

Identificação e Caracterização de Reclamações Duplicadas por Consumidores em Múltiplas Plataformas

Gestefane Rabbi¹, Marcelo M. R. Araújo¹, Gabriel Kakizaki², Julia Viterbo¹,
Julio C. S. Reis², Raquel O. Prates¹, Marcos André Gonçalves¹

¹Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

²Departamento de Informática – Universidade Federal de Viçosa (UFV)

{gestefane,marceloaraujo,juliapaes,rprates,mgoncalv}@dcc.ufmg.br

{gabriel.kakizaki,jreis}@ufv.br

Abstract. *The growing volume of data in complaints repositories of consumers poses significant challenges for the effective management of this information. Among these challenges, the fact that many complaints are registered more than once, by the same consumer, to put pressure on companies stands out, which can impact the management of these records and distort analyses based on this data. This study proposes an approach to identify duplicates using temporal analysis and attributes such as consumer, supplier and object of the complaint from consumers records on different platforms. In this sense, natural language processing techniques are explored, specifically the BERTimbau model, to detect semantic similarities between complaints. The results show that 95% of duplicates are posted within 30 days of the original. The proposed approach contributes to improving the accuracy in detecting duplicates and the efficiency in managing this type of (unstructured) data, benefiting conflict resolution and complaints administration by competent entities.*

Resumo. *O crescente volume de dados em repositórios de reclamações de consumidores impõe desafios significativos para a gestão eficaz dessas informações. Dentre estes desafios destaca-se o fato de que muitas reclamações são registradas mais de uma vez, por um mesmo consumidor, para pressionar as empresas, o que pode impactar a gestão desses registros e distorcer análises baseadas nestes dados. Este estudo propõe uma abordagem para identificar duplicatas usando análise temporal e atributos como consumidor, fornecedor e objeto da reclamação a partir de reclamações registradas por consumidores em diferentes plataformas. Neste sentido são exploradas técnicas de processamento de linguagem natural, especificamente o modelo BERTimbau, para detectar similaridades semânticas entre reclamações. Os resultados mostram que 95% das duplicatas são postadas em até 30 dias após a original. A abordagem proposta contribui para melhorar a precisão na detecção de duplicatas e a eficiência na gestão desse tipo de dado (não-estruturado), beneficiando a resolução de conflitos e a administração das reclamações por entidades competentes.*

1. Introdução

A Internet se popularizou em todo mundo, e, com isso, é crescente o número de usuários que utilizam sistemas Web para realização de compras online. Pesquisas recentes reve-

laram que 62% dos consumidores fazem de duas a cinco compras online mensalmente¹. Mais do que isso, em países como o Brasil, cerca de 85% da população com acesso à Internet faz pelo menos uma compra online por mês, o que já indica uma preferência dos brasileiros para compras online em comparação às realizadas em lojas físicas².

Com o aumento do número de compras realizadas nesse ambiente, cresce também o número de reclamações dos consumidores em relação aos produtos e/ou serviços adquiridos devidos a inúmeros problemas (*e.g.*, defeitos, atrasos, cobranças indevidas, etc). Com isso, emergiram várias plataformas que permitem aos usuários manifestarem seu descontentamento. Exemplos dessas plataformas incluem “Reclame AQUI”, “Consumidor.gov.br” e “Denúncia”³. Com o tempo, o uso desses serviços se expandiu para registrar também reclamações de consumidores sobre compras ou serviços contratados fora do ambiente online, como telefonia e bancos.

De forma geral, essas plataformas fornecem mecanismos para usuários expressarem suas opiniões e/ou (in)satisfações em relação a determinado produto ou serviço, permitindo aos consumidores exigirem (e muitas vezes receberem) um posicionamento das empresas acerca do fato ocorrido. Em outras palavras, a partir de uma reclamação cadastrada por um consumidor qualquer, empresas também registradas nessas plataformas são publicamente pressionadas a se manifestarem. Logo, para fins de credibilidade e/ou reputação desses estabelecimentos reclamados, um posicionamento adequado e em tempo hábil em relação à reclamação é de extrema importância.

Entretanto, isso nem sempre ocorre em um intervalo de tempo razoável para o reclamante – ou sequer ocorre, o que faz com que consumidores cadastrem novas reclamações, a fim de aumentar a pressão para receberem um posicionamento por meio da insistência. Com isso, cresce o número de reclamações registradas em duplicidade nas plataformas⁴, impondo desafios significativos para a gestão adequada dessas informações. Essa multiplicidade de dados pode prejudicar a resolução de conflitos, uma vez que uma contagem de ocorrências errada, para maior, pode induzir a conclusões sobre alta incidência de reclamações, por exemplo, sobre um produto ou empresa, e de fato não retratar a realidade. Assim, mecanismos que auxiliem essas plataformas na gestão adequada de tal conteúdo são de suma importância. É nesse contexto em que se insere o objetivo deste estudo.

