Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.5959

Detecting and Analysing Duplicate Consumer Complaints and Collective Demands Across Multiple Platforms

Gestefane Rabbi [**Universidade Federal de Minas Gerais (UFMG)** | ***gestefane@dcc.ufmg.br***]
Júlia Viterbo [**Universidade Federal de Minas Gerais (UFMG)** | ***juliapaes@dcc.ufmg.br***]
Gabriel Kakizaki [**Universidade Federal Viçosa (UFV)** | ***gabriel.kakizaki@ufv.br***]
Zilton Cordeiro Junior [**Universidade Federal de Minas Gerais (UFMG)** | ***zilton@dcc.ufmg.br***]
Raquel O. Prates [**Universidade Federal de Minas Gerais (UFMG)** | ***rprates@dcc.ufmg.br***]
Julio C. S. Reis [**Universidade Federal Viçosa (UFV)** | ***jreis@ufv.br***]
Marcos André Gonçalves [**Universidade Federal de Minas Gerais (UFMG)** | ***mgoncalv@dcc.ufmg.br***]

Department of Computer Science, Universidade Federal de Minas Gerais (UFMG), Av. Presidente Antonio Carlos, 6627, Pampulha, Belo Horizonte, MG, 31270-901, Brazil

Received: 25 April 2026 • **Published:** 13 March 2026

Abstract The increasing volume of data in consumer complaint repositories poses considerable challenges for the effective management and analysis of this information. A primary issue is the prevalence of duplicate complaints, often submitted by the same consumer across different platforms as a strategy to exert pressure on service providers. Furthermore, the identification of collective consumer demands embedded within these complaints is essential for revealing systemic issues affecting broader consumer groups. This study proposes a computational framework to address these dual challenges: (i) the detection of duplicate complaints through temporal correlation and cross-platform matching of key attributes—such as consumer identity, service provider, and complaint subject—and (ii) the identification of collective demands via clustering techniques based on semantic similarity. To this end, natural language processing (NLP) methods are employed to extract and represent semantic content from unstructured complaint texts. Empirical results indicate that 95% of duplicate complaints are submitted within a 30-day window from the original entry. Additionally, the proposed clustering approach demonstrates validated effectiveness to enhance the management of unstructured consumer complaint data, facilitating more efficient conflict resolution and informed decision-making for regulatory agencies and service providers.

Keywords: Consumer Complaints, Duplicates, Collective Demands, PROCON, Consumidor.gov, Sindec

1 Introduction

The worldwide widespread use of the Internet has led to more and more users using Web systems for online shopping. Recent studies have shown that in 2023 alone 62% of consumers make between two and five online purchases per month[1]. Furthermore, in countries such as Brazil, around 85% of the population with Internet access makes at least one online purchase per month. A recent study also found that 61% of Brazilian consumers shop more frequently online than in physical stores[2].

With the increasing number of purchases in online environments, the number of consumer complaints about products and/or services due to various issues (e.g., defects, delays, incorrect charges, etc.) has also increased. As a result, several platforms have emerged in Brazil where users can express their dissatisfaction. Examples include “Reclame AQUI”[3], “Consumidor.gov”[4] and “Denuncio”[5], to name just a few. Over time, the use of these services has expanded to include complaints about purchases or services made outside the online environment, for instance, in telecommunications and banking. These phenomena have contributed to the increase of technical challenges regarding the effective management of “digital complaints”.

In practice, responses to initial complaints are not always timely or forthcoming, leading consumers to submit repeated complaints — often across multiple platforms — as a strategy to escalate their demands and compel a resolution. As a result, the number of duplicate complaints on these platforms keeps growing, posing significant challenges for effective information management. This data redundancy can hinder conflict resolution, as inflated complaint counts may lead to incorrect or unrealistic conclusions about the prevalence of issues related to a specific product or company [de Carvalho *et al.*, 2008].

In the event of unresolved issues, consumers can take legal action against the companies that caused the problem. In order to facilitate access to justice and reduce the number of lawsuits, the competent bodies such as the Public Prosecutor’s Office and the Public Defender’s Office can file the so-called *collective demands* based on the number of complaints targeting a same company or product. These lawsuits are aimed at defending the interests of a group of consumers who face similar situations. In this work, we present an

[1] https://forbes.com.br/forbes-money/2023/07/62-dos-consumidores-fazem-ate-cinco-compras-online-por-mes-aponta-pesquisa/

[2] https://g1.globo.com/economia/noticia/2022/12/14/61percent-dos-brasileiros-compram-mais-pela-internet-do-que-em-lojas-fisicas-aponta-estudo.ghtml

[3] `reclameaqui.com.br`

[4] `consumidor.gov.br`

[5] `denuncio.com.br`

approach to address both issues: (i) **removal of complaint redundancy** and (ii) **automatic detection of collective demands**, which is an extension of our previous work [Rabbi *et al.*, 2024]. In particular, this study has the following guiding research question (RQ): *How can we automatically detect and group duplicate and semantically similar consumer complaints, written in Brazilian Portuguese, across multiple platforms to support the identification of collective consumer demands in a practical, scalable, and interpretable way?* Here, we hypothesize that combining BERT-based semantic similarity with topic modeling techniques enables the identification of latent collective issues from noisy, large-scale consumer complaint data in a manner that is useful and actionable for public institutions.

In order to do this, we analyzed 1,723,245 complaints registered by consumers on three different platforms from 2006 to 2024 (see details in Table 1), all originating from cities in the state of Minas Gerais, Brazil: Consumidor.gov, PROCON-MG and Sindec. We proposed an approach to identify duplicate complaints that involves the temporal analysis [Mourão *et al.*, 2008; Salles *et al.*, 2010] and the study of attributes such as the consumer (complainant), the provider (complained party) and the subject of the complaint.

Our proposed approach utilizes advanced natural language processing (NLP) techniques for Brazilian Portuguese, specifically the BERTimbau model [Souza *et al.*, 2020], which is superior to traditional methods such as fuzzy matching [Wang *et al.*, 2017] or other machine learning techniques based on *soft computing*, being more robust to the variability of natural language in consumer complaints [de Carvalho *et al.*, 2006, 2008]. The overall goal is to recognize semantic similarities between complaints by setting appropriate thresholds for the identification of duplicates.

The results obtained from our duplicate complaint identification approach show that 95% of duplicates were posted within 30 days after the initial submission, indicating a temporal range of consumer persistence in seeking resolution for their issues. Additionally, the length and word clouds of duplicates across the three data repositories highlight significant differences in the content and posting patterns of duplicated complaints. These differences include references to legal fragments or more mentions of services and companies, reflecting the unique characteristics of each dataset.

Duplicate complaints may artificially inflate analytical outcomes by misrepresenting repeated individual cases as distinct, collective grievances and thereby exaggerate their prevalence. Thus, having identified duplicate complaints, we propose a method for systematically identifying genuine collective demands. Particularly, we apply BERTopic [Grootendorst, 2022], a method that incorporates algorithms for automatically searching dense topics in a collection of documents (in the case, complaints), assuming that semantically similar documents form topics (i.e., potential collective demands). Experimental results demonstrate that the proposed method produces semantically coherent clusters of complaints. A manual evaluation of these groupings yielded substantial inter-annotator agreement (Fleiss' Kappa[6] = 0.6580), indicating the method's effectiveness in identifying potential collective demands. These findings highlight the approach's potential to enhance the management and prioritization of consumer complaints recorded across digital platforms.

The remainder of this article is organized as follows. Sections 2 and 3 present definitions and related work. Section 4 describes the details of the proposed approach for identifying duplicate complaints. Next, section 5 describes the manual validation conducted. The main results are discussed in Section 6. Section 7 concludes the study and outlines directions for future research.

2 Background

We here define the concepts of duplicate consumer complaints and collective demands adopted in this work.

When engaging with online complaint registration platforms, consumers may submit duplicate records, often as a means of exerting pressure on the company in question to prompt a timely response. This is typically achieved by entering a new complaint containing identical text to a previously submitted record or by making minor alterations that either preserve the original message or further emphasize it.

Formally, the task of identifying duplicate consumer complaints can be defined as follows: given a pair of complaints $[c_i, c_j]$, made by the same user u about a product or service p in a time interval t, a similarity threshold τ and a similarity function (e.g., cosine) $s(c_i, c_j) \in [0, 1]$ indicating the extent to which c_i and c_j are (semantically) similar, we define a parameterized similarity function $F_{s,\tau}$ between two complaints c_i and c_j as:

$$F_{s,\tau}(c_i, c_j) = \begin{cases} 1 & \text{if } s(c_i, c_j) \geq \tau \\ 0 & \text{otherwise.} \end{cases}$$

A further challenge in the domain of consumer complaints is the growing volume of reports concerning specific companies, products, or services, which frequently point to systemic issues affecting a broad range of individuals. These recurring complaints, commonly referred to as collective demands, signify widespread dissatisfaction that extends beyond isolated incidents. If not addressed properly, such claims can significantly undermine consumer trust, damage corporate reputations, and disrupt market stability, thereby underscoring the importance of their timely identification in effective complaint management.

Identifying collective demands is essential to reveal patterns indicative of systemic issues and facilitate data-driven interventions. By detecting clusters of similar complaints, companies are better positioned to implement targeted enhancements to their products and services, while regulatory bodies can derive actionable insights to uphold consumer rights and enforce relevant legal protections. However, this process presents significant challenges due to the unstructured nature of the complaint data, the variability in how issues are articulated, and the dispersion of complaints across multiple platforms, all of which hinder the reliable detection of meaningful patterns.

