.gov.br*, none provide a publicly available, large-scale textual corpus suitable for reuse, comparison, and benchmarking in Portuguese NLP research.

Other Portuguese-language resources related to opinion mining and complaint analysis are typically derived from product reviews or social media data (Souza et al., 2016; Cardoso et al., 2021). For instance, Hartmann et al. (2014) introduce a large corpus of Brazilian Portuguese product reviews collected from e-commerce websites, focusing on lexical variation and out-of-vocabulary phenomena. Despite its size, this dataset consists mainly of short, informal texts and does not capture dialogic interactions between consumers and organizations.

Beyond the complaint domain, several benchmark corpora have been proposed for Portuguese NLP tasks. Fakepedia (Charles et al., 2022) and FakeRecogna (Garcia et al., 2022) introduce datasets for fake news detection, emphasizing balanced collections, reproducibility, and update mechanisms. Although these works exemplify best practices in corpus construction and documentation, they focus on news articles rather than on consumer–company interactions.

Large-scale datasets for sentiment analysis and user behavior prediction have also been explored in other languages, such as user reviews (Zhou and Ganesan, 2016; Yang, 2025), customer complaint mining (Ghazzawi and Alharbi, 2019), and purchase behavior prediction on social media (Sakaki et al., 2016). However, these resources are either unavailable in Portuguese or originate from interaction settings that differ substantially from formal consumer dispute-resolution platforms.

In contrast to existing studies and datasets, **ConsumerBR** focuses explicitly on the construction and characterization of a large-scale corpus of consumer complaints and company responses in Brazilian Portuguese, derived from a public governmental platform. By prioritizing scale, textual completeness, and rich metadata, the corpus fills a critical gap in the ecosystem of Portuguese NLP resources and enables a wide range of downstream research tasks, including complaint understanding, sentiment analysis, and longitudinal analyses of con-

sumer–company interactions.

3 Data Collection

The *Consumidor.gov.br* platform provides multiple public interfaces for accessing information on consumer disputes. Among them, the "Indicadores" panel[2] presents aggregated statistics such as company rankings, resolution rates, and average response times, and includes a "Dados abertos" section[3] for downloading complaint records. However, this channel only provides structured metadata and does not include the textual content of consumer complaints or company responses.

Access to full textual interactions is available exclusively through the "Relato do Consumidor" section,[4] which displays individual complaints and their corresponding company responses. This interface supports filtering by attributes such as company, topic, problem category, region, and satisfaction score, and presents the consumer narrative, the company response (when available), the complaint status, the consumer evaluation, and temporal and location metadata.

Complaints are listed in reverse chronological order and loaded incrementally via an infinite-scroll mechanism, in which only a limited number of entries are initially displayed, and additional items are retrieved dynamically. Because this mechanism does not rely on explicit pagination, the total number of available complaints cannot be directly inferred, and no URL parameters are exposed to enable straightforward page-based crawling. As a result, conventional automated crawling strategies are not applicable and require a dedicated data extraction approach.

Given the absence of textual data in the "Dados abertos" section, data collection is performed through the "Relato do Consumidor" interface. Rather than relying on browser automation, we identified and directly accessed the underlying endpoint used by the public web interface to retrieve complaint records.[5] This strategy enables a more stable and efficient extraction process while preserving the structure of the returned data. Although the endpoint supports optional filtering parameters, the collection procedure retrieves all available complaints by default, and each request includes a timestamp parameter to prevent cached responses.

The endpoint returns HTML fragments containing rendered complaint entries. These fragments are parsed using the `BeautifulSoup` library[6] to extract the relevant fields. The extracted attributes include company name, complaint status, full consumer narrative, submission date and location, company response and response date (when available), and the consumer's final evaluation, consisting of a numerical rating and an optional comment.

Because the platform does not provide conventional pagination, the extraction process simulates the infinite scrolling behavior by iteratively adjusting the parameter `indicePrimeiroResultado`, which controls the starting index of each batch. Each request retrieves 10 complaints at a time, and the process continues until either an empty batch or a duplicate batch is detected. To ensure robustness and avoid server-side blocking, the extraction includes request throttling, duplicate detection across batches, and explicit handling of HTTP 429 responses via cooldown periods.