Particularmente, exploramos 1.723.245 reclamações cadastradas por consumidores desde 2006 em 3 plataformas distintas, a saber: Consumidor.gov.br, Procon e Sindec. Neste contexto, propusemos uma abordagem para identificação de reclamações duplicadas que considera a análise temporal [Mourão et al. 2008] das reclamações, bem como o exame de atributos tais como: o consumidor (reclamante), o fornecedor (reclamado) e o objeto da reclamação. Além disso, a abordagem proposta neste estudo explora técnicas avançadas de processamento de linguagem natural (PLN) para a Língua Portuguesa do Brasil, especificamente o modelo BERTimbau [Souza et al. 2020], que são mais robustas à variabilidade da linguagem natural presente nas reclama-

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>

³reclameaqui.com.br, consumidor.gov.br, denuncia.com.br

⁴No contexto deste projeto, consideramos como *duplicata* uma reclamação feita por um mesmo consumidor, para a mesma empresa, em um intervalo de tempo de 30 dias. Essas definições foram efetuadas com base em interação com usuários especialistas do Ministério Público de Minas Gerais (MPMG).

ções dos consumidores, em comparação a abordagens clássicas como *fuzzy matching* [Wang et al. 2017] ou outras técnicas de aprendizado de máquina baseadas em *soft computing* [de Carvalho et al. 2006, de Carvalho et al. 2008]. A ideia geral é detectar similaridades semânticas entre reclamações por meio do estabelecimento de limiares adequados para a identificação dessas duplicatas.

Nossos resultados revelam que 95% das duplicatas foram postadas com intervalos de até 30 dias após a primeira postagem, indicando uma faixa temporal de persistência dos consumidores em buscar resolução para seus problemas. Além disso, o comprimento e nuvens de palavras das duplicatas entre os três repositórios de dados destacam diferenças significativas no conteúdo e nos padrões de postagens das reclamações duplicadas, como, por exemplo, referências a fragmentos de leis ou mais menções sobre serviços ou empresas, refletindo as particularidades de cada base de dados. Logo, as descobertas deste estudo podem contribuir para identificar diferentes padrões de comportamento de usuários nos três órgãos analisados, auxiliando no gerenciamento mais adequado dos dados.

Vale destacar certos aspectos práticos e legais da solução proposta. Por exemplo, a detecção de duplicatas é um passo essencial para identificar possíveis *demandas coletivas* a partir de manifestações individuais. Demandas coletivas, segundo o Artigo 81º do Código de Defesa do Consumidor, são “de natureza indivisível de que seja titular grupo, categoria ou classe de pessoas ligadas entre si ou com a parte contrária por uma relação jurídica base” e são passíveis de defesa legal. Dados duplicados podem “inflar” indevidamente certas análises, por meio da multiplicação de casos particulares, identificando (erroneamente) demandas como sendo coletivas, já que supostamente afetam uma parte significativa da população, quando na verdade são poucos os indivíduos afetados, mas que atuam de forma contundente nas plataformas. Nosso estudo também pode fornecer embasamento para futuras análises focadas na identificação da coexistência de reclamações sobre o mesmo fornecedor e/ou produto em diferentes órgãos (intra ou entre repositórios), suportando outros tipos de análises e ações legais.

O restante deste artigo está organizado da seguinte forma. A próxima seção apresenta trabalhos relacionados. Na Seção 3, descrevemos detalhes da abordagem proposta para identificação de reclamações duplicadas. Os principais resultados são discutidos na Seção 4. A Seção 5 conclui o estudo e apresenta direções para trabalhos futuros.

2. Trabalhos Relacionados

Descrevemos aqui trabalhos relacionados considerando duas dimensões principais: (i) reclamações de consumidores e, (ii) estratégias gerais para remoção de informações duplicadas em repositórios de dados de diferentes contextos (não exclusivamente reclamações).

(i) Reclamações de Consumidores. Plataformas online têm sido amplamente exploradas por consumidores para expressarem sua insatisfação acerca de determinado produto e/ou serviço [Almeida and Ramos 2012]. Esse fenômeno proporciona a geração de uma quantidade significativa de informação que tem atraído a atenção de pesquisadores de diversas áreas com objetivos distintos. Em [Sargiani et al. 2020], por exemplo, os autores exploram dados de reclamações de consumidores com o propósito de identificar setores de mercado mais problemáticos. Para isso, o estudo coletou reclamações da base de dados do Sincdec, no período de 2013 a 2017, e, após identificar o setor bancário como tendo o maior número de reclamações, em torno de 90 mil, focou as análises nessas reclamações. Os resultados destacam que bancos de varejo possuem o maior número de

reclamações, e que a maior parte delas, cerca de 10 mil, são sobre cobranças indevidas.

Já Félix *et. al* [2018] explora técnicas de processamento de linguagem natural para classificação de tópicos e/ou assuntos comumente presentes em reclamações sobre operadoras de telefonia. Com essa finalidade, foram coletados dados do Procon, Reclame AQUI e também do Twitter, totalizando cerca de 300 mil reclamações. Em seguida, foi aplicado o algoritmo *Latent Dirichlet Allocation* (LDA) [Jelodar et al. 2019] para a modelagem de tópicos. Os resultados apontam que os problemas se concentram principalmente nos serviços de telefonia e Internet, e que a base do Reclame AQUI pode fornecer melhores informações do que as outras para auxiliar as empresas a solucionarem problemas. Uma abordagem para automatização e limpeza de dados brutos de um serviço de atendimento de reclamações de consumidores é apresentada em [Freitas and Andreão 2021], com o intuito de colaborar na construção de classificadores automáticos. Foram utilizadas nesse estudo cerca de 13 mil reclamações sobre empresas de telefonia e TV por assinatura, coletadas da plataforma Anatel Consumidor⁵. Como resultado da aplicação da metodologia proposta, houve uma redução no tempo de treinamento de modelos de classificação, sem alterações significativas na acurácia.