Formally, the task of identifying collective demands can be defined as follows: given a set of complaints

[6] Fleiss' Kappa is a statistical measure of inter-rater agreement for categorical labels [Fleiss *et al.*, 1971].

$C = \{c_1, c_2, \ldots, c_n\}$ registered by distinct users about the same product or service p in a time interval t, a similarity threshold λ and a scoring function $r(C) \in [0, 1]$ indicating the extent to which the complaints in C refer to a shared issue affecting multiple consumers, we define a binary classification function $G_{r,\lambda}(C)$ as:

$$G_{r,\lambda}(C) = \begin{cases} 1 & \text{if } r(C) \geq \lambda \\ 0 & \text{otherwise.} \end{cases}$$

3 Related Work

This section discusses related work along two main dimensions: (i) consumer complaints and (ii) general strategies for removing duplicate information in data repositories across different contexts (not limited to complaints).

3.1 Consumer Complaints

Consumers increasingly utilize online platforms to voice their dissatisfaction with specific products and/or services [Almeida and Ramos, 2012]. This widespread behavior has resulted in the generation of substantial volumes of data, which has garnered the attention of researchers across various disciplines for a range of analytical purposes. For instance, Sargiani *et al.* [2020] conducted a study leveraging consumer complaint data to identify problematic sectors within the market. The researchers analyzed complaints submitted to the Sindec database between 2013 and 2017, identifying the banking sector as the most frequently targeted, with approximately 90,000 complaints. A more detailed examination indicated that retail banks were responsible for the majority of these complaints, including around 10,000 related to incorrect fees.

Similarly, [Félix *et al.*, 2018] conducted an investigation into the application of natural language processing techniques for identifying prevalent topics in consumer complaints related to telecommunications service providers. The study utilized a dataset comprising approximately 300,000 complaints collected from PROCON-MG, Reclame AQUI, and Twitter (currently referred to as X). For topic modeling, the Latent Dirichlet Allocation (LDA) algorithm was employed [Jelodar *et al.*, 2019]. The findings indicated that the predominant issues related to telephone and Internet services, with the Reclame AQUI platform yielding more actionable insights than the other sources, thus offering greater utility for businesses seeking to address consumer concerns.

In [Freitas and Andreão, 2021], the authors proposed a methodology for the automation and preprocessing of raw data obtained from a consumer complaint service, with the objective of facilitating the development of automatic classifiers to determine whether service providers successfully contacted consumers in response to their complaints. The study utilized approximately 13,000 complaints related to telecommunications and pay-TV companies, collected from the Anatel Consumidor platform. The proposed methodology demonstrated a reduction in training time for classification models, while maintaining classification accuracy at a satisfactory level.

3.2 Strategies for Removing Duplicate Information

Approaches to identify and remove duplicate data have been proposed in various domains, such as healthcare records [Elmagarmid *et al.*, 2007], product catalogs in e-commerce [Ripon *et al.*, 2010], social media and Web content [Mansoor *et al.*, 2020], and bibliographic databases [de Carvalho *et al.*, 2011]. In fact, poor handling of duplicate information can lead to various problems, such as increased processing costs and reduced performance of models trained on such data [Barz and Denzler, 2020].

Related research in this domain exhibits similarities with the present study by analyzing or evaluating the discriminative capacity of attributes for the purpose of entity resolution [Mangaravite *et al.*, 2022; Carvalho *et al.*, 2022; Belém *et al.*, 2023; Silva *et al.*, 2019]. While these prior studies primarily focus on structured data and investigate which attributes (such as name, identification number, or address) most effectively distinguish between entities, the approach presented in this study extends this line of inquiry to unstructured textual data—specifically, consumer complaints. In this context, the central challenge involves identifying semantically similar content within noisy and heterogeneous textual descriptions.

Previous research on the elimination of duplicate information has predominantly focused on structured data, such as records within tabular databases, and has often employed rule-based approaches — either manually defined or generated through machine learning techniques [Elmagarmid *et al.*, 2007; Mangaravite *et al.*, 2022; Carvalho *et al.*, 2022]. However, these methods are generally inadequate for handling high-dimensional textual data, particularly noisy text generated by end users. In such contexts, recent studies have explored the use of semantic embeddings and deep learning models to address the task more effectively [Mansoor *et al.*, 2020; de Andrade *et al.*, 2023]. In particular, Mansoor *et al.* [2020] proposed a hybrid model that integrates Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs) to assess semantic similarity between sentence pairs, achieving an accuracy of 87.5% on the Quora question pairs dataset. Related to this, de Andrade *et al.* [2023] examined the class separability of contextual embedding representations of texts, demonstrating that well-optimized embeddings can produce highly separable feature spaces, thereby reducing the dependence on complex classifiers in text classification tasks.

3.3 Research Gap

This study distinguishes itself from the existing literature in two primary ways: (i) by focusing on the analysis of duplicate consumer complaints sourced from multiple government platforms, which are examined collectively; and (ii) by employing data preprocessing techniques to enhance the accuracy of quantitative analyses. Notably, the study incorporates natural language processing (NLP) methods, which have not been extensively utilized in prior research for this specific type of analysis.

With respect to duplicate information removal strategies, traditionally centered on the deduplication of structured data, this study adopts a distinct approach by targeting unstructured data, specifically consumer complaint texts,

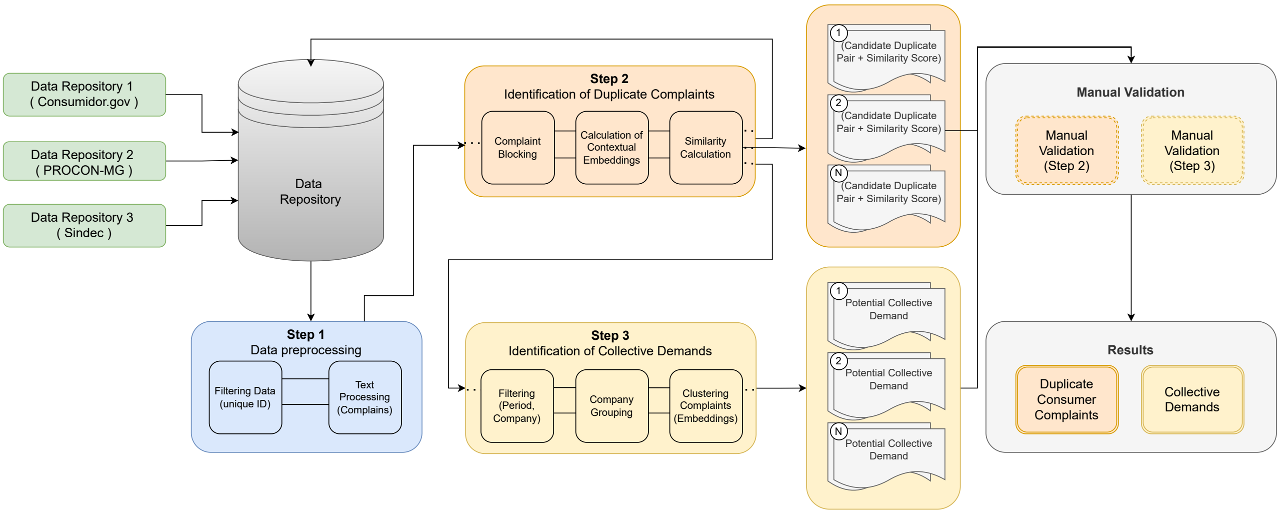

Data Repository 1
(Consumidor.gov)
Data Repository 2
(PROCON-MG)
Data Repository 3
(Sindec)
Data Repository
Step 1
Data preprocessing
Filtering Data (unique ID)
Text Processing (Complains)
Step 2
Identification of Duplicate Complaints
Complaint Blocking
Calculation of Contextual Embeddings
Similarity Calculation
Step 3
Identification of Collective Demands
Filtering (Period, Company)
Company Grouping
Clustering Complaints (Embeddings)
1 (Candidate Duplicate Pair + Similarity Score)
2 (Candidate Duplicate Pair + Similarity Score)
N (Candidate Duplicate Pair + Similarity Score)
1 Potential Collective Demand
2 Potential Collective Demand
N Potential Collective Demand
Manual Validation
Manual Validation (Step 2)
Manual Validation (Step 3)
Results
Duplicate Consumer Complaints
Collective Demands

Figure 1. Overview of proposed approach.

and addressing the linguistic nuances inherent to this context. A further distinguishing feature of the study is its focus on texts written in Brazilian Portuguese, a language for which significantly fewer linguistic resources are available compared to English.

While deep entity matching methods — as surveyed in Neural Networks for Entity Matching (Barlaug & Gulla, [Barlaug and Gulla, 2021]) - have advanced neural architectures for structured record linkage, those techniques depend on labeled data and structured schemas. Our pipeline, by contrast, processes unstructured consumer complaint texts in Brazilian Portuguese, performing deduplication using BERTimbau embeddings and cosine similarity thresholds without supervised labeling.

Indeed, BERTopic has been applied successfully to consumer complaint corpora, such as the U.S. Consumer Financial Protection Bureau (CFPB) dataset [Vaishnav *et al.*, 2024], demonstrating greater topic coherence and interpretability than traditional methods like LDA and LSA. To our knowledge, our implementation is one of the first to apply BERTopic at scale to Brazilian Portuguese consumer complaints [dos Santos *et al.*, 2023; Bastani *et al.*, 2019].

Finally, our study diverges from previous ones by proposing a novel approach to identify potential collective demands. To the best of our knowledge, this specific task has not yet been addressed in the existing literature. The proposed approach is expected to offer valuable support to data management professionals operating in this context.

