Predição de Desfechos no Consumidor.gov.br: Uma Abordagem Híbrida Tabular-Textual Baseada em Aprendizado de Máquina

Felipe Bandeira da Silva[1], **Dimas Cassimiro do Nascimento Filho**[2], **Bruno Nogueira**[3]

[11]Programa de Pós-Graduação em Engenharia da Computação (PPGEC)
Universidade de Pernambuco (UPE) – Recife – PE – Brasil

[2]Universidade Federal do Agreste de Pernambuco (UFAPE) – Garanhuns – PE – Brasil

[3]Universidade Federal de Alagoas (UFAL) – Maceió – AL – Brasil

fbs@ecomp.poli.br, dimas.cassimiro@ufape.edu.br, bruno@ic.ufal.br

Abstract. *Service failures generate operational costs, legal risks, and reputational damage. This study uses machine learning to predict the outcome of complaints on Consumidor.gov.br, assessing whether including complaint text in a hybrid tabular-textual approach improves performance over a tabular baseline. The method combines structured variables, Portuguese legal embeddings from Serafim-335m, and sentiment analysis, with the best performance achieved by XGBoost and Qwen2 used to explain the main themes. The results indicate an F1-Macro of 0.5928, with a 26.76% gain over the baseline, highlighting the model's usefulness for complaint prioritization and for the early identification of cases with greater risk of unsuccessful resolution.*

Resumo. *Falhas no atendimento geram custos operacionais, riscos legais e danos à reputação corporativa. Esta pesquisa utiliza aprendizado de máquina para prever o desfecho de reclamações no Consumidor.gov.br, investigando sua viabilidade e avaliando se a inclusão do texto da reclamação em uma abordagem híbrida tabular-textual melhora o desempenho em relação a um baseline tabular. O método combina variáveis estruturadas, embeddings jurídicos em português do Serafim-335m e análise de sentimento, utilizando o XGBoost para classificação e Qwen2 para explicar os principais temas. Os resultados indicam um F1-Macro de 0,5928, com ganho de 26,76% em relação ao baseline, evidenciando o potencial do modelo para priorização e antecipação de casos.*

1. Introdução

A gestão de conflitos de consumo no ambiente digital tornou-se um fator determinante para a sustentabilidade econômica e reputacional das organizações contemporâneas. Na última década, a digitalização massiva de serviços e o crescimento do comércio eletrônico transformaram as interações de mercado, frequentemente resultando em falhas de serviço cujo volume excede a infraestrutura de atendimento convencional [Vaishnav et al. 2024]. Nesse contexto, a insatisfação do consumidor emerge da lacuna entre a expectativa e o desempenho real do serviço, e a recuperação do serviço (*service recovery*) atua como o principal fator de retenção nos sistemas modernos de análise de queixas [Yang et al. 2018].

No Brasil, a plataforma Consumidor.gov.br institucionalizou a mediação desses conflitos no ambiente digital, sendo gerida pela Secretaria Nacional do Consumidor como ferramenta de *Online Dispute Resolution* que promove a desjudicialização e a transparência nas relações de consumo [Carneiro et al. 2024]. Embora o Consumidor.gov.br tenha registrado volumes massivos em anos anteriores, como as 818.157 queixas de 2023 [Senacon 2024], os dados analisados neste estudo, referentes a 2024 e 2025, evidenciam novos desafios de classificação que demandam abordagens híbridas para capturar a subjetividade dos relatos.

No entanto, apesar da elevada taxa de resolução geral, persiste um volume de casos não resolvidos, especialmente no setor de telecomunicações [Senacon 2024], o que indica quebra de confiança e risco de judicialização [Carneiro et al. 2024]. O desafio central para a automação reside na natureza predominantemente reativa das estratégias de triagem, que dependem de revisões manuais e análises retrospectivas de narrativas frequentemente ambíguas, incompletas e emocionalmente carregadas [Alarifi et al. 2023].