(ii) Estratégias para Remoção de Informações Duplicadas. Abordagens para a identificação e remoção de dados duplicados têm sido propostas considerando diferentes domínios de aplicação [Elmagarmid et al. 2007, Ripon et al. 2010, Mansoor et al. 2020, de Carvalho et al. 2011]. Em suma, o tratamento inadequado de informações duplicadas pode culminar em vários problemas, como o aumento do custo de processamento e degradação do desempenho de modelos treinados sobre esses dados [Barz and Denzler 2020]. Trabalhos nessa linha se assemelham ao nosso ao analisarem ou avaliarem a capacidade discriminativa dos atributos no processo de resolução de entidades [Mangaravite et al. 2022, Carvalho et al. 2022, Belém et al. 2023, Silva et al. 2019] Além disso, estudos focados na remoção de informações duplicadas historicamente tiveram maior ênfase em dados estruturados, como registros de bases de dados tabulares, e geralmente utilizando métodos baseados em regras, manualmente confeccionadas ou aprendidas por meio de técnicas de aprendizado de máquina [Elmagarmid et al. 2007, Mangaravite et al. 2022, Carvalho et al. 2022]. Esses métodos, porém, não são adequados para aplicação em dados textuais de alta dimensionalidade, principalmente em textos ruidosos gerados por usuários finais. Nesse caso, estudos mais recentes têm aplicado *embeddings* semânticos e modelos de *deep learning* [Mansoor et al. 2020, de Andrade et al. 2023a] para a tarefa.

Lacuna de Pesquisa. Este estudo se diferencia dos trabalhos existentes na literatura apresentados em (i), uma vez que está focado na análise de reclamações duplicadas de consumidores, provenientes de diferentes plataformas governamentais, as três citadas anteriormente, de forma conjunta. Para tal, foram aplicadas técnicas de pré-processamento de dados a fim de garantir maior acurácia nas medições quantitativas realizadas, como, por exemplo, o uso de técnicas de PLN, as quais não foram tão amplamente exploradas em trabalhos anteriores para a realização deste tipo de análise. Em relação aos trabalhos focados em estratégias para remoção de informações duplicadas (ii), que comumente focam na deduplicação de dados estruturados, incluindo atributos tais como CPF, CNPJ, etc, nossa abordagem lida com dados não estruturados, especificamente textos de

⁵<https://apps.anatel.gov.br/AnatelConsumidor/>

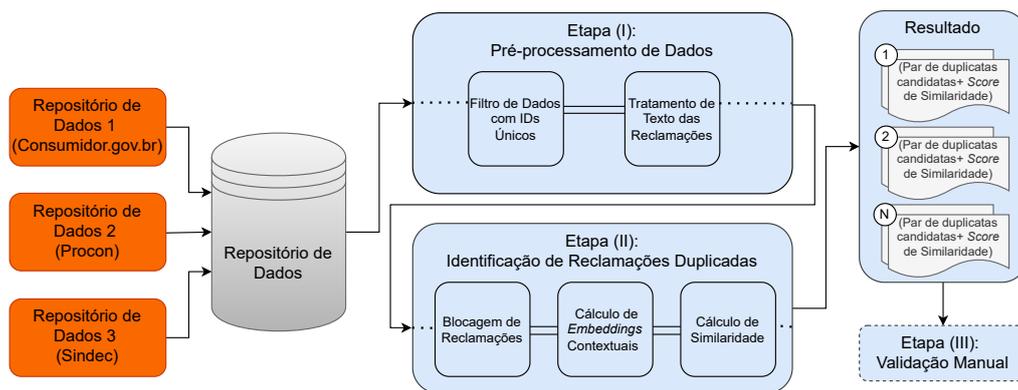


Figura 1. Abordagem proposta para identificação de reclamações duplicadas.

reclamações de consumidores, abrangendo nuances linguísticas que são bastante comuns neste contexto. Por fim, o fato do trabalho explorar textos na Língua Portuguesa, para o qual não existem tantos recursos linguísticos disponíveis em comparação com a Língua Inglesa, é também um diferencial.

3. Abordagem Proposta

Detalhes relativos à abordagem proposta neste estudo para identificação de reclamações duplicadas em múltiplos repositórios de dados são apresentados nesta seção. A Figura 1 apresenta uma visão geral, envolvendo cada uma das principais etapas detalhadas a seguir.

3.1. Repositórios de Dados

Neste artigo, exploramos dados oriundos de 3 repositórios distintos: Consumidor.gov.br, Procon e Sindec⁶. Uma descrição de cada repositório é apresentada a seguir.