4 Methodology

This section outlines the proposed approach for addressing two key aspects of the problem investigated in this study: (i) the identification of duplicate complaints across multiple data repositories, and (ii) the detection of collective demands within the remaining complaint data. The first phase is dedicated to identifying duplicate and near-duplicate complaints to ensure data consistency and reduce redundancy. In the second phase, the remaining complaints are organized by company, service, and/or product, utilizing semantic representations to generate potential clusters of collective demands. Both phases underwent manual validation to ensure reliability. An overview of the proposed approach is presented in Figure 1, with each step described in detail below.

4.1 Data Repository

In this study, we explored data from three different repositories: **Consumidor.gov**, **PROCON-MG**, and **Sindec**. A description of these repositories is given below:

4.1.1 Consumidor.gov

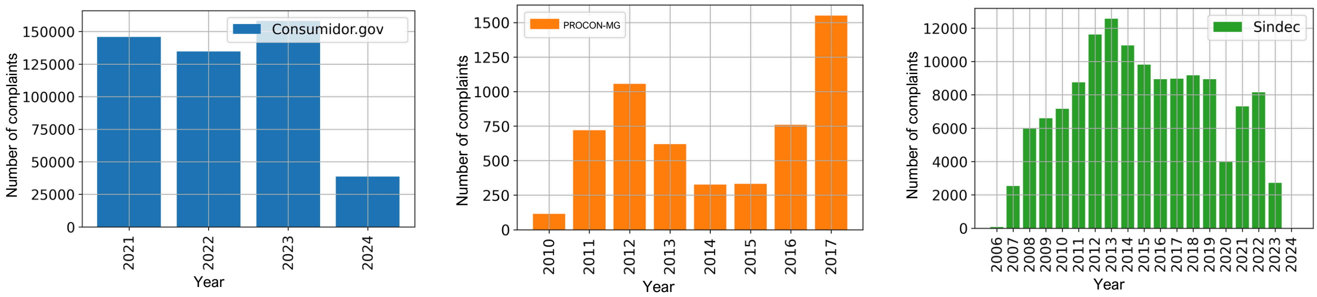
A public platform managed by Senacon (the National Consumer Secretariat), an agency under the Ministry of Justice and Public Security responsible for consumer protection policies in Brazil. Through this platform, consumers can file complaints directly against companies and seek resolution. To participate, companies must register and sign a commitment agreement. Consumers must also have a Silver or Gold account on Gov.br to access the service. The platform has been available since 2021 and can be accessed at `consumidor.gov.br`.

4.1.2 PROCON-MG

A State Consumer Protection and Defense Program that, among other responsibilities, receives consumer complaints and mediates potential solutions with companies. The agency was established to safeguard consumer rights and also takes action against practices that harm collective interests. Complaints against a company can be filed online, by phone, or in person at a PROCON-MG office.

4.1.3 Sindec

The National System for Consumer Protection Information is a set of technologies to integrate and consolidate information on consumer protection authorities, providing a robust management tool to meet these authorities' dynamic needs.



(a) Consumidor.gov

(b) PROCON-MG

(c) Sindec

Figure 2. Distribution of complaints by repository per year.

Table 1. Summary of complaints per repository with unique consumer (or complainant) and company (or complained party) identifiers.

Repository	Complaints	Complainants	Complained Parties	Avg. per Complainant	Avg. per Complained Party	Period
Consumidor.gov	477,436	213,478	845	2.2	565.0	01/2021–03/2024
PROCON-MG	5,476	5,065	2,814	1.1	1.9	08/2010–11/2017
Sindec	134,198	98,426	13,754	1.4	9.8	10/2006–02/2024
Total	617,110	316,969	17,413	1.9	35.4	10/2006–03/2024

4.1.4 *Data Summary*

In total, we collected 951,814; 427,239; and 344,192 complaints from the platforms Consumidor.gov, PROCON-MG and Sindec, respectively, dating back to 2006, all originating from cities in the state of Minas Gerais, Brazil. The distribution is shown in Figure 2, where a clear increase in the volume of complaints can be observed from 2021 onwards, probably related to the launch of Consumidor.gov. Finally, these 1,723,245 complaints are consolidated into a single data repository that serves as input for the subsequent steps.

4.2 Step 1: Data Preprocessing

The initial phase of our methodology involves data preprocessing, which leverages unique identifiers for complainants and complained parties (i.e., the company or organization against which the complaint was filed) derived through a Master Data Management (MDM) process [Loshin, 2010]. Comprehensive details regarding the implementation adopted in this study are available in [Mangaravite *et al.*, 2022]. In essence, this process ensures the consistent and unambiguous identification of entities across all integrated data sources, thereby facilitating more accurate and efficient analytical procedures.

Table 1 presents the number of complaints recorded in each individual data repository following the preprocessing stage. Consumidor.gov accounts for the highest volume of complaints, totaling 477,436 entries, followed by Sindec and PROCON-MG, with 134,198 and 5,476 records, respectively. Notably, Consumidor.gov also reports the greatest number of unique complainants (213,478), whereas Sindec registers the highest number of distinct complaint targets (13,754). On average, each platform received approximately two complaints per complainant and 35 complaints per complained party.

4.3 Step 2: Identification of Duplicate Complaints

The objective of this step is to identify complaints from the same consumers that pertain to the same company or product within a specified time frame. This step is essential to ensure the accuracy of the subsequent quantitative data analysis. Consumers may submit multiple complaints to the same or different organizations in response to a single incident requiring intervention by these authorities, which can lead to an artificial inflation of statistics for certain companies or industries. To address this, the proposed approach is divided into three primary steps described in detail below.

4.3.1 *Complaints Blocking*

The first stage involves a blocking procedure, in which complaints submitted by the same user and directed to the same company are grouped together. This blocking process is performed using unique user and company identifiers that have been previously filtered. As a result, multiple blocks are created, each comprising the complaints of a specific consumer directed to a particular company within a one-month period. The primary objective of this step is to reduce the number of necessary comparisons, thereby optimizing computational efficiency and lowering processing costs.

4.3.2 *Representing Complaints Using Contextual Embeddings*

Following the blocking phase, a pre-trained sentence embedding model based on the BERT architecture was employed to generate vector representations of the complaints, capturing their underlying semantic content. These embeddings facilitate the identification of complaint pairs that exhibit a level of similarity exceeding a specified threshold. The model adopted in this study was BERTimbau [Souza *et al.*, 2020], which is specifically trained on Brazilian Portuguese data, in conjunction with the sBERT library [Reimers and Gurevych, 2019], which encodes each complaint into a semantic vector within a latent space. This representation enables the application of vector-based similarity metrics, such as cosine similarity or distance measures, to detect instances of semantically equivalent or near-duplicate complaints.

In this context, it is important to underscore the necessity of employing deep learning-based approaches in place of traditional methods, such as fuzzy matching [Wang *et al.*, 2017;

Miller *et al.*, 2009], given the inherent complexity and variability of natural language found in consumer complaints. These texts frequently contain informal language, spelling mistakes, lexical variation (e.g., synonyms), and idiomatic expressions. Such linguistic nuances pose significant challenges for traditional techniques, which lack the robustness needed to effectively handle the diverse and often noisy language typical of user-generated content. Moreover, conventional models are limited in their capacity to capture semantic similarity, as different individuals may express complaints about the same entity using distinct phrasing while conveying equivalent meanings.

4.3.3 Calculating Similarity

Subsequently, the similarity between the embedding vectors generated in the previous step is calculated. This is accomplished by computing the cosine distance between pairs of complaint vectors, resulting in a similarity score for each pair. The scores are ranked in descending order, allowing a domain expert to determine an appropriate similarity threshold for classifying complaints as duplicates. Based on empirical evaluations conducted in this study, a similarity threshold of 90% was established for the analysis. Accordingly, two complaints are considered duplicates if their similarity score meets or exceeds this threshold and if they were submitted within a 30-day period, as defined by the domain expert.

The similarity threshold of $\tau = 0.90$ was not arbitrarily chosen. It was based on empirical evaluations during our preliminary experiments and confirmed by domain experts, aiming to minimize false positives in institutional contexts. This conservative setting was further supported by the manual validation in Section 5, where annotators achieved substantial agreement (Fleiss' $\kappa = 0.748$).

As a result of this process, pairs of candidate duplicate complaints are identified along with their corresponding similarity scores. In the context of this study, a total of 17,583 duplicate complaints were identified among the 617,110 complaints analyzed (as presented in Table 1), spanning multiple data repositories. It is important to highlight that the similarity threshold adopted in this analysis was intentionally conservative. In practical applications, the identified complaint pairs could be ranked in descending order of similarity to facilitate further expert review, which may enhance the overall effectiveness and adaptability of the proposed approach.

4.4 Step 3: Identification of Collective Demands

After identifying duplicate complaints, the next phase of the proposed approach focuses on detecting potential collective demands. As outlined in Step 3 of Figure 1, this phase involves a three-step process: (i) filtering complaints based on contextual parameters (e.g., time period and company), (ii) grouping them by company or service, and (iii) applying semantic clustering techniques to identify recurring issues affecting multiple consumers. These clusters are subsequently interpreted as potential collective demands, which are further validated by experts.

The process begins with predefined parameters, such as the desired time period, data repository, and target company. In a practical application, the expert user can input these parameters through a user-friendly interface. The "Company" parameter is optional, allowing for flexible analysis. Based on the inputs provided, the approach filters complaints that meet the defined criteria and organizes them accordingly. Relevant complaint information is retrieved, including complaint identifiers and content.