Identifica-se, portanto, uma lacuna metodológica para o desenvolvimento de modelos preditivos *ex-ante*, capazes de estimar o risco de não resolução no momento do registro da queixa [Baldo et al. 2022]. A premissa central desta pesquisa é que abordagens estritamente tabulares são insuficientes para capturar a complexidade do conflito, pois se limitam a metadados que ignoram a essência da reclamação [Alarifi et al. 2023]; o diferencial do modelo híbrido reside na capacidade de lidar com a carga semântica das narrativas textuais para distinguir um caso resolvível de um potencial processo judicial [Silva et al. 2024].

A contribuição principal deste trabalho é demonstrar empiricamente que a fusão de *embeddings* jurídicos em português (Serafim-335m), análise de sentimento (LeIA) e categorização semântica via LLM (Qwen2) com metadados *ex-ante* apresenta desempenho superior, no recorte avaliado, a linhas de base tabulares e textuais clássicas na predição de desfechos de reclamações no Consumidor.gov.br [Gomes et al. 2024, Almeida 2018, An et al. 2024]. Como contribuições secundárias, destacam-se: (i) a aplicação inédita dessa combinação ao domínio governamental brasileiro de defesa do consumidor; (ii) um protocolo de avaliação com divisão cronológica que evita vazamento temporal; e (iii) uma análise temática das macrocategorias de insatisfação como subsídio interpretativo à priorização.

2. Trabalhos relacionados

A literatura sobre predição de desfechos em reclamações de consumidores pode ser organizada em dois eixos: predição do resultado da mediação e priorização de demandas. No primeiro eixo, Jondhale et al. [Jondhale et al. 2024] predizem se o consumidor aceitará ou escalará a resposta da empresa em reclamações financeiras, enquanto Vaishnav et al. [Vaishnav et al. 2024] buscam prever tanto o tempo de resposta quanto o desfecho final, ambos no contexto norte-americano e sem semântica densa em português. No segundo eixo, Vairetti et al. [Vairetti et al. 2023] propõem priorização baseada em *deep learning* e critérios multicritério, enquanto Blümel e Zaki [Blümel and Zaki 2022] comparam abordagens clássicas e baseadas em *deep learning* para ordenar reclamações por urgência, sem explorar *embeddings* densos adaptados ao domínio jurídico-administrativo.

No domínio específico do Consumidor.gov.br, Sousa et al. [Sousa et al. 2020]

analisaram reclamações do setor de telecomunicações brasileiro com modelos puramente tabulares, sem reportar F1-Macro ou métricas por classe, o que limita a aplicabilidade em conjuntos desbalanceados e inviabiliza comparação direta com a abordagem proposta [Sujon et al. 2025].

A literatura recente sugere que integrar dados estruturados com representações textuais eleva o desempenho além dos limites das variáveis de cadastro [Alarifi et al. 2023]. Métodos baseados em frequência de palavras, como o TF-IDF, foram usados com relativo sucesso [Jondhale et al. 2024], mas falham em capturar nuances de gravidade e o jargão jurídico-administrativo brasileiro [Silva et al. 2024]. Este trabalho avança ao substituir representações esparsas por *embeddings* densos adaptados ao português, como a família Serafim-335m [Gomes et al. 2024], combinando essa profundidade semântica com a categorização assistida pelo Qwen2-7B [An et al. 2024]. A Tabela 1 sistematiza essa comparação com a literatura e a originalidade da abordagem proposta.