Consumidor.gov.br. O governo federal possui disponível um serviço público e gratuito que permite a comunicação direta entre consumidores e empresas para a solução de problemas de consumo. Em linhas gerais, consumidores acessam esse serviço para se comunicarem com as empresas. Por se tratar de um serviço do governo, é necessário que as empresas se cadastrem e assinem um termo de comprometimento, assim como os consumidores devem possuir uma conta nível prata ou ouro no gov.br. Este serviço, acessível por meio do endereço `consumidor.gov.br` está disponível desde 2021.

Procon. Órgão público de defesa do consumidor que, dentre diversas outras funções, recebe reclamações de consumidores para mediar possíveis soluções com a empresa. Concebido com o objetivo de proteger o consumidor, esse órgão também atua contra práticas de empresas que prejudiquem o coletivo. É possível se manifestar contra uma empresa pela Internet, telefone, ou também presencialmente em alguma unidade do Procon.

Sindec. O Sistema Nacional de Informações de Defesa do Consumidor (Sindec) se refere a um conjunto de tecnologias utilizado para integrar e consolidar informações sobre os órgãos de defesa ao consumidor, visando proporcionar um instrumento de gestão adequado ao dinamismo típico de seus setores de atendimento.

No total, coletamos 951.814, 427.239 e 344.192 cadastradas nas plataformas Consumidor.gov.br, Procon e Sindec, respectivamente, registradas desde 2006, conforme

⁶É importante destacar os dados explorados neste estudo foram disponibilizados pela entidade financeira e contém informações sensíveis/sigilosas não compartilháveis.

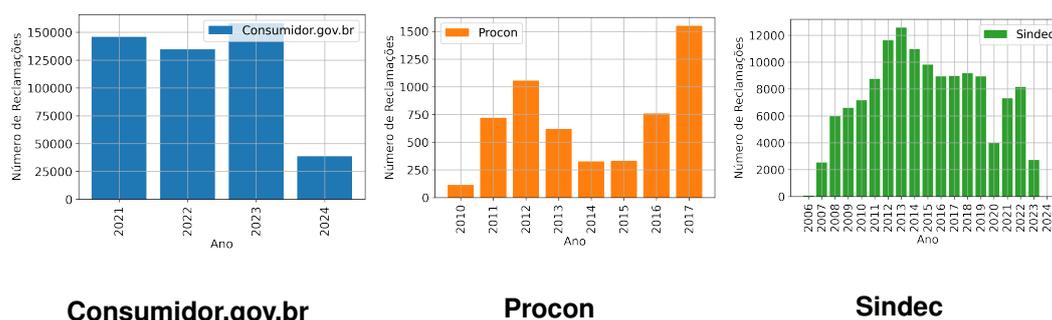


Figura 2. Número de reclamações por ano por repositório.

distribuição apresentada na Figura 2. Notamos um aumento significativo no volume de reclamações a partir de 2021 provavelmente associado ao lançamento da plataforma Consumidor.gov.br. Por fim, as 1.723.245 reclamações, são armazenadas em um repositório de dados único, utilizado como entrada para as etapas subsequentes.

3.2. Etapa I: Pré-Processamento dos Dados

A primeira etapa do processo de identificação de duplicatas é o pré-processamento, que é realizado utilizando os identificadores únicos de reclamantes e reclamados, obtidos através do processo de *Master Data Management* (MDM) [Loshin 2010]. Detalhes da implementação utilizada neste estudo podem ser obtidos em [Mangaravite et al. 2022]. Em resumo, a execução deste processo garante que cada entidade utilizada seja identificada de forma única e consistente em todos os repositórios de dados integrados, permitindo uma análise mais precisa e eficiente. A Tabela 1 apresenta a quantidade de reclamações oriundas de cada um dos repositórios de dados explorados neste trabalho, depois da execução desta etapa. Em suma, podemos observar que a plataforma Consumidor.gov.br possui o maior volume de reclamações cadastradas (477.436), seguido de Sindec e Procon, com 134.198 e 5.476 registros, respectivamente. Curiosamente, o Consumidor.gov.br possui o maior número de reclamantes (213.478), enquanto o Sindec, o maior número de alvo de reclamações (13.754). Além disso, cada plataforma recebeu em média ≈ 2 reclamações por reclamante e 35 por reclamado (*i.e.*, empresa alvo da reclamação).

3.3. Etapa II: Estratégia para Identificação de Reclamações Duplicadas

O objetivo desta etapa é detectar reclamações provenientes de diversos consumidores, direcionadas a uma mesma empresa ou produto em um intervalo de tempo específico. Essa etapa é fundamental para permitir a geração, de maneira mais precisa, de análises quantitativas dos dados. Isso se dá porque os consumidores podem, em face de um mesmo evento que gera a necessidade de acionamento destes órgãos, criar diversas chamadas em um mesmo órgão ou em órgãos distintos, podendo inflar artificialmente as estatísticas de reclamação de empresas e setores específicos. Nesse sentido, a proposta é dividida em 3 etapas principais descritas a seguir.

Tabela 1. Sumário de reclamações por repositório com identificadores únicos de consumidor (ou reclamante) e empresa (ou reclamado).