This approach employs semantic grouping techniques, utilizing BERTopic [Grootendorst, 2022], a topic modeling algorithm, to cluster complaints based on similarity. It identifies clusters that represent potential collective demands by grouping complaints with common semantic features related to a shared problem, product, or service.

Before detailing the semantic grouping technique, it is important to clarify that BERTopic was not used for deduplication or entity matching. That process was fully addressed in the prior step using BERTimbau embeddings and cosine similarity with a high threshold. In the current phase, BERTopic is applied solely to the set of complaints that have already been deduplicated. Its role is to group them into interpretable clusters representing potential collective demands, enabling a more actionable interpretation for human validation.

Unlike the previous step, this phase compares complaints from different complainants regarding the same company or product, excluding previously identified duplicates. Complaints are initially filtered by the target company using input parameters such as the time period and company name, if available. If company information is absent, an additional internal filtering process ensures that all analyzed complaints refer to the same company.

The clusters generated by our approach represent potential collective demands, each corresponding to complaints grouped by a shared issue. An example is provided in Table 2. These clusters offer valuable insights for identifying systemic problems within companies, services, or products, forming a data-driven foundation for collective consumer protection actions. However, the approach is intended to assist domain experts, who must manually validate the results and determine the appropriate actions.

Table 2. Example of clustering results for a telecommunications company with associated keywords and representative complaints in Portuguese.

Topic Id	Associated Keywords	Representative Complaints
-1	[<company>, internet, plano, ...]	[O reclamante é consumidor...]
0	[cancelamento, plano, controle ...]	[Em outubro fiz a solicitação...]
1	[gb, assinantes, 00, bonus, ...]	[Localizei uma oferta de...]
2	[cashback, liuz, reclamac, ...]	[Eu já fiz uma reclamação...]
3	[internet, cmicos, 18 dias, ...]	[Contratei a <company>...]
...	...	...
252	[contestac, dificuldade, art, ...]	[Boa noite!\n Entrei em...]
253	[31, meros consigo, mero 1056, ...]	[Estou recebendo diversas...]
254	[ditos recarga, encargos, ...]	[Fiz uma consulta no...]
255	[eusta, rua padre, padre...]	[Em outubro do ano de 2023...]

Based on the results presented in Table 2, the approach successfully identified 256 potential collective demands, each represented by a cluster of complaints grouped around common keywords. The first row (Topic or Demand ID -1) corresponds to a complaint that did not align with any specific group, remaining isolated and not forming a collective demand. As an example, the first valid collective demand (i.e., Demand ID 0), shown in Table 3, was derived from 36 individual complaints and is characterized by representative terms such as "cancellation", "plan", "control", and "billing". Terms enclosed in angle brackets in Table 2 (e.g., `<company>`) indicate omitted named entities.

Table 3. Example of complaints used to build a collective demand.

Complaint Id	Index	Opening Date	Text (in Portuguese)
125589	0	2024-04-30	Meu plano teve aumento de preço...
127875	1	2024-04-22	empresa cancelou numero e passou...
129660	2	2024-03-31	Na noite de hoje tentei realizar o...
131294	3	2024-03-26	Já faz um tempo que estou tentando...
...	...	...	...
134959	34	2024-03-18	Solicitei o cancelamento dos servi...
136900	35	2024-03-13	Bom dia\n\nPor favor o cancelam...

4.5 Validation Measure

The evaluation of our approach required not only quantitative analysis of system outputs but also qualitative assessment of their interpretability and practical relevance. In both tasks — duplicate complaint identification and collective demand detection — the final step involved human annotators judging whether the results produced by the system were correct. To transform this subjective evaluation into a reliable measure of quality, we employed an inter-rater agreement statistic. Specifically, we adopted Fleiss' Kappa [Fleiss *et al.*, 1971], which quantifies the degree of consensus among multiple raters while correcting for the level of agreement expected by chance.

To formally assess inter-rater reliability in the manual validation, we calculated Fleiss' Kappa [Fleiss *et al.*, 1971]. Given N items independently classified into k categories by n raters, let n_{ij} denote the number of raters who assigned item i to category j. The proportion of raters who agreed on item i is:

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij}-1).$$

The overall agreement is the average across all N items:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i.$$

Let $p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$ be the overall proportion of assignments to category j, and define

$$\bar{P_e} = \sum_{j=1}^{k} p_j^2.$$

Finally, Fleiss' Kappa is given by:

$$\kappa = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}}.$$

This statistic corrects for chance agreement and ranges from -1 (complete disagreement) to $+1$ (perfect agreement), with 0 corresponding to chance-level agreement. According to McHugh [McHugh, 2012], values between 0.61 and 0.80 indicate substantial agreement, while values above 0.80 indicate almost perfect agreement.

Table 4. Results of the manual labeling performed by three independent evaluators over candidate duplicate pairs. The table reports agreement levels (total, partial, none) and the proportion of instances confirmed as duplicates (Yes) or rejected (No), per repository.

Repository	Agreement (%)			Are duplicates (%)	
	Total	Partial	None	1 - Yes	2 - No
Consumidor.gov	87.00	10.00	3.00	78.00	9.00
PROCON-MG	90.48	9.52	0.00	90.48	0.00
Sindec	85.00	12.00	3.00	55.00	30.00

5 Manual Validation

First, we performed a manual evaluation of the results obtained from the execution of Steps 2 and 3 (Figure 1) of the methodology proposed for carrying out this study, as presented below.

5.1 Duplicate Complaints

To qualitatively assess the performance of the proposed approach, a random sample of candidate duplicate pairs identified in each data repository was manually labeled. Specifically, from the 17,583 duplicate pairs identified, 100 pairs were randomly selected from each repository, with the exception of PROCON-MG, where only 21 candidate duplicate pairs were identified. In total, 221 complaint pairs were subjected to manual evaluation. Each pair was independently reviewed by three volunteers, who classified the instances into one of three categories: (1) "Duplicate" for cases that clearly represented a duplicate record of an original complaint; (2) "Not duplicate" for cases that did not; and (3) "Inconclusive" when the evaluator was unable to confidently assign a label.

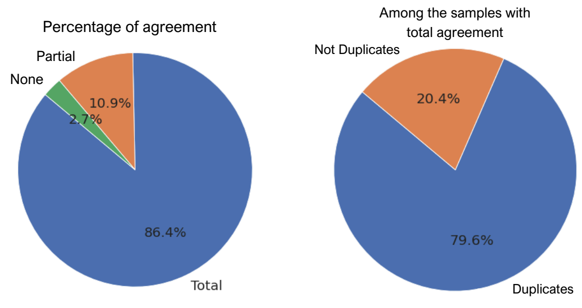
To assess the level of agreement among the evaluators, Fleiss' Kappa was calculated. The degree of agreement was categorized as follows: total agreement when all evaluators assigned the same label, partial agreement when at least one evaluator assigned a different label, and no agreement when all three evaluators assigned distinct labels to the same instance.

The results of this evaluation are summarized in Table 4.

In the case of the Consumidor.gov repository, 78% of the instances were labeled as duplicates by the evaluators, while 9% were labeled as non-duplicates. This indicates that 9% of the cases identified as duplicates by the model were, in fact, not considered duplicates by the evaluators. For the Sindec repository, 30% of the model-identified duplicate samples were classified as non-duplicates by the evaluators, suggesting a higher rate of false positives. In contrast, for the PROCON-MG repository, all samples identified by the model as duplicates received unanimous agreement from the evaluators, confirming their classification as true duplicates.

Overall, the *Fleiss' Kappa* score was 0.748, indicating substantial agreement among evaluators. As illustrated in Figure 3(a), total agreement was achieved for 86.4% of the evaluated samples, corresponding to 191 out of 221 samples. Partial agreement was observed in 10.9% of the samples (24 out of 221), and no agreement was reached in only 2.7% of the samples (6 out of 221).

An instance of complete disagreement among the three evaluators is presented in Table 5. The complaints exhibit a high degree of textual similarity, including mentions to the same order number, some overlapping dates, and an identical protocol number. However, the first complaint refers to only



(a) Agreement among evaluators (b) Duplicate confirmation

Figure 3. Agreement among evaluators and duplicate confirmation.

one order and one protocol, whereas the second includes references to two additional orders and two additional protocols. The evaluator who classified the pair as duplicates may have interpreted the first complaint as being subsumed within the second. Nevertheless, among the four key dates cited in each complaint—return date, order arrival date, refund date, and stated deadline — only two dates coincide. Given that the second complaint refers to multiple orders, it is possible that the overlapping dates reflect the earliest or latest dates associated with those additional transactions.

In this case, the evaluator who classified the texts as non-duplicates may have concluded that the first text, which contained fewer and differing details compared to the second, could not be regarded as a duplicate. On the other hand, the evaluator who labeled the texts as inconclusive may have considered both interpretations to be valid and, as a result, was unable to make a definitive judgment.

Figure 4 presents the distribution of labels assigned by the three evaluators during the manual validation process. A notable level of consistency is observed among the evaluators concerning the number of samples labeled as duplicates (label 1) and non-duplicates (label 2), while inconclusive labels (label 3) were relatively rare. Additionally, Figure 3(b) illustrates that, among the samples exhibiting total agreement, 79.6% of the instances identified as duplicates by the model were also confirmed as duplicates during the manual evaluation. This outcome reflects the satisfactory performance of the approach proposed in this study.