Ethical and Legal Considerations. The *Consumidor.gov.br* platform operates under official Terms of Use and data protection policies that define which information is publicly accessible and how personal data must be handled. While complaint narratives, company responses, and final consumer evaluations are publicly visible through the platform interface, personally identifiable information (PII), such as names, identification numbers and contact details, is not publicly disclosed.

Although the platform's Terms of Use and Data Policy do not explicitly address or prohibit automated access to publicly available content, they clearly distinguish between public and confidential data and restrict the use of personal information to legally authorized contexts. In compliance with these policies, the data collection process employed in this work retrieves only publicly visible information from the interface and does not access restricted content or bypass authentication mechanisms. No data outside the scope of publicly displayed fields is accessed or collected.

The construction of the **ConsumerBR** corpus

[2] https://www.consumidor.gov.br/pages/indicador/geral/abrir

[3] https://www.consumidor.gov.br/pages/dadosabertos/externo/

[4] https://www.consumidor.gov.br/pages/indicador/relatos/abrir

[5] Although the platform does not provide an official public API for accessing textual complaint data, the identified endpoint corresponds to the same data retrieval mechanism employed by the public interface.

[6] https://beautiful-soup-4.readthedocs.io/en/latest/

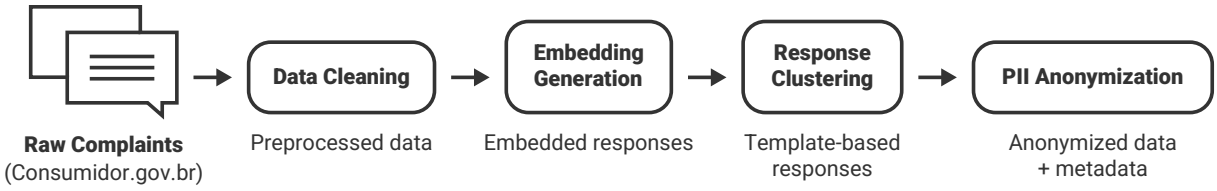

Figure 1: Overview of the **ConsumerBR** data processing pipeline.

adheres to these platform policies and to the principles established by the Brazilian General Data Protection Law (LGPD). Data collection is limited to publicly accessible content, and only anonymized versions of the textual data are released for research use. The preprocessing pipeline incorporates anonymization procedures to mitigate the risk of disclosing sensitive or identifying information, and no attempts are made to re-identify individuals. Although derived from a public interface, the corpus is intended exclusively for research purposes and responsible use.

4 Data Processing

The data processing pipeline is designed to transform raw complaint records into a structured, anonymized, and analysis-ready corpus. The pipeline consists of four main steps: (i) data cleaning, (ii) embedding generation, (iii) clustering of company responses, and (iv) anonymization of personally identifiable information (PII). Figure 1 provides an overview of the complete workflow.

4.1 Data Cleaning

The first step focuses on cleaning and standardizing the raw data collected from the platform. Duplicate complaint records are removed, and missing or incomplete fields are handled in accordance with predefined consistency rules. Textual fields, including company names, consumer complaints, and company responses, are minimally preprocessed to normalize whitespace, remove encoding artifacts, and ensure consistent formatting across records. Each complaint is assigned a unique internal identifier, enabling traceability across subsequent stages of the pipeline.

4.2 Embedding Generation

To support semantic comparison between company responses, each response is converted into a dense vector representation. We rely on the Serafim 100M Portuguese (PT) Sentence Encoder,[7] a sentence transformer model trained for Brazilian Portuguese (Gomes et al., 2024). Embeddings are generated in batches to ensure computational efficiency and stored in HDF5 format, along with linked metadata, such as complaint identifiers and company names. The resulting representations map each response to a fixed-dimensional vector space, enabling semantic similarity computation using cosine similarity.

4.3 Response Clustering

The clustering step aims to identify standardized or template-based responses reused by companies across multiple complaints. Here, responses are grouped by company, and clustering is applied independently within each company-specific subset. Such a methodology decision reflects the observation that response templates are typically specific to individual organizations.