Repositório	Reclamações	Reclamantes	Reclamados	Média por Reclamante	Média por Reclamado	Período
Consumidor.gov.br	477.436	213.478	845	2.2	565.0	01/01/2021–31/03/2024
Procon	5.476	5.065	2.814	1.1	1.9	11/08/2010–10/11/2017
Sindec	134.198	98.426	13.754	1.4	9.8	10/10/2006–20/02/2024
Total	617.110	316.969	17.413	1.9	35.4	10/10/2006 - 31/03/2024

Blocagem de Reclamações. Primeiramente realizamos o processo de blocagem de reclamações, que consiste no agrupamento de reclamações realizadas por um mesmo usuário e direcionadas a uma mesma empresa. Essa blocagem é realizada utilizando os identificadores únicos de usuários e empresas, filtrados anteriormente. Dessa forma, criam-se diversos blocos, cada um constituído pelas reclamações de um dado consumidor direcionados a uma empresa específica, em um intervalo de um mês. Essa etapa visa reduzir a quantidade de comparações a serem feitas, reduzindo o custo computacional do processo.

Representação de Reclamações por Meio de *Embeddings* Contextuais. Depois de blocar as reclamações, aplicamos um modelo pré-treinado de *sentence-embeddings*, que utiliza a arquitetura *BERT* para encontrar *embeddings*, *i.e.*, representações vetoriais de sentenças que codificam o sentido semântico das reclamações, a fim de localizar pares de reclamações que possuam similaridade definida por um dado limiar. O modelo utilizado foi o BERTimbau [Souza et al. 2020], um modelo treinado com dados em Português Brasileiro, em conjunto com a biblioteca sBERT [Reimers and Gurevych 2019], que possibilita codificar uma reclamação num vetor semântico em um espaço latente, permitindo o cálculo de métricas vetoriais, como distância e similaridade entre reclamações, a fim de detectar ocorrências que apresentam um grau suficiente de similaridade.

Neste contexto, é importante mencionar que a necessidade de utilizar uma abordagem de aprendizado profundo em detrimento de abordagens clássicas, como *fuzzy matching* [Wang et al. 2017, Miller et al. 2009], dá-se devido à complexidade e variabilidade da linguagem natural presente nas reclamações dos consumidores, as quais incluem textos informais, possuindo muitas vezes erros ortográficos, variações de sinônimos e expressões idiomáticas. Essas nuances tornam a detecção de duplicatas um desafio significativo para métodos tradicionais, sendo menos robustos às variações linguísticas que podem estar presentes nas reclamações de consumidores. Ademais, modelos tradicionais são limitados na capacidade de detectar similaridade semântica entre reclamações, já que indivíduos podem realizar reclamações direcionadas a uma mesma entidade (reclamada) de maneiras diferentes, mas com um sentido semântico convergente.

Cálculo de Similaridade. A seguir, realizamos o cálculo de similaridade entre os vetores de *embeddings* obtidos na etapa anterior. Para isso, computamos distância de cosseno entre as reclamações para gerar o *score* de similaridade⁷. Após essa etapa, as similaridades são ordenadas de maneira decrescente, deixando a cargo do especialista de domínio definir o *threshold* de similaridade das reclamações para serem consideradas duplicatas. No contexto deste trabalho, a partir de uma avaliação experimental, definimos um *threshold* de 90% para *score* de similaridade, para realização das análises. Ou seja, caso duas reclamações possuam similaridade igual ou acima desse limiar e sejam realizadas num período de até 30 dias, conforme definido por especialista de domínio, serão consideradas duplicatas. Como resultado, são relacionados pares de duplicatas candidatas com o acréscimo do *score* de similaridade fornecido pela abordagem. Neste estudo, dentre as 617.110 reclamações analisadas (ver Tabela 1), identificamos 17.583 reclamações duplicadas nos diferentes repositórios. É importante notar que o *threshold* definido é bastante conservador. Em um cenário prático, esses resultados podem ser ordenados de forma decrescente para análise, o que pode beneficiar ainda mais o desempenho da estratégia proposta.

⁷Os *scores* assumem valores entre -1 e 1, sendo 1 indicando similaridade máxima, 0 ortogonalidade entre os vetores, isto é, nenhuma similaridade e -1 similaridade mínima, ou seja, vetores em direções opostas.

3.4. Etapa III: Validação Manual dos Resultados

Por fim, com o objetivo de avaliar qualitativamente o desempenho da abordagem proposta, rotulamos uma amostra aleatória de pares de duplicatas candidatas identificadas em cada um dos repositórios de dados. Especificamente, a partir dos resultados obtidos (*i.e.*, 17.583 duplicatas) selecionamos randomicamente 100 pares em cada repositório, exceto para o Procon, onde o total de pares de duplicatas candidatas identificadas pela abordagem é igual a 21. Ao todo, avaliamos manualmente 221 pares de reclamações. Todos os pares foram avaliados de forma independente por 3 voluntários que classificaram as instâncias como: (1)“Duplicata”, para os casos onde claramente percebe-se o registro duplicado de uma reclamação de origem; (2)“Não duplicata”, caso contrário, ou; (3)“Inconclusivo”, para os casos onde o avaliador foi incapaz de definir um rótulo. Para medir a concordância entre os avaliadores, calculamos o *Fleiss’ Kappa* [Fleiss et al. 1971]. As concordâncias entre os avaliadores foram classificadas como *total*, quando a atribuição de rótulos foi unânime entre os avaliadores; *parcial*, quando pelo menos um dos avaliadores rotulou diferentemente dos demais; e *nenhuma*, quando cada avaliador atribuiu um rótulo diferente a uma dada amostra. Os resultados desta etapa são apresentados na Tabela 2.