5.2 Potential Collective Demands

Finally, to assess the effectiveness of our proposed approach in identifying collective demands, we applied it to complaints registered by consumers within a predefined time period, specifically from November 28, 2024, to December 31, 2024. This period was selected because it includes Black Friday 2024, a major retail event that typically occurs on the Friday following Thanksgiving in the United States and has gained widespread popularity globally, including in Brazil. Known for substantial discounts and special sales both in stores and online [BlackFriday.com.br, 2024], Black Friday induces a significant increase in consumer shopping activity, thus increasing the likelihood of post-purchase complaints.

Table 5. Example of total disagreement among evaluators.

Text 1
Order XXXXXXXXXXXX3116 was returned on 12/12 and arrived at <COMPANY> on 12/13. The refund was supposed to be processed by 12/23 according to the informed deadline, but they stated it would be processed by 12/27, as per protocol XXXXX358. More than 10 days have passed since the refund deadline, and the refund has still not been processed.
Text 2
Orders XXXXXXXXXXXX6431, XXXXXXXXXXXX3116, and XXXXXXXXXXXX9871 were returned on 12/12 and arrived at <COMPANY> on 12/14. The refund was supposed to be processed by 12/23 according to the informed deadline, but they stated it would be processed by 12/27, as per protocols XXXXX942, XXXXX813, and XXXXX358. More than 10 days have passed since the refund deadline, and the refunds have still not been processed.

Table 6. Volume of complaints, number of identified clusters, and average size of clusters per domain (Bank, Card, and Online Shopping) from November 28 to December 31, 2024 .

Group	Total Complaints	Clusters	Avg Complaints/Cluster
Bank	4740	468	10.13
Card	1394	92	15.15
Online Shopping	223	20	11.15

To facilitate the manual validation process of the identified results, we randomly selected potential collective demands from three key areas of interest: “Banking Services,” “Credit Card Services,” and “Online Shopping Platforms.” These areas were chosen based on recent research indicating that they received the highest number of complaints in 2024 [Reclame Aqui, 2024]. As shown in Table 6, overall we find that the **Bank** segment has the highest total number of complaints (4,740) and the largest number of potential collective demands (468 clusters), with an average of 10.13 complaints per cluster. The **Card** segment follows with 1,394 complaints spread across 92 clusters, resulting in a higher average of 15.15 complaints per cluster. Finally, **Online Shopping** segment has the lowest total number of complaints (223) and 20 clusters, with an average of 11.15 complaints per cluster.

In summary, Table 6 offers an overview of the volume of potential collective demands across different segments, emphasizing the Bank segment as the most significant, both in terms of total complaints and potential collective demands. The Card segment, while featuring fewer clusters, exhibits a higher concentration of complaints per cluster, suggesting a greater potential for more targeted collective actions. The Online Shopping segment, although smaller in scale, remains a relevant area for collective demand analysis due to its relatively high average number of complaints per cluster.

Based on the results obtained, we randomly selected three potential collective demands from each area (for a total of nine) and analyzed ten associated random complaints for each demand. The groups of potential collective demands and complaints were manually evaluated by three volunteers, who classified them into the following categories: (i) *coherent with the cluster* **(Yes)**, (ii) *not related to the cluster* **(No)**, or (iii) *undecidable due to insufficient or ambiguous informa-*

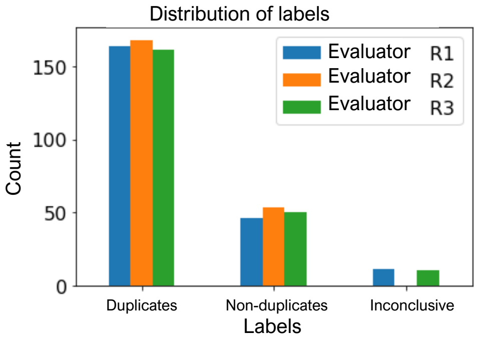


Figure 4. Distribution of labels among evaluators.

tion **(Inconclusive)**. This annotation process was designed to assess the alignment between the complaints within each cluster according to human judgment. To evaluate the consistency of the annotations, we also calculated *Fleiss' Kappa*. The results from this step are further discussed in the next section (Section 6.2.2).

6 Results and Discussion

This section presents the main results of the proposed approaches for identifying (i) duplicate consumer complaints from different data repositories and (ii) potential collective demands, based on the conducted manual validation.

6.1 Duplicate Consumer Complaints

The identification of duplicate complaints is a vital component in improving data quality and ensuring the integrity of consumer protection analyses. Such duplicates can emerge due to various factors, including individual user behavior, redundancies at the system level, or inconsistencies in data integration practices across different platforms. For the purposes of this study, duplicate complaints are defined as separate records submitted by the same consumer against the same company that demonstrate a high degree of textual similarity and are filed within a specified time frame. Effectively detecting these duplicates is essential not only to avoid inflating complaint volumes but also to enhance the detection of recurring issues that may signal broader patterns of consumer harm.

In this section, we present the results of our duplicate detection methodology, applied to three heterogeneous consumer complaint repositories: *Consumidor.gov*, *Sindec*, and *PROCON-MG*. We begin with an overview of the prevalence and distribution of duplicate complaints across these datasets, followed by a multi-dimensional analysis focusing on (i) textual content, (ii) user behavior, and (iii) temporality. Each dimension provides distinct insights into the nature and dynamics of duplication phenomena within and across platforms. The results offer a fundamental understanding of complaint redundancy patterns, which can support both data curation strategies and subsequent analytical tasks, such as the identification of potential collective claims.

6.1.1 General Analysis

A preliminary characterization was conducted across the three repositories under analysis, encompassing a total of 617,110 complaints with identifiable complainants and defendants, as detailed in Table 1. This dataset constitutes the full scope of data utilized for the duplicate detection analysis. Following the application of the proposed methodology, 17,583 complaints were identified as duplicates, accounting for approximately 2.8% of the total dataset.

The Consumidor.gov repository accounted for the majority of identified duplicates, with 16,185 cases, representing 94.3% of the total. In comparison, the Sindec repository contained 816 duplicate cases (4.6%), while the PROCON-MG repository had the fewest, with only 182 instances, corresponding to 1.0% of the total duplicates.

Building upon the general analysis, we conducted a more detailed examination of the characteristics of duplicate complaints across the three repositories (Consumidor.gov, Sindec, and PROCON-MG). This analysis was structured around three key dimensions:

1. Textual content – focusing on the primary textual features of the complaints;
2. User behavior – investigating the frequency with which individual users submitted duplicate complaints; and
3. Temporality – evaluating the time intervals between the original complaint and its corresponding duplicate.

6.1.2 Textual Content

To visually characterize the textual content of duplicate complaints across the repositories, word clouds were generated using the text from the earliest complaint in each duplicate pair — that is, the initial record within the defined temporal window. These visualizations, presented in Figure 5, illustrate notable differences in the thematic focus of the complaints across the datasets.

In the Consumidor.gov repository, the content of complaints was directed primarily toward companies. By contrast, the PROCON-MG data exhibited a more balanced emphasis on both services and products. The Sindec repository was distinguished by the frequent inclusion of legal references, such as articles and clauses, often quoted verbatim. Additionally, many complaints in Sindec were written in the third person, suggesting they may have been submitted by representatives on behalf of consumers. The presence of formal legal language likely contributed to longer complaint texts, a pattern examined in further detail below.

As shown in Table 7, the average lengths of duplicate complaints in the Consumidor.gov and PROCON-MG repositories were comparable, at 139 and 151 words, respectively. In contrast, complaints in the Sindec repository were significantly longer, a difference that can be attributed to the frequent inclusion of detailed legal references.

Table 7. Mean and standard deviation of the number of words and characters in complaints registered in the analyzed repositories.

Repository	**#Words**	**#Characters**
Consumidor.gov	139 (±119)	835 (±727)
PROCON-MG	151 (±153)	903 (±947)
Sindec	343 (±433)	2977 (±3974)

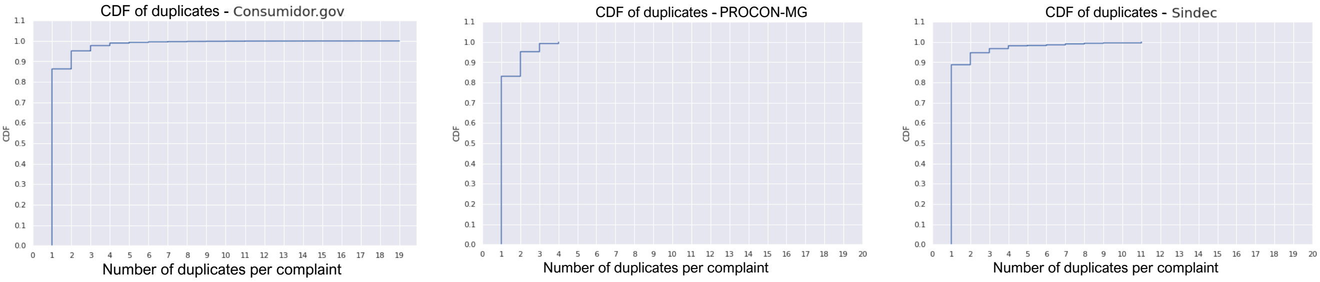
6.1.3 User Behavior

Subsequently, we analyzed user behavior, specifically the actions of complainants, in submitting duplicate complaints across the examined platforms. Overall, 95% of duplicate complaints consisted of no more than two additional copies beyond the original, although in rare cases, up to 19 duplicates were observed. As illustrated in Figure 6, the cumulative distribution functions for each repository confirm that the vast majority of complaints—95%—had at most two duplicates, consistent across all datasets.