Tabela 1. Comparação com a literatura e contribuição deste trabalho

Estudo	Domínio	Foco	Dados	LLM / Embeddings	Diferencial / Lacuna
Jondhale et al. [Jondhale et al. 2024]	Financeiro	Desfecho	Híbrido	Não	Prediz escalada da reclamação, mas não no momento do registro e sem LLM.
Vaishnav et al. [Vaishnav et al. 2024]	Financeiro	Desfecho	Tabular	Não	Estima desfecho em contexto norte-americano, sem semântica densa em português.
Vairetti et al. [Vairetti et al. 2023]	Reclamações	Priorização	Texto	Sim	Prioriza reclamações urgentes sem integração com dados tabulares estruturados.
Blümel & Zaki [Blümel and Zaki 2022]	Reclamações	Priorização	Texto	Não	Compara NLP clássico e *deep learning* para priorizar, sem *embeddings* densos em português.
Sousa et al. [Sousa et al. 2020]	Telecom Brasil	Análise	Tabular	Não	Analisa reclamações do setor de telecom brasileiro sem modelagem preditiva de desfecho.
Este Trabalho	Consumidor .gov	Desfecho	Híbrido (Tab + Txt)	Sim (Serafim + Qwen2)	Une LLM, análise de sentimento e dados tabulares no contexto governamental brasileiro.

3. Protocolo Experimental

Esta seção detalha o protocolo experimental desenvolvido para a predição de desfechos no portal Consumidor.gov.br [Senacon 2024], ilustrado na Figura 1, que or-

ganiza os componentes da abordagem em etapas sequenciais para assegurar a reprodutibilidade e enfrentar os desafios técnicos de bases de dados governamentais ruidosas [Vaishnav et al. 2024, Baldo et al. 2022].

O processo inicia-se com a ingestão dos registros provenientes do portal Consumidor.gov.br [Senacon 2024], que são submetidos a um saneamento inicial para mitigar o ruído do jargão administrativo e a presença de duplicatas. Na sequência, o diagrama detalha a arquitetura híbrida adotada: os dados de perfil (variáveis estruturadas) seguem o caminho de processamento tabular estabelecido na literatura [Silva et al. 2024], enquanto o texto livre é direcionado para o enriquecimento semântico via *embeddings* densos [Gomes et al. 2024] e categorização assistida por LLM [An et al. 2024]. Esses vetores distintos são então concatenados para alimentar o modelo de classificação, cujo treinamento respeita um *split* cronológico fixo (treino e validação em 2024, teste em 2025) para assegurar a estabilidade das métricas e evitar o vazamento de dados (*data leakage*) [Vairetti et al. 2023]. As próximas subseções detalham os passos do fluxo proposto.

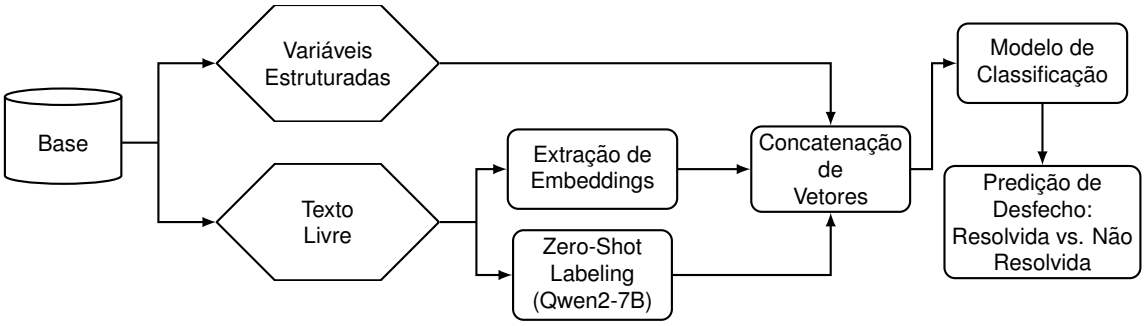

Figura 1. Protocolo experimental proposto

3.1. Ingestão e Saneamento Inicial de Dados

A extração inicial contemplou as reclamações de 2024 e 2025, resultando em um volume bruto de 89.321 registros. O saneamento seguiu etapas sequenciais, iniciando pela remoção de 2.233 duplicatas exatas (2,50%). Posteriormente, realizou-se a filtragem de desfechos para manter apenas os casos categorizados como Resolvido ou Não Resolvido, o que eliminou 7.893 registros (9,06%) e eliminou o ruído decorrente de situações de abandono ou processos inconclusivos [Jondhale et al. 2024].