Considerando o repositório de dados do Consumidor.gov.br, observamos que 78% das instâncias foram rotuladas como duplicatas e 9% como não duplicatas, o que indica que 9% das ocorrências que o modelo classificou como duplicatas não são de fato duplicatas (segundo os avaliadores). Já para o Sindec o modelo classificou como duplicatas 30% das amostras consideradas não duplicadas pelos avaliadores. Por último, no Procon o modelo apontou como duplicatas a totalidade das amostras nas quais houve concordância total entre os avaliadores.

Em geral, foi obtido o índice *Fleiss’ Kappa* de 0.748, o que indica uma concordância substancial entre os avaliadores. Como ilustra a Figura 3b, houve concordância total em 86.4% do total de amostras avaliadas, percentual este correspondente a 191 amostras dentre as 221. A concordância parcial, em geral, ficou em 10.9%, ou seja, em 24 das 221 amostras. E não houve concordância alguma em apenas 6 das 221 amostras, o que representa 2.7%. A título de exemplo de discordância total entre os três avaliadores são apresentados dois textos na Tabela 3. Analisando as reclamações, nota-se alta semelhança entre os textos, incluindo a citação de um mesmo número de pedido, de algumas datas iguais e de um mesmo número de protocolo. Enquanto, no primeiro texto, cita-se apenas um pedido e um protocolo; no segundo, entretanto, citam-se mais dois pedidos e mais dois protocolos além dos presentes no outro texto. Nesse sentido, o rotulador que acreditou serem duplicatas pode ter considerado que o primeiro texto estaria englobado no segundo. Contudo, dentre as quatro datas presentes em cada reclamação (data de devolução, de chegada do pedido ao estabelecimento, de reembolso e de prazo informado), apenas duas delas coincidem nos textos, já que, por englobar mais pedidos, o segundo texto talvez tenha tomado as datas mínima ou máxima de um deles. Nesse caso, o avaliador que avaliou os textos como não duplicatas pode ter entendido que o primeiro texto, por possuir menos informações e até informações diferentes do que

Tabela 2. Resultados da rotulação manual de três avaliadores.

Repositório	Concordância (%)			São duplicatas (%)	
	Total	Parcial	Nenhuma	1 - Sim	2 - Não
Consumidor.gov.br	87.00	10.00	3.00	78.00	9.00
Procon	90.48	9.52	0.00	90.48	0.00
Sindec	85.00	12.00	3.00	55.00	30.00

Tabela 3. Exemplo de discordância total entre os avaliadores.

Texto 1	Texto 2
O pedido XXXXXXXXXXXXXXX3116 foi devolvido no dia 12/12 e chegou na <RECLAMADA>no dia 13/12, o reembolso deveria ocorrer até o dia 23/12 conforme o prazo informado, porém informaram que iria ocorrer até o dia 27/12, conforme os protocolos XXXXX358. Já se passaram mais de 10 dias do prazo para reembolso e ainda não ocorreu o reembolso do pedido.	Os pedidos de nºs XXXXXXXXXXXXXXX6431, XXXXXXXXXXXXXXX3116 e XXXXXXXXXXXXXXX9871 foram devolvidos no dia 12/12 e chegaram na <RECLAMADA>no dia 14/12, o reembolso deveria ocorrer até o dia 23/12 conforme o prazo informado, porém informaram que iria ocorrer até o dia 27/12, conforme os protocolos XXXXX942, XXXXX813 e XXXXX358. Já se passaram mais de 10 dias do prazo para reembolso e ainda não ocorreu o reembolso dos pedidos.

o segundo, não poderia ser considerado uma cópia dele. E, ao mesmo tempo, o avaliador que considerou as informações nos textos como inconclusiva pode ter considerado essas duas lógicas como cabíveis e não foi capaz de chegar a um veredito.

A Figura 3a mostra a distribuição dos rótulos atribuídos pelos três avaliadores no processo de avaliação manual. Observa-se uma consistência razoável entre os avaliadores com respeito às quantidades de amostras rotuladas como duplicatas (rótulo 1) e não duplicatas (rótulo 2) enquanto as inconclusivas (rótulo 3) apresentaram uma baixa ocorrência. A Figura 3c mostra que, dentre as amostras nas quais houve concordância total entre os avaliadores, 79.6% das que foram identificadas como duplicatas pelo modelo, são de fato duplicatas, confirmadas na avaliação manual. Isso mostra um desempenho satisfatório na abordagem proposta neste estudo. Futuramente, e considerando um cenário prático, a ideia é que esses resultados possam ser utilizados para melhoria da gestão de reclamações realizadas por essas plataformas.

4. Resultados

Apresentamos nesta seção os principais resultados obtidos pela abordagem proposta considerando reclamações de consumidores oriundas dos diferentes repositórios de dados.