Clustering is performed using the DBSCAN algorithm over normalized embeddings, with cosine distance as the similarity metric. DBSCAN is particularly suitable for this task, as it does not require specifying the number of clusters in advance and naturally identifies outliers. Under this formulation, dense clusters correspond to frequently reused boilerplate responses, while isolated points represent more individualized replies.

As a result, each response is assigned a cluster identifier, which can be used to filter template-based messages or to analyze response patterns at the company level. This information is preserved as metadata in the processed dataset. The distinction between template-based and non-template responses is therefore derived from unsupervised clustering, rather than from manual annotation.

4.4 PII Anonymization

The final step of the pipeline addresses privacy concerns by detecting and anonymizing personally identifiable information (PII) present in the

[7] https://huggingface.co/PORTULAN/serafim-100m-portuguese-pt-sentence-encoder

textual fields. Anonymization follows a hybrid approach that combines rule-based detection with model-based named entity recognition.

Structured PII is identified using regular expressions targeting patterns such as email addresses, phone numbers (including toll-free numbers), dates, URLs, numerical identifiers, and already masked values. In parallel, unstructured entities are detected using a Transformer-based NER model fine-tuned for Portuguese,[8], which identifies personal names and organization names (Souza et al., 2019).

Detected entities are merged and passed through a post-processing step that filters false positives using external dictionaries of common Portuguese names[9] and the list of companies present in the dataset metadata. Overlapping entities are resolved according to a predefined priority hierarchy, ensuring consistent labeling.

In the final labeling step, all validated PII spans are replaced with standardized placeholder tags (e.g., `[EMAIL]`, `[NOME]`, `[ORGANIZACAO]`, `[MASCARADO]`). This process preserves the linguistic structure of the text while preventing the disclosure of sensitive information, resulting in anonymized versions of the complaints and responses suitable for downstream NLP tasks and public corpus release.

5 ConsumerBR Corpus

This section presents the structure and a detailed characterization of the **ConsumerBR** corpus. We describe the organization of the dataset and its constituent fields, followed by a statistical and exploratory analysis of its main textual and metadata attributes. The characterization focuses on corpus-level properties and distributional patterns relevant for NLP research, without exposing sensitive individual-level information.

5.1 Structure

Table 1 summarizes the schema of the **ConsumerBR** corpus. The dataset is available as a single CSV file of approximately 5.2 GiB, containing 3,150,552 records and 12 fields. Each record corresponds to a unique consumer complaint and its associated metadata, including anonymized textual content, temporal information, complaint status, and optional consumer evaluation fields.

[8] https://huggingface.co/marquesafonso/bertimbau-large-ner-selective

[9] https://brasil.io/dataset/genero-nomes/nomes/

Table 1: Schema of the **ConsumerBR** corpus.

Field	Type	Description
`ID`	str	Unique identifier
`id_reclamacao`	int	Complaint identifier
`empresa`	str	Company name
`reclamacao_labeled`	txt	Anonymized complaint text
`resposta_labeled`	txt	Anonymized response
`padronizada`	bool	If is template-based response
`status`	cat	Complaint status
`data_abertura`	ts	Submission timestamp
`localizacao`	str	Consumer-reported location
`data_resposta`	ts	Response timestamp
`avaliacao_nota`	int	Evaluation score (1–5)
`avaliacao_comentario`	txt	Evaluation comment

Table 2: Global statistics of the **ConsumerBR** corpus.

Statistic	Value
Total number of complaints	3,150,552
Time span	05/2021 – 04/2025
File size	5.2 GiB
Number of companies	1,445
Average complaints per company	2,180
Median complaints per company	131
Max complaints (single company)	131,381
Complaints with company response	3,090,951 (98.11%)
Complaints without response	59,601 (1.89%)
Complaints with evaluation score	1,463,328 (46.5%)
Complaints without evaluation score	1,687,224 (53.6%)
Complaints with evaluation comment	1,063,124 (33.7%)
Mean complaint length	772.27 (chars)
Median complaint length	567.00 (chars)
Mean response length	587.76 (chars)
Median response length	488.00 (chars)
Template-based responses	2,907,712 (92.29%)
Non-template responses	242,840 (7.71%)

The released version of the corpus provides exclusively anonymized textual fields alongside structured metadata, enabling large-scale linguistic analysis while mitigating privacy risks. Raw, non-anonymized textual content is retained internally for corpus construction and validation purposes but is not distributed as part of the public release.