Despite the significantly larger dataset in Consumidor.gov compared to the other repositories, the pattern of up to two duplicates per user was consistent across all domains.

(a) Consumidor.gov (b) PROCON-MG (c) Sindec

Figure 5. Word clouds (in Portuguese) for duplicated complaints in each repository.



(a) Consumidor.gov (b) PROCON-MG (c) Sindec

Figure 6. Cumulative distribution of duplicates per repository.

6.1.4 Temporality

The analysis then turned to the temporal dynamics of duplicate complaint submissions across the different platforms. Notably, nearly half of all identified duplicates were submitted on the same day as the corresponding original complaint.

Descriptive statistics for the time intervals between duplicate submissions revealed the following: for Consumidor.gov, mean interval was 8.54 days (±0.07), with standard deviation of 10.21 days (±0.05); for Sindec, mean interval was 4.83 days (±0.25), with standard deviation of 9.23 days (±0.18); and for PROCON-MG, mean interval was 0.34 days (±0.14), with standard deviation of 2.06 days (±0.10).

As illustrated in Figure 7, 95% of duplicate complaints were submitted within 30 days of the original across all repositories. However, significant variation was observed in the proportion of same-day duplicates: in Consumidor.gov, between 40% and 50% of duplicates were submitted on the same day; in PROCON-MG, this proportion exceeded 90%; and in Sindec, it was approximately 70%.

One possible explanation for these differences lies in the disparity in the volume of identified duplicates across the repositories. Consumidor.gov accounted for over 17,000 duplicate cases, whereas PROCON-MG and Sindec had only 816 and 182, respectively. This discrepancy suggests that the figures for PROCON-MG and Sindec may not fully capture broader temporal patterns. Unlike the user behavior analysis, where consistent trends were supported by the data, the current findings do not conclusively demonstrate distinct temporal characteristics across the repositories. Further research would be required to validate this hypothesis, indicating a potential direction for future study.

These findings answer the first part of our Research Question by showing that automatic detection of duplicates is feasible with high reliability through BERTimbau embeddings combined with a conservative similarity threshold and a 30-day temporal window, capturing the vast majority of repeated complaints with empirical validation.

6.2 Collective Demands

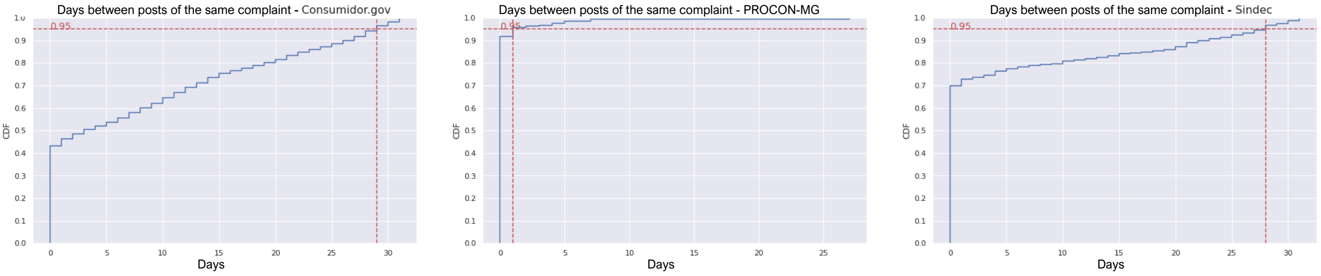
Consumer complaints often point not only to individual grievances, but also to broader, recurring issues that may justify collective legal action. In this subsection, we examine how such patterns emerge within the data by analyzing the frequency, thematic content, and distribution of complaints across different service domains. By combining quantitative insights with qualitative annotations, we aim to identify signs of potential collective demands that warrant further legal or institutional attention.

6.2.1 Overview

In this section, we show a distribution of complaints by company, separated by domain. This distribution is directly relevant to the identification of collective demands: in practice, even a relatively small number of complaints directed at a few companies may be sufficient to trigger a class action or institutional intervention. Thus, understanding whether dissatisfaction is concentrated or dispersed across providers helps contextualize the clusters of potential collective demands identified in subsequent analyses.

Figure 8 provides a visual representation of how consumer dissatisfaction is distributed within each sector, whether it is concentrated among a few companies or distributed across many. The figure highlights the ten companies with the highest number of complaints in three distinct domains: banking (a), credit card services (b), and online shopping (c).

In the banking sector, complaints are more evenly distributed among several institutions. The bank with the highest number of complaints received nearly 500, while others followed closely with totals ranging between 300-400. This



(a) Consumidor.gov (b) PROCON-MG (c) Sindec

Figure 7. Cumulative distribution of duplicates per day in each repository.

pattern suggests a relatively widespread level of consumer dissatisfaction affecting multiple entities within the sector.

By contrast, the credit card services domain exhibits a higher concentration of complaints. In this category, just two companies account for a substantial proportion of the total complaints, indicating that issues may be more localized, potentially stemming from specific service deficiencies or isolated user experience problems.

The online shopping sector presents a markedly different pattern, with a single company dominating the complaint landscape. This company received over 170 complaints—more than three times the number reported by the second-ranked one. This significant centralization of consumer complaints points to the possibility of systemic issues concentrated within a particular platform.

To examine the prevailing themes and linguistic patterns within consumer complaints, we employed a combined approach that integrates frequency-based and score-based analyses of terms associated with each service domain. Figure 9 presents word clouds depicting the most frequently mentioned terms in complaints related to banking (a), credit card services (b), and online shopping (c). In parallel, Figure 10 displays the top-ranked words based on their average importance scores, which reflect the discriminative relevance of each term within its respective category.

In the banking domain, both frequency and scoring analyses reveal a blend of textual noise — such as terms like “agg”, “vendo”, and “cc”, which may originate from automated systems or data artifacts — and substantively meaningful words including “desconformidade” (non-compliance), “encaminhar” (to forward), and “notificação” (notification). These terms suggest recurring issues related to document processing, procedural inconsistencies, and communication failures between institutions and consumers.

For the credit card services category, there is a stronger alignment between frequently occurring and highly scored terms. The word “notificação” (notification) emerges as both the most common and impactful, highlighting the centrality of formal notices and billing disputes in consumer dissatisfaction. Other relevant terms, such as “prazo” (deadline), “reduzido” (reduced), and “extrajudicial”, point to a heightened focus on legal proceedings and time-sensitive concerns.

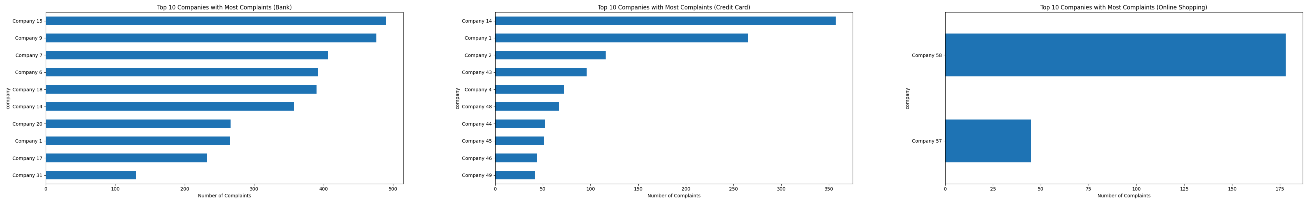
In the online shopping sector, both types of analysis converge around financial and legal terminology. Frequently cited and highly weighted terms such as “cobradas” (charged), “dividas” (debts), and “serasa” (credit bureau) underscore the prominence of financial grievances. The presence of terms like “stj” (Supreme Court), “cancelar” (to cancel), and various product or brand names reflects ongoing issues related to transaction disputes, customer service deficiencies, and the potential impact on consumer credit.

This dual analysis offers insights not only into the most common complaint topics but also into the linguistic patterns through which consumers express dissatisfaction. These findings may aid in the early identification of potential collective claims by highlighting recurring terms and themes that reflect broader consumer concerns.

6.2.2 Qualitative Analysis

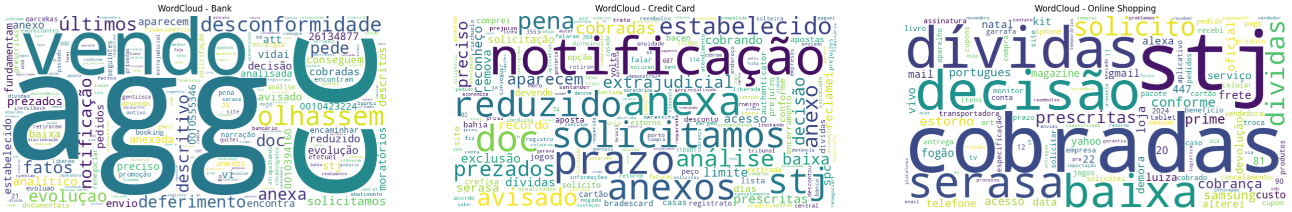
We conduct a qualitative assessment of the results, as defined in Section 5.2. Table 8 presents the Fleiss’ Kappa values per cluster, along with their respective 95% confidence intervals (CI) obtained via bootstrap resampling. In addition to the point estimates, these intervals offer insight into the statistical stability of the agreement measures. The results show variation in inter-rater agreement across clusters. Notably, Cluster 7 achieved perfect agreement ($\kappa = 1.000$), reflecting unanimous consensus among annotators. In contrast, Cluster 6 has a wide confidence interval (CI = $[-0.111, 1.000]$), indicating substantial uncertainty and inconsistency. Other clusters display a range of agreement levels, from slight to substantial, with varying degrees of statistical confidence.