4.1. Análise Geral

Primeiramente, conduzimos uma caracterização inicial dos resultados obtidos. Nos três repositórios analisados, houve um total de 617.110 reclamações que possuíam identificadores de reclamante e reclamado, conforme Tabela 1. Essa quantidade representa o universo de dados considerados para a análise de possíveis duplicatas. Assim, depois de executarmos a abordagem proposta, das 617.110 reclamações, 17.583 foram identificadas como duplicatas, o que corresponde a 2,8% do total. O repositório Consumidor.gov.br apresentou a maior quantidade de duplicatas, com 16.185 desses casos, representando 94,3% do total de duplicatas identificadas. No Sindec foram encontradas 816 duplica-

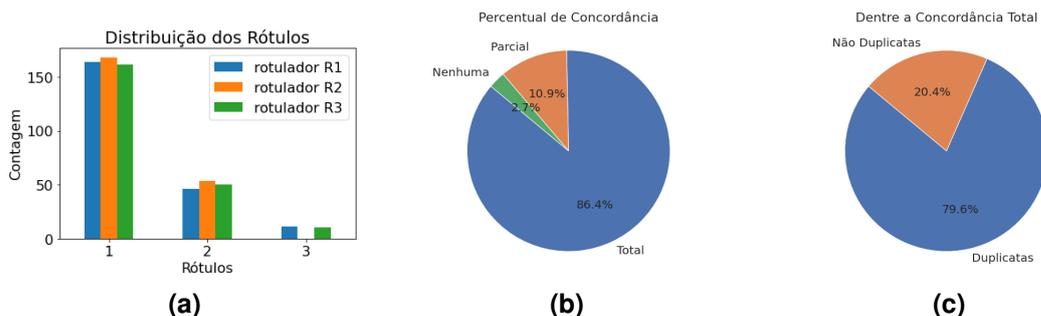


Figura 3. (a) Distribuição dos rótulos entre avaliadores. (b) Concordância entre avaliadores. (c) Confirmação de duplicatas.



Consumidor.gov.br Procon Sindec
Figura 5. Distribuição cumulativa de duplicatas por repositório.



Consumidor.gov.br Procon Sindec
Figura 6. Distribuição cumulativa de duplicatas por dia em cada repositório.

formas analisadas neste estudo. De forma geral, observamos que 95% das reclamações duplicadas têm no máximo 2 cópias além da original, havendo casos de até 19 duplicatas. A Figura 5 mostra as distribuições cumulativas em cada repositório, evidenciando, de modo individual, o fato de que 95% das reclamações tem até 2 cópias. Apesar que a base do Consumidor.gov.br ter uma quantidade de dados significativamente maior do que a dos outros domínios, observa-se que o padrão de até 2 duplicatas ocorrerem majoritariamente é consistente para todos os repositórios.

4.4. Temporalidade

Em seguida, focamos em entender aspectos temporais relacionados a duplicação de um registro das diferentes plataformas. Nota-se que quase a metade das duplicatas estudadas foram postadas no mesmo dia. As estatísticas para os dias entre postagens duplicadas mostram, para o Consumidor.gov.br, uma média de 8.54 (± 0.07) com desvio padrão de 10.21 (± 0.05) dias. Para o Sindec, a média é 4.83 (± 0.25) e desvio padrão de 9.23 (± 0.18) dias. No caso do Procon, a média é 0.34 (± 0.14) com desvio padrão de 2.06 (± 0.10) dias.

Na Figura 6 são apresentados os indicadores que evidenciam que 95% das postagens duplicadas são realizadas com menos de 30 dias em cada repositório separadamente. Chama a atenção a diferença substancial dos casos de duplicatas postadas no mesmo dia quando comparadas as diferentes bases de dados. No caso do Consumidor.gov.br, entre 40 e 50% das reclamações repetidas foram postadas no mesmo dia; no do Procon, mais de 90%; No Sindec, aproximadamente 70%. Uma possível explicação para isso seria a grande diferença na quantidade de duplicatas encontradas em cada um dos domínios. Para o Consumidor.gov.br, mais de 17 mil casos foram encontrados, contra 816 para o Procon e 182 para o Sindec, possivelmente indicando que as porcentagens encontradas para Procon e Sindec sejam pouco representativas do comportamento temporal mais geral das duplicatas, diferentemente do que foi observado na análise por usuário da seção acima, em que os dados corroboraram um padrão. Nesse contexto, não é conclusivo se os repositórios têm ou não características diferentes quanto aos três aspectos analisados

as quais expliquem divergências como a da temporalidade. Tal hipótese necessitaria de uma avaliação mais profunda, sendo cabível a uma possível continuação deste estudo.

5. Conclusão e Trabalhos Futuros

Neste trabalho, apresentamos uma caracterização de reclamações duplicadas por consumidores em três repositórios de dados distintos. Para isso, propusemos uma abordagem que leva em conta uma análise temporal das reclamações, similaridades semânticas entre elas, bem como os atributos: reclamante, reclamado e objeto da reclamação, explorando técnicas de PLN para a Língua Portuguesa.