Uma etapa metodológica essencial envolveu a exclusão de operadoras fora do grupo Tier-1, reduzindo a base em 21.806 registros (27,53%). Esta filtragem delimita o escopo do estudo às operadoras que detêm a maior fatia de mercado e concentram o maior volume de demandas no portal [Senacon 2024]. A escolha justifica-se pela necessidade de garantir a homogeneidade dos processos de atendimento e evitar o ruído estatístico de prestadoras de pequeno porte, cujos fluxos de mediação e prazos divergem significativamente dos grandes grupos econômicos [Senacon 2024].

A etapa seguinte consistiu na remoção de 1.048 relatos com menos de 10 palavras (1,83%) para assegurar a densidade semântica mínima necessária aos modelos de linguagem [Gomes et al. 2024]. Por fim, aplicou-se um algoritmo de janela deslizante de 30 dias utilizando similaridade de cosseno sobre os *embeddings* do Serafim-335m, resultando na exclusão de 978 duplicatas semânticas com similaridade maior ou igual a 0,95 (1,74%).

Ao final deste protocolo, a base foi consolidada com 55.363 reclamações, representando uma retenção de 61,98% do volume original.

3.2. Definição do Baseline Tabular

A definição do *baseline* tabular fundamenta-se na seleção de atributos estruturados que caracterizam a queixa no momento do registro. Este modelo utiliza metadados corporativos e geográficos, especificamente o nome da operadora, a unidade federativa e a cidade, permitindo capturar variações regionais e políticas de atendimento das empresas [Senacon 2024]. A escolha desses atributos é intencional: eles estão disponíveis no momento do registro da queixa, antes de qualquer interação de mediação, o que garante que o modelo opere estritamente com informações ex-ante e não incorra em vazamento de dados [Vaishnav et al. 2024]. A utilização desses atributos estabelece o patamar de desempenho das ferramentas tradicionais, viabilizando a comparação direta com a abordagem híbrida proposta [Alarifi et al. 2023]. Trabalhos como o de Vairetti et al. [Vairetti et al. 2023] reforçam que modelos baseados exclusivamente em metadados estruturados atingem um teto de desempenho que só pode ser superado com a incorporação de representações textuais, motivando a extensão híbrida proposta neste estudo.

3.3. Extração de Características e Representação Semântica

O processamento textual emprega o Serafim-335m, que gera vetores de 1024 dimensões por sentença [Gomes et al. 2024]. Adotamos os *embeddings* congelados (*feature extractor*), sem ajuste fino: essa escolha preserva a integridade das representações pré-treinadas, reduz o risco de sobreajuste em uma base ruidosa e garante eficiência de inferência [Vaishnav et al. 2024]. Pela mesma razão de preservar a integridade dessas representações, optou-se por manter as narrativas em formato bruto, com pontuação e conectivos originais, permitindo que o mecanismo de atenção explore plenamente o conhecimento pré-treinado e capture a estrutura sintática completa dos relatos [Silva et al. 2024].

3.4. Categorização e Modelagem de Tópicos via LLM

Para garantir a interpretabilidade da abordagem, aplicamos o modelo Qwen2-7B [An et al. 2024] para a rotulagem automática de amostras, convertendo os textos livres em variáveis categóricas que enriquecem os modelos tabulares com contexto inteligível [Vairetti et al. 2023]. Este componente é responsável por identificar sete macrocategorias de insatisfação que organizam o domínio das telecomunicações conforme os eixos temáticos mais recorrentes na plataforma [Senacon 2024]. Esta classificação permite que o modelo capture as nuances específicas de cada tipo de demanda antes da etapa final de inferência, auxiliando na construção de mecanismos de suporte à decisão mais precisos [Blümel and Zaki 2022].

As sete macrocategorias de insatisfação foram estabelecidas com base nas classificações adotadas pelo sistema de atendimento da Anatel, estruturadas em alinhamento às práticas de *CRM* do setor. Os tópicos recorrentes das reclamações foram agrupados, revisados manualmente e nomeados pelos autores; esses rótulos foram então fornecidos ao Qwen2 [An et al. 2024], que classificou cada relato em uma das categorias. A confiabilidade da rotulagem foi auditada em uma amostra aleatória de 200 reclamações por um especialista do setor, obtendo concordância de 91,5% e Kappa de Cohen de 0,83 (IC 95% [0,77; 0,89]). As sete macrocategorias estão detalhadas na Tabela 2.