5.2 Characterization

Table 2 reports the main descriptive statistics of the **ConsumerBR** corpus, which spans complaints submitted between May 2021 and April 2025 and comprises over 3.1 million consumer–company interactions involving 1,445 distinct companies.

Data completeness is high for the core interaction fields: 98.11% of complaints include a company response, while only 1.89% remain unanswered. In contrast, evaluation-related fields show substantial sparsity, reflecting the optional nature

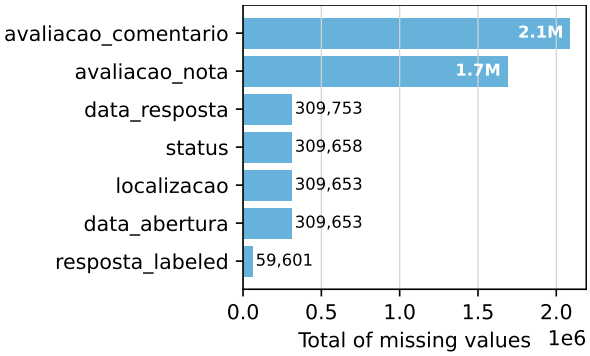

Figure 2: Distribution of missing values across fields.

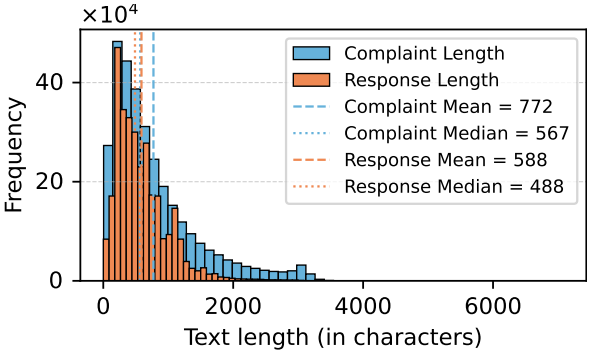

Figure 3: Distribution of text lengths for complaints and company responses.

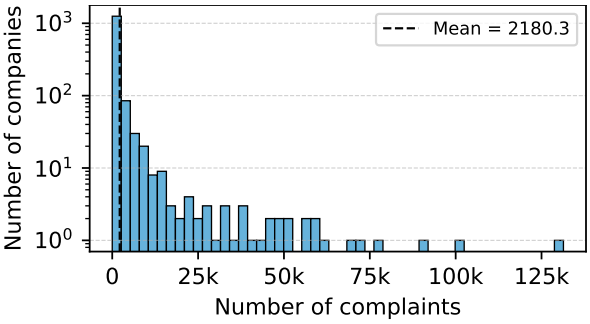

Figure 4: Distribution of complaints per company.

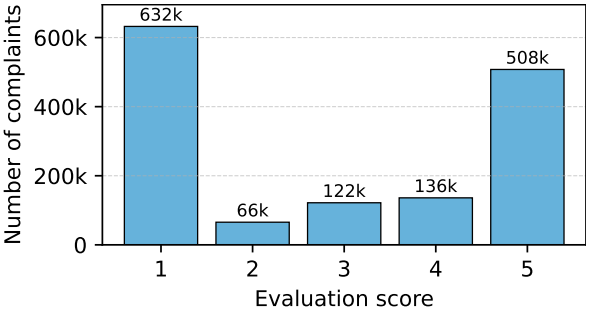

Figure 5: Distribution of consumer satisfaction ratings.

of consumer feedback on the platform. Approximately 46.5% of complaints include an evaluation score, and 33.7% include an evaluation comment.