Nossos resultados revelam uma série de descobertas interessantes que podem auxiliar essas plataformas na gestão adequada das reclamações. Mostramos que: 95% das duplicadas são postadas até 30 dias após a reclamação de origem, o que evidencia uma janela temporal em que majoritariamente se insere o universo das reclamações repetidas; o conteúdo das duplicatas sugere um padrão de comportamento característico dos reclamantes de cada domínio, os quais tendem a reclamar com recorrência acerca de assuntos-foco a depender da plataforma usada; a quantidade de duplicatas tende até duas cópias da reclamação original, o que pode levantar hipóteses aos gerenciadores dessas plataformas para os casos em que há mais do que isso; e as duplicatas postadas no mesmo dia correspondem a parte significativa do conjunto total das duplicatas na análise temporal, indicando que os reclamantes tem alguma tendência a repetir suas reclamações rapidamente.

Além disso, um padrão de postagem de reclamações observado foi o uso de *templates* de textos para criar uma reclamação, com a substituição apenas dos dados pessoais do reclamante como nome, CPF e endereço. Nestes casos o modelo entende que, semanticamente, os documentos tem alto *score* de semelhança e identifica como possível duplicata, sendo que de fato não é duplicata, pois trata-se de uma reclamação do mesmo tipo porém de consumidores distintos. Uma possível solução desse problema seria incluir essas informações sobre os usuários (entidades) por meio de um processo de *Named Entity Recognition* (NER), como o apresentado, por exemplo em [Belém et al. 2022, de Andrade et al. 2023b, Belém et al. 2023]. Complementarmente, podemos ponderar diferentemente partes (ou entidades mencionadas) distintas da reclamação para refletir a sua importância relativa na identificação de duplicatas [de Oliveira et al. 2007]. Estas são algumas das direções que pretendemos investigar em trabalhos futuros.

Ainda como trabalhos futuros, também planejamos explorar outros repositórios de dados de reclamações como o Reclame AQUI, bem como outros modelos de linguagem (*e.g.*, BERTabaporu [Costa et al. 2023]), incluindo os de larga escala (LLMs), como Sabiá, LLaMa (3.1) e GPT (4.0)⁸ e outras técnicas de representação de palavras (*e.g.*, word2vec [Sienčnik 2015], doc2vec [Le and Mikolov 2014]) e métricas de similaridade além da similaridade de cosseno, para comparação dos resultados e enriquecimento das análises. Pretendemos ainda realizar estudos de ablação neste contexto. Por fim, almejamos conduzir uma análise de eficiência da abordagem proposta com o objetivo de mensurar seu potencial de aplicação em um cenário real.

Agradecimentos. Esse trabalho foi parcialmente financiado pelo Ministério Público de Minas Gerais (MPMG), projeto Capacidades Analíticas, pelo CNPq, CAPES, FAPEMIG e AWS.

⁸maritaca.ai/sabia-2, llama.meta.com, chatbotapp.ai

Referências

- 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.
- Barz, B. and Denzler, J. (2020). Do We Train on Test Data? Purging CIFAR of Near-Duplicates. *Journal of Imaging*, 6(6):41.
- 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).
- Belém, F. M., Ganem, M. A. S., França, C., Carvalho, M., Laender, A. H. F., and Gonçalves, M. A. (2022). Reforço e delimitação contextual para reconhecimento de entidades e relações em documentos oficiais. In *Anais do Simp. Bras. de Banco de Dados (SBBDD)*, pages 292–303.
- 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, 2022*, pages 2206–2213.
- Costa, P. B., Pavan, M. C., Santos, W. R., Silva, S. C., and Paraboni, I. (2023). Bertabaporu: assessing a genre-specific language model for portuguese nlp. In *Proc. of the Int. Conf. on Recent Advances in Natural Language Processing (RANLP)*, pages 217–223.
- de Andrade, C. M. V., Belém, F., Cunha, W., França, C., Viegas, F., Rocha, L., and Gonçalves, M. A. (2023a). 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 Andrade, C. M. V., França, C., Belém, F., Jallais, G., Ganem, M. A. S., Texeira, G., Laender, A. H. F., and Gonçalves, M. A. (2023b). PromptNER: Uma Abordagem para Reconhecimento de Entidades Nomeadas em Dados Sensíveis a Partir de Instâncias Rotuladas Automaticamente. In *Anais do Simp. Bras. de Banco de Dados (SBBDD)*, pages 269–281.
- 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.
- de Oliveira, D. F., de Moura, E. S., Ribeiro-Neto, B. A., da Silva, A. S., and Gonçalves, M. A. (2007). Computing block importance for searching on web sites. In *Proc. ACM Conference on Information and Knowledge Management (CIKM)*, pages 165–174.
- Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16.

- 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.
- 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.
- Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.
- Loshin, D. (2010). *Master data management*. Morgan Kaufmann.
- Mangaravite, V., Carvalho, M., Cantelli, L., Ponce, L. M., Campoi, B., Nunes, G., Lander, A. H. F., and Goncalves, M. A. (2022). DedupeGov: Um Ambiente para Duplicação de Grandes Volumes de Dados de Pessoas Físicas e Jurídicas em Âmbito Governamental. In *Anais do Simp. Bras. de Banco de Dados (SBBDD)*, 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.
- 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.
- 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.
- Sargiani, V., de Castro, L. N., and Silva, L. A. (2020). A data mininf 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.
- Sienčnik, S. K. (2015). Adapting word2vec to named entity recognition. In *Proc. of the Nordic Conference of Computational Linguistics (NODALIDA)*, pages 239–243.
- 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 (SBBDD)*, 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.
- 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.