Tabela 2. Macrocategorias de insatisfação (Qwen2-7B)

Categoria	Escopo e Temas Abrangidos	Exemplos de Demandas
Financeiro e Cobrança	Cobranças indevidas, faturas e sanções.	Conta, multa e negativação.
Planos e Ofertas	Gestão de contratos e promoções.	Ativação, migração e bônus.
Linha, Chip e Portabilidade	Conectividade física e troca de operadora.	Bloqueios e troca de SIM.
Qualidade e Falha Técnica	Performance de rede e infraestrutura.	Sinal de internet e suporte.
Atendimento	Relacionamento e jornada do cliente.	Qualidade e tempo de espera.
Cadastro e Aparelho	Dados do usuário e hardware.	Dados cadastrais e serviços.
Jurídico e Regulatório	Questões legais e normativas.	Ações judiciais e conformidade.

3.5. Protocolo de Avaliação e Prevenção de Data Leakage

A integração dos dados é realizada por meio da concatenação do vetor de metadados estruturados (baseline tabular) com o vetor de representação semântica do texto. Esse procedimento unifica as variáveis de contexto operacional com a carga subjetiva dos relatos dos consumidores [Gomes et al. 2024]. A escolha pela concatenação direta, em detrimento de estratégias de fusão mais complexas, justifica-se por sua transparência interpretativa e pela facilidade de isolar a contribuição individual de cada bloco de atributos na análise de desempenho [Vairetti et al. 2023]. Essa fusão permite que os classificadores processem simultaneamente informações geográficas, temporais e semânticas [Alarifi et al. 2023].

Para evitar vazamento de dados (*data leakage*), a avaliação respeitou a ordem cronológica dos registros. Os registros de 2024 foram destinados ao treinamento e ao ajuste de hiperparâmetros, enquanto os registros de 2025 foram reservados exclusivamente para o teste cego. Essa estratégia garante que o modelo seja avaliado apenas sobre eventos futuros em relação ao período de treinamento, aproximando o experimento das condições reais de uso no portal governamental [Vaishnav et al. 2024].

3.6. Avaliação e Métricas de Desempenho

Para lidar com o desequilíbrio significativo do banco de dados, adotou-se o F1-Macro como métrica principal. Recomendado para classes desproporcionais em contextos de negócios, ele equilibra precisão e recall, penalizando falhas em classes minoritárias e oferecendo uma medida mais adequada do poder preditivo em cenários desbalanceados [Sujon et al. 2025]. Essa escolha é corroborada por trabalhos recentes de predição de desfecho em reclamações, que também apontam o F1-Macro como métrica mais apropriada nesse tipo de base [Vaishnav et al. 2024].

A Equação 1 apresenta o cálculo do F1 para uma classe individual, definido como a média harmônica entre *Precision* e *Recall*. No caso deste estudo, o F1-Macro corresponde à média aritmética dos valores de F1 calculados separadamente para cada classe, sendo utilizado como métrica central na comparação entre os modelos avaliados.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{1}$$

4. Arquitetura e Implementação dos Modelos Preditivos

Esta seção detalha a configuração dos algoritmos de aprendizado de máquina e a estruturação do espaço de características para a predição binária do desfecho, comparando o desempenho entre o baseline puramente tabular e o modelo híbrido enriquecido semanticamente, com base no protocolo experimental e no pipeline de extração de características descritos na Seção 3. A metodologia adota uma abordagem de classificação supervisionada binária, em que o modelo busca antecipar o status da queixa. Optou-se pelo XGBoost por sua eficiência com dados heterogêneos em larga escala [Baldo et al. 2022], sua robustez ao desbalanceamento de classes [Sujon et al. 2025] e sua capacidade de capturar interações não lineares entre metadados operacionais e *embeddings* textuais densos [Jondhale et al. 2024].