An analysis of missing values (Figure 2) reveals systematic patterns induced by the platform workflow. Records lacking a company response (*resposta_labeled* and *data_resposta*) correspond to unresolved or unanswered complaints. Missing values in *avaliacao_nota* almost always imply missing values in *avaliacao_comentario*, whereas the inverse does not necessarily hold, indicating that consumers may assign a numerical score without providing textual feedback.

Figure 3 presents the distribution of text lengths for complaints and company responses, measured in number of characters. Complaint narratives exhibit a wide range of lengths, from very short descriptions to detailed reports exceeding several thousand characters, with a mean length of 772 characters and a median of 567 characters. This long-tailed distribution reflects the heterogeneity of consumer writing styles and levels of detail. Company responses are generally shorter and more concentrated, with a mean length of 588 characters and a median of 488 characters, suggesting more standardized communication practices.

The corpus includes complaints directed at 1,445 distinct companies. The distribution of complaints across companies is highly imbalanced and follows a pronounced long-tail pattern (Figure 4). Approximately 76.7% of companies received fewer than 1,000 complaints during the observed period. A percentile-based analysis shows that 75% of companies have at most 909 complaints, whereas the top 5% exceed 8,283 complaints, and the top 1% surpass 47,408 complaints.

Complaint outcomes, captured in the *status* field, are distributed as follows: 28.06% of complaints are marked as resolved, 23.45% as not resolved, and 48.49% as not evaluated. Although half of the complaints lack an explicit final evaluation, the subset of evaluated cases is relatively balanced between resolved and unresolved outcomes, indicating comparable volumes of successful and unsuccessful resolutions when feedback is provided.

Among complaints with an evaluation score, the *avaliacao_nota* field exhibits a strongly polarized distribution, with ratings concentrated at the extremes of the scale (1 and 5), as shown in Figure 5. This pattern suggests that consumers are more likely to provide feedback when their experience is either highly satisfactory or highly unsatisfactory, a behavior commonly observed in voluntary rating systems.

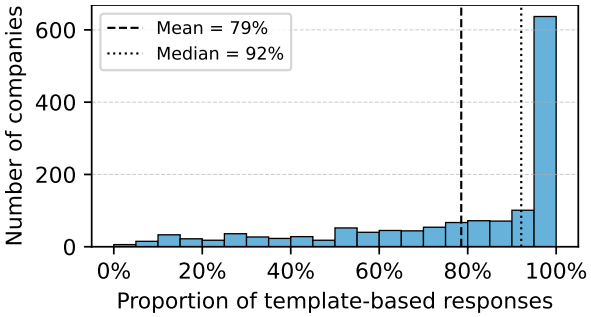

Figure 6: Distribution of the proportion of template-based responses across companies.

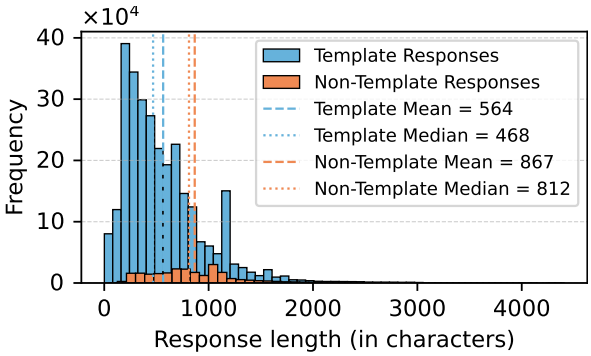

Figure 7: Distribution of text lengths for template-based and non-template responses.

5.3 Template-based vs. Non-template Responses

A salient characteristic of the **ConsumerBR** corpus is the prevalence of template-based responses. As described in Section 4, an unsupervised clustering-based approach is applied during preprocessing to identify responses that follow repetitive textual patterns, typically associated with copy-and-paste practices or automated reply systems.

Overall, 92.29% of all company responses are classified as template-based, while the remaining 7.71% correspond to non-template, more individualized replies. This proportion varies substantially across companies, with some organizations relying almost exclusively on standardized messages, whereas others exhibit a higher proportion of personalized responses (Figure 6).