These differences may stem from the clarity or ambiguity of the complaint texts within each cluster or from differences



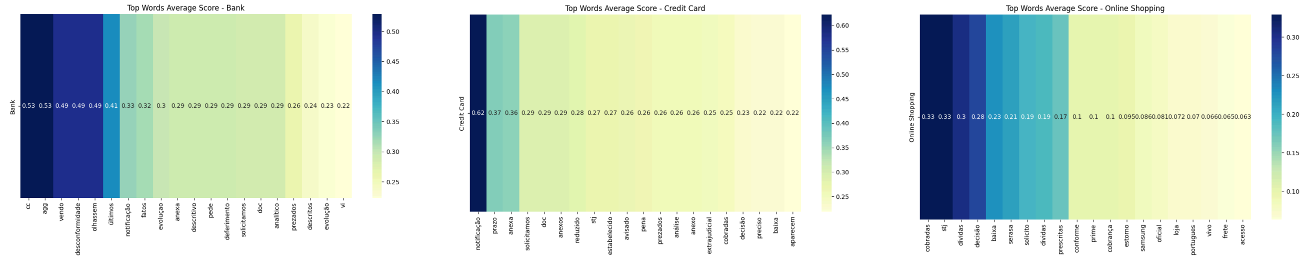
(a) Bank (b) Credit Card (c) Online Shopping

Figure 8. Top companies per area.



(a) Bank (b) Credit Card (c) Online Shopping

Figure 9. Word clouds by domain (Bank, Credit Card, Online Shopping). Larger words indicate higher within-domain relative frequency after deduplication; frequent terms summarize the dominant themes raised by consumers.



(a) Bank (b) Credit Card (c) Online Shopping

Figure 10. Top-ranked terms by domain based on average importance scores. Darker bars indicate higher within-domain discriminative importance; scores summarize each term's contribution to distinguishing that domain's complaints. This highlights domain-specific vocabulary beyond raw frequency.

in topic complexity. The confidence intervals were computed using non-parametric bootstrapping: for each cluster, 1,000 random resamples (with replacement) of the annotated items were generated, and the Kappa value was recalculated for each sample. The CI corresponds to the 2.5th and 97.5th percentiles of the resulting distribution.

Table 8. Fleiss' Kappa values per cluster, with 95% confidence intervals, for the three domains (Credit Card, Bank, and Online Shopping). The measure quantifies inter-rater agreement in the manual validation of collective demand clusters.

Group	Cluster	Kappa	CI Low	CI Up
Credit Card	1	0.7000	0.2039	1.0000
Credit Card	2	0.3096	-0.2097	0.7285
Credit Card	3	0.8295	0.4231	1.0000
Bank	4	0.5982	0.1346	1.0000
Bank	5	0.3096	-0.2097	0.7285
Bank	6	0.5833	-0.1111	1.0000
Online Shopping	7	1.0000	1.0000	1.0000
Online Shopping	8	0.7012	0.2581	1.0000
Online Shopping	9	0.8661	0.5200	1.0000

Figure 11 shows the Fleiss' Kappa values per cluster, grouped by topic, along with 95% confidence intervals (CI) represented as error bars. The length of each bar reflects the statistical stability of the agreement estimate. Clusters with shorter intervals, such as Clusters 3, 8, and 9, indicate more robust agreement and greater consistency among annotators. In contrast, Clusters 2, 5, and 6 exhibit wide confidence intervals, suggesting substantial uncertainty and variability in the ratings. Cluster 7 stands out with perfect agreement ($\kappa = 1.000$) and no interval, reflecting unanimous decisions across all raters. Despite variation in point estimates, the overall trend suggests that many clusters achieved moderate to substantial agreement.

In addition to the per-cluster analysis, we also computed Fleiss' Kappa values aggregated by group and across the entire dataset. As shown in Table 9, the Online Shopping group achieved the highest level of agreement among annotators, with a mean Kappa of 0.8558, indicating near-perfect consensus. The Credit Card group followed with a mean Kappa of 0.6131, reflecting substantial agreement, while the Bank group showed a moderate level of agreement with a mean Kappa of 0.4971. When considering all annotated instances together, the overall Fleiss' Kappa was **0.6580**, which also corresponds to a substantial agreement level. These results suggest that, although some variation exists between topics, the annotation process was overall consistent and reliable across different domains.

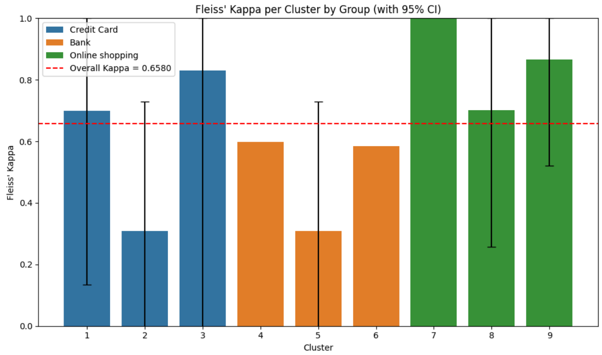


Figure 11. Fleiss' Kappa per cluster with 95% confidence intervals. Error bars were computed using 1,000 bootstrap resamples per cluster.

Differences in inter-rater agreement within groups may stem from topic-specific challenges in both manual interpretation and automated clustering. Some domains involve more ambiguous or heterogeneous complaint narratives, making it harder for annotators to consistently judge cluster coherence. These contextual factors likely contributed to the observed variation in Kappa values across groups.

Figure 12 illustrates the distribution of labels assigned by annotators across the entire dataset and within each thematic

Table 9. Mean Fleiss' Kappa values per domain (Bank, Credit Card, and Online Shopping) and overall, with 95% confidence intervals. The results summarize inter-rater agreement in the manual validation of collective demand clusters across groups.

Group	Mean Kappa	CI Low	CI Up
Bank	0.4971	0.3233	0.5975
Credit Card	0.6131	0.3291	0.8231
Online Shopping	0.8558	0.7094	0.9933
Overall	**0.6580**	**0.5281**	**0.7690**

group. In the overall distribution (Figure 12a), the proportions of "No" (49.4%) and "Yes" (49.0%) are nearly identical, while the "Inconclusive" label accounts for only 1.5% of cases. This near-equal distribution between positive and negative judgments suggests frequent disagreement among annotators, which is reflected in the moderate variability observed in the Fleiss' Kappa values.

Upon examining each group individually, distinct patterns emerge. In the Banking category (Figure 12b), the distribution remains relatively balanced between "Yes" and "No," which may reflect ambiguities in the financial complaint texts and the inherent challenges in interpreting cluster coherence within that domain. The Credit Card group (Figure 12c) demonstrates a slight predominance of "Yes" labels, suggesting greater perceived cohesion among the complaints in those clusters. Conversely, the Online Shopping group (Figure 12d) exhibits a clearer dominance of "Yes" labels, with very few "No" or "Inconclusive" responses, indicating stronger agreement and clearer thematic consistency.

These results address the second part of our Research Question, indicating that clustering semantically similar complaints with BERTopic yields interpretable groups of potential collective demands. The manual validation with substantial agreement (Fleiss' $\kappa = 0.658$, CI 95%) confirms the practical and interpretable value of this grouping for institutional use.

These distributions provide insights into the observed variation in inter-rater agreement across the groups. Topics characterized by more homogeneous and well-defined content tend to yield higher agreement, whereas more ambiguous or diverse topics result in more balanced or uncertain label assignments.

7 Conclusion and Future Work

This study introduced a computational approach addressing two major challenges: (i) the identification of duplicate consumer complaints by combining temporal patterns with key attributes — such as complainant identity, service provider, and complaint subject — across distinct platforms, and (ii) the detection of collective demands through clustering methods based on semantic similarity.

For the first challenge (i), we conducted a comprehensive characterization of duplicate consumer complaints submitted across three repositories. The proposed methodology integrated temporal analysis, semantic similarity assessment, and attribute-based matching, supported by natural language processing (NLP) techniques tailored for Brazilian Portuguese. This enabled accurate detection and interpretation of duplicate records.

The analysis yielded several key findings with practical implications for complaint management:

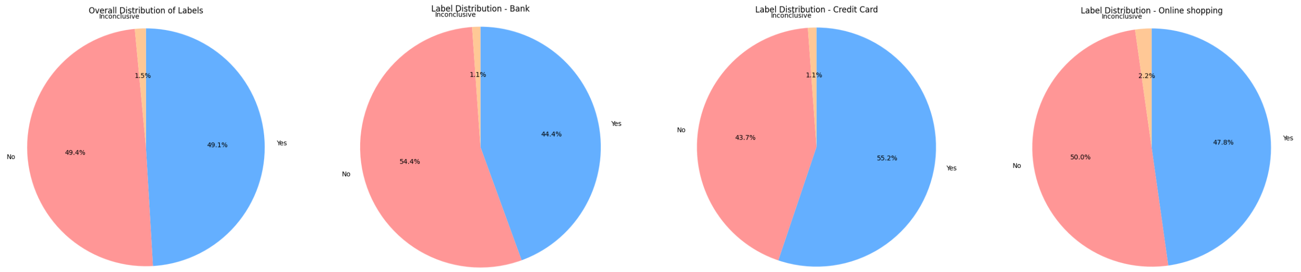
- 95% of duplicates were submitted within 30 days of the original complaint, establishing a clear temporal window for detecting repeated entries.
- The textual content of duplicates reflected domain-specific reporting behaviors, with issue frequency varying across platforms.
- In most cases, duplicates did not exceed two per original complaint, suggesting a normative duplication pattern.
- A significant portion of duplicates was submitted on the same day as the original, evidencing the common practice of rapid resubmission.