Template-based responses tend to be considerably shorter and lexically less diverse than non-template responses, as illustrated in Figure 7. On average, template-based responses contain 564 characters, compared to 867 characters for non-template responses. This gap reflects the repetitive, formulaic nature of standardized replies, which typically prioritize procedural information and generic explanations over case-specific details, whereas non-template responses more often incorporate contextualized information and individualized problem resolution.

This distinction has important implications for downstream NLP tasks: while template-based replies may introduce redundancy and bias in language model training, non-template responses provide richer linguistic material for tasks such as response generation and negotiation analysis. By explicitly identifying and labeling standardized responses, the **ConsumerBR** corpus enables controlled experimentation at scale, allowing researchers to include or exclude templated content depending on the target application.

5.4 Data Availability and Access

The **ConsumerBR** corpus is publicly available through the Zenodo repository under a research-only license (Duarte et al., 2025). The dataset includes exclusively anonymized textual content and metadata derived from publicly accessible interfaces of the *Consumidor.gov.br* platform.

To comply with ethical, legal, and contractual constraints, the original non-anonymized textual content and the data collection and preprocessing code are not publicly released. The corpus was constructed as part of a collaborative research project involving academic and industry partners, and its distribution is subject to data governance and intellectual property agreements.

Access to the corpus is restricted to research purposes. The license explicitly prohibits attempts at re-identification, redistribution for non-research purposes, or any use that violates applicable data protection regulations, including the Brazilian General Data Protection Law (LGPD). Users are required to comply with these terms and to cite the dataset appropriately in any resulting publications.

6 Potential Use Cases

The **ConsumerBR** corpus supports a wide range of NLP research tasks by providing large-scale, real-world consumer–company interactions in Brazilian Portuguese. The availability of paired complaints and responses, combined with structured metadata and outcome information, enables applications such as complaint classification, topic modeling, sentiment analysis, and outcome prediction.

The corpus is also suitable for training and adapting language models to the customer dispute and negotiation domain. It can be employed in tasks

such as response generation, complaint summarization, and assistance in drafting replies for human agents. The explicit distinction between template-based and non-template responses further allows controlled experimentation on response personalization and standardization strategies.

Finally, the presence of consumer ratings and resolution status enables research on preference modeling, evaluation, and alignment, as well as studies on fairness and robustness in automated systems that mediate interactions between users and institutions. These characteristics make **ConsumerBR** a versatile resource for both NLP research and applied studies in socially relevant domains.

7 Conclusion

This paper introduces **ConsumerBR**, a large-scale corpus of consumer complaints and company responses in Brazilian Portuguese, constructed from publicly available data on the *Consumidor.gov.br* platform. The corpus comprises more than 3.1 million consumer–company interactions collected over four years and combines anonymized textual content with rich structured metadata, including temporal information, complaint outcomes, and consumer satisfaction indicators. In addition to documenting the dataset structure, we provided a comprehensive statistical and linguistic characterization of its main properties.

A key contribution of this work lies in the design of a data processing pipeline tailored to the challenges of real-world data. In particular, the pipeline incorporates cleaning and standardization procedures, large-scale embedding and clustering to identify template-based responses, and a hybrid anonymization strategy that ensures compliance with ethical and legal constraints. By explicitly distinguishing between standardized and non-template company responses, the corpus enables controlled experimentation and helps mitigate redundancy-driven bias in downstream NLP applications.

Despite its scale and richness, the corpus has some limitations. First, the dataset is restricted to a single platform and reflects the specific interaction dynamics and policies of *Consumidor.gov.br*, which may limit generalization to other consumer dispute settings. Second, although the core textual fields are largely complete, a substantial proportion of complaints lack explicit outcome evaluations, reflecting the voluntary nature of consumer feedback on the platform. Finally, while anonymization procedures reduce privacy risks, they may also remove contextual cues that could be relevant for certain fine-grained linguistic analyses.

Future work includes extending the corpus with new data releases, enriching the metadata with sector-level and complaint-type annotations, and exploring cross-platform extensions to support comparative studies. We also plan to provide additional resources to facilitate corpus access and usability, as well as to define benchmark tasks and evaluation protocols derived from the dataset, further consolidating **ConsumerBR** as a reference resource for research in Brazilian Portuguese NLP.

Acknowledgements

This work was supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Henrique Lopes Cardoso, Tomás Freitas Osório, Luís Vilar Barbosa, Gil Rocha, Luís Paulo Reis, João Pedro Machado, and Ana Maria Oliveira. 2021. Robust complaint processing in portuguese. *Information*, 12(12):525.

Anderson Cordeiro Charles, Lívia Ruback, and Jonice Oliveira. 2022. Fakepedia corpus: A flexible fake news corpus in portuguese. In *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21-23, 2022, Proceedings*, volume 13208 of *Lecture Notes in Computer Science*, pages 37–45. Springer.

Luis Antônio Duarte, Pedro Giacomin, Vitória Bispo, Mariana Silva, Adriano Pereira, and Gisele Lobo Pappa. 2025. Consumerbr: Brazilian portuguese consumer complaint corpus (1.0). [Dataset] Zenodo.

Gabriel Lino Garcia, Luis C. S. Afonso, and João P. Papa. 2022. Fakerecogna: A new brazilian corpus for fake news detection. In *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21-23, 2022, Proceedings*, volume 13208 of *Lecture Notes in Computer Science*, pages 57–67. Springer.

Amani Ghazzawi and Basma Alharbi. 2019. Analysis of customer complaints data using data mining techniques. *Procedia Computer Science*, 163:62–69. 16th Learning and Technology Conference 2019Artificial Intelligence and Machine Learning: Embedding the Intelligence.

Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. 2024. Open sentence embeddings for portuguese with the serafim pt* encoders family. In *Progress in Artificial Intelligence - 23rd EPIA Conference on Artificial Intelligence, EPIA 2024, Viana do Castelo, Portugal, September 3-6, 2024, Proceedings, Part III*, volume 14969 of *Lecture Notes in Computer Science*, pages 267–279. Springer.

Nathan Hartmann, Lucas Avanço, Pedro Paulo Balage Filho, Magali Sanches Duran, Maria das Graças Volpe Nunes, Thiago Alexandre Salgueiro Pardo, and Sandra M. Aluísio. 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 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3865–3871. European Language Resources Association (ELRA).

Leandro Divino Miranda de Oliveira, Marcos Inácio Severo de Almeida, and Jussara Goulart da Silva. 2025. Reclamações online no brasil: Uma análise abrangente da plataforma consumidor.gov.br. *Revista ADMPG*, 14(2).

Adriana Silvina Pagano, Ana Paula Couto da Silva, Wesley Maciel, and Yohan Gumiel. 2022. User profile characterization within a brazilian online dispute resolution platform. In *LatinX in Natural Language Processing Research Workshop*.

Shigeyuki Sakaki, Francine Chen, Mandy Korpusik, and Yan-Ying Chen. 2016. Corpus for customer purchase behavior prediction in social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Daniel Silva, William Betker, Daniel Gonçalves, and Ugo Dias. 2024. Modelos transformers para a análise automática de satisfação na plataforma consumidor.gov.br. In *Anais do XII Workshop de Computação Aplicada em Governo Eletrônico*, pages 1–13, Porto Alegre, RS, Brasil. SBC.

Ellen Souza, Tiago Alves, Ingryd Teles, Adriano L. I. Oliveira, and Cristine Gusmão. 2016. TOPIE: an open-source opinion mining pipeline to analyze consumers' sentiment in brazilian portuguese. In *Computational Processing of the Portuguese Language - 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings*, volume 9727 of *Lecture Notes in Computer Science*, pages 95–105. Springer.

Fábio Souza, Rodrigo Nogueira, and Roberto A. Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *CoRR*, abs/1909.10649.

Yun Yang. 2025. Sentiment analysis of consumer reviews on online shopping platforms using integrated deep learning models. *ICT Express*, 11(5):881–887.

Guangyu Zhou and Kavita Ganesan. 2016. Linguistic understanding of complaints and praises in user reviews. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 109–114. The Association for Computer Linguistics.