The proposed methodology also demonstrated high potential for identifying collective consumer demands within unstructured complaint data (i.e., challenge (ii)). By integrating semantic embeddings with topic modeling techniques — using BERTopic — the approach effectively grouped complaints into semantically coherent clusters that captured recurring grievances across various service domains. Manual validation performed by independent annotators confirmed a significant degree of internal consistency within the clusters, yielding a Fleiss' Kappa score of 0.6580. Among the domains analyzed, the credit card sector exhibited the highest average number of complaints per cluster, indicating a greater concentration of dissatisfaction around a narrower set of well-defined issues.

The designed solution has been successfully integrated and is undergoing evaluation within the production environment of PROCON-MG, the Consumer Protection Agency of the State of Minas Gerais. It has been fully incorporated into the ProconData System, a platform developed by this same research group and adopted as a primary technological infrastructure by PROCON-MG for the processing and analysis of consumer complaints. Among its core functionalities, the system incorporates the automated identification of collective demands, enabling the detection of recurrent patterns and emerging consumer issues at scale. This successful integration and ongoing evaluation demonstrate the solution's potential as a strategic instrument for regulatory and consumer protection authorities, as it enhances the capacity for early identification of collective demands and provides a robust foundation for data-driven, targeted interventions in consumer defense.

A main limitation of the study lies in handling complaints generated from standardized text templates, which often led to erroneous duplicate classifications. This issue could be mitigated by incorporating user-specific information through Named Entity Recognition (NER) and by applying differential weighting to named entities, thereby improving both duplicate detection and the identification of collective demands.

Another limitation concerns the detection of collective demands, which relied on manually predefined keywords and filters. While effective for controlled experiments, this approach reduces adaptability to dynamic scenarios. Furthermore, while the results and insights presented are grounded in real deployments, we acknowledge that they are specific



(a) Overall (b) Bank (c) Credit Card (d) Online Shopping

Figure 12. Distribution of labels assigned by annotators for each group and overall. The pie charts show the proportion of votes across categories: *Yes*, *No*, and *Inconclusive*.

to Brazilian Portuguese consumer complaints and regulatory frameworks. The system's performance and utility may vary in different jurisdictions, languages, or domains. Future work is needed to assess generalizability across broader datasets and institutional contexts.

Future research should explore automated or semi-automated mechanisms capable of generalizing beyond fixed keyword sets and identifying emerging issues in real time. We also plan to broaden the scope of the analysis by incorporating additional complaint repositories, such as Reclame AQUI[7], and by experimenting with alternative language models such as BERTabaporu, and assess the robustness and feasibility of deployment, conducting efficiency analysis.

Acknowledgements

This work was partially funded by Ministério Público do Estado de Minas Gerais (MPMG), the Analytical Capabilities Project, and supported by CNPq, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional e Tratamento e Disseminação de Informação (INCT-TILD-IAR), CAPES, FAPEMIG, and AWS.

Authors' Contributions

All authors contributed equally to the conception of this study. GR, JV and GK planned, implemented the approaches, and performed the analysis. JR, ZC, RP and MG contributed with the design of methodology and experiments. GR is the main contributor and writer of this article. All authors read, and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

Almeida, T. N. V. d. and Ramos, A. S. M. (2012). Os impactos das reclamações on-line na lealdade dos consumidores: um estudo experimental. *Revista de Adm. Contemporânea*, 16:664–683.

Barlaug, N. and Gulla, J. A. (2021). Neural networks for entity matching: A survey. *ACM Transactions on Knowledge Discovery from Data*, 15(3):1–37. DOI: 10.1145/3442200.

Barz, B. and Denzler, J. (2020). Do We Train on Test Data? Purging CIFAR of Near-Duplicates. *Journal of Imaging*, 6(6):41.

Bastani, K., Namavari, H., and Shaffer, J. (2019). Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, 127:256–271. DOI: https://doi.org/10.1016/j.eswa.2019.03.001.

Belém, F. M., de Andrade, C. M. V., França, C., Carvalho, M., Ganem, M. A. S., Teixeira, G., Jallais, G., Laender, A. H. F., and Gonçalves, M. A. (2023). Contextual reinforcement, entity delimitation and generative data augmentation for entity recognition and relation extraction in official documents. *J. Inf. Data Manag.*, 14(1).

BlackFriday.com.br (2024). How does Black Friday work? Accessed on: April 20, 2025.

Carvalho, M., Mangaravite, V., Ponce, L. M., Cantelli, L., Campoi, B., Nunes, G., de Paiva, B. B. M., Laender, A. H. F., and Gonçalves, M. A. (2022). Deduplicating large volumes of data from natural and legal entities in the governmental field. In *IEEE International Conference on Big Data*, pages 2206–2213.

de Andrade, C. M. V., Belém, F., Cunha, W., França, C., Viegas, F., Rocha, L., and Gonçalves, M. A. (2023). On the class separability of contextual embeddings representations - or "the classifier does not matter when the (text) representation is so good!". *Inf. Process. Manag.*, 60(4):103336.

de Carvalho, A. P., Ferreira, A. A., Laender, A. H. F., and Gonçalves, M. A. (2011). Incremental unsupervised name disambiguation in cleaned digital libraries. *J. Inf. Data Manag.*, 2(3):289–304.

de Carvalho, M. G., Gonçalves, M. A., Laender, A. H. F., and da Silva, A. S. (2006). Learning to deduplicate. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 41–50.

de Carvalho, M. G., Laender, A. H. F., Gonçalves, M. A., and da Silva, A. S. (2008). Replica identification using genetic programming. In *Proc. of the ACM Symposium on Applied Computing (SAC)*, pages 1801–1806.

dos Santos, A., Alves, D., and Braga, R. (2023). Topic modelling on consumer financial protection bureau data: An approach using bert-based embeddings. *ResearchGate*.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection: A Survey.

[7] `https://www.reclameaqui.com.br/`

IEEE Transactions on Knowledge and Data Engineering, 19(1):1–16.

Félix, L. G. S., Silveira, J. V., Luiz, W., Dias, D., and Rocha, L. (2018). Avaliação Automática de Conteúdo de Aplicações de Reclamação Online. In *Anais do Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, pages 49–56.

Fleiss, J. *et al*. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Freitas, M. d. S. and Andreão, R. V. (2021). Automatização do Processamento do Texto Bruto Oriundo de um Serviço de Atendimento de Reclamações. In *Anais da Escola Regional de Informática do Rio de Janeiro (ERI-RJ)*, pages 72–79.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211.

Loshin, D. (2010). *Master data management*. Morgan Kaufmann.

Mangaravite, V., Carvalho, M., Cantelli, L., Ponce, L. M., Campoi, B., Nunes, G., Laender, A. H. F., and Goncalves, M. A. (2022). DedupeGov: Um Ambiente para Deduplicação de Grandes Volumes de Dados de Pessoas Físicas e Jurídicas em Âmbito Governamental. In *Anais do Simp. Bras. de Banco de Dados (SBBD)*, pages 90–102.

Mansoor, M., Rehman, Z. U., Shaheen, M., Khan, M. A., and Habib, M. (2020). Deep Learning based Semantic Similarity Detection using Text Data. *Information Technology And Control*, 49(4):495–510.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282. DOI: 10.11613/BM.2012.031.

Miller, F. P., Vandome, A. F., and McBrewster, J. (2009). *Levenshtein Distance: Information theory, Computer science, String (computer science), String metric, Damerau?Levenshtein distance, Spell checker, Hamming distance*. Alpha Press.

Mourão, F., Rocha, L., Araújo, R. B., Couto, T., Gonçalves, M. A., and Jr., W. M. (2008). Understanding temporal aspects in document classification. In *Proc. of the Int. Conf. on Web Search and Web Data Mining (WSDM)*, pages 159–170.

Rabbi, G., Araújo, M. M., Kakizaki, G., Viterbo, J., Reis, J. C., Prates, R. O., and Gonçalves, M. A. (2024). Identificação e caracterização de reclamações duplicadas por consumidores em múltiplas plataformas. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 313–326.

Reclame Aqui (2024). Black Friday 2024: Consumers report an increase in problems with online purchases. Accessed on: April 20, 2025.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv*.

Ripon, K. S. N., Rahman, A., and Rahaman, G. A. (2010). A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates. *Journal of Computers*, 5(12):1800–1809.

Salles, T., Rocha, L., Pappa, G. L., Mourão, F., Meira, W., and Gonçalves, M. (2010). Temporally-aware algorithms for document classification. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 307–314.

Sargiani, V., de Castro, L. N., and Silva, L. A. (2020). A data mining study of sindec complaints in the period 2013-2017. In *Proc. of the Int. Conf. on Internet Techn. & Society (ITS) and Sustainability, Techn. and Education (STE)*, pages 35–45.

Silva, L. S., Canalle, G. K., Salgado, A. C., Lóscio, B. F., and Moro, M. M. (2019). Uma Análise Experimental do Impacto da Seleção de Atributos em Processos de Resolução de Entidades. In *Anais do Simp. Bras. de Banco de Dados (SBBD)*, pages 37–48.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Braz. Conf. on Intelligent Systems (BRACIS)*, pages 403–417.

Vaishnav, D., Neethinayagam, M., Khaire, A., and Woo, J. (2024). Predictive analysis of cfpb consumer complaints using machine learning.

Wang, Y., Qin, J., and Wang, W. (2017). Efficient approximate entity matching using jaro-winkler distance. In *Web Inf. Systems Engineering (WISE)*, pages 231–239.