Para garantir a imparcialidade na comparação entre o desempenho do baseline (apenas metadados estruturados) e da abordagem híbrida proposta (metadados integrados ao texto), a definição dos hiperparâmetros de cada arquitetura foi realizada por meio de otimização bayesiana de hiperparâmetros (*Optuna*) sobre o conjunto de treinamento [Akiba et al. 2019]. Este processo permitiu identificar as configurações ideais para os coeficientes de regularização e a profundidade das árvores, assegurando a robustez dos resultados de F1-Macro obtidos [Sujon et al. 2025]. O uso do status final como variável dependente justifica-se pela necessidade de identificar falhas de resolução que podem levar à judicialização, sendo o alvo ideal para a predição de risco [Senacon 2024].

Os atributos estruturados utilizados contemplam o Nome da Operadora, a Unidade Federativa e a Cidade, que servem para capturar o contexto operacional e geográfico mínimo da demanda [Sousa et al. 2020]. Embora representem um baseline conciso, essas variáveis permitem isolar e quantificar o ganho real de performance proporcionado exclusivamente pela inclusão das narrativas textuais, concentrando-se na correlação entre o perfil da operadora e a semântica do relato para antecipar o risco de não resolução no momento do registro da queixa [Alarifi et al. 2023, Rabbi et al. 2024].

A seguir, as Seções 4.1, 4.2 e 4.3 detalham as especificações das arquiteturas de classificação, enquanto a Seção 4.4 descreve as técnicas de processamento de linguagem natural que sustentam a representação híbrida.

4.1. Regressão Logística

Este modelo serve como um dos baselines comparativos fundamentais por ser robusto e interpretável [Sousa et al. 2020]. Estudos sobre mineração de dados em ouvidorias indicam que a Regressão Logística é uma ferramenta eficaz para estabelecer uma referência para o desfecho de reclamações [Alarifi et al. 2023]. Neste trabalho, utilizamos a técnica para avaliar como classificadores lineares se comportam diante dos espaços de alta dimensão gerados pelo Serafim-335m [Gomes et al. 2024]. Implementamos o modelo com regularização L2, máximo de 3.000 iterações, parâmetro $C = 0,5$, pesos de classes balanceados, resolvedor liblinear e estado aleatório fixo para fins de reprodutibilidade, visando equilibrar a importância dos atributos e evitar o sobreajuste em dados ruidosos [Alarifi et al. 2023].

4.2. Random Forest

Este modelo utiliza a técnica de Bagging para assegurar resistência a outliers e processar variáveis mistas sem normalizações complexas [Alarifi et al. 2023]. No con-

texto do Consumidor.gov.br, o modelo é valioso por sua resistência a outliers e capacidade de processamento simultâneo de variáveis mistas (estruturadas e não estruturadas), mantendo a integridade dos atributos tabulares [Sousa et al. 2020]. Para garantir o equilíbrio entre viés e variância, o modelo operou com uma floresta de 300 estimadores, sem limite de profundidade máxima, divisão mínima de 5 amostras, folhas mínimas de 2 amostras, peso de classes balanceado por subamostra, estado aleatório fixo para garantir a reprodutibilidade dos experimentos [Baldo et al. 2022].

4.3. XGBoost

O XGBoost combina árvores por reforço sequencial, corrigindo erros residuais a cada iteração, o que o torna eficaz em dados tabulares ruidosos [Baldo et al. 2022]. O modelo foi configurado com 500 estimadores, profundidade máxima de 6, taxa de aprendizado de 0,05, subamostragem de 0,8 (amostras e colunas), regularizações L1 de 0,1 e L2 de 1,0, peso mínimo por nó filho de 5 e peso da classe positiva ajustado pela razão entre classes. O limiar de decisão foi otimizado no conjunto de validação para maximizar o F1-Macro [Sujon et al. 2025]. Essa configuração equilibra complexidade e generalização sob o protocolo de divisão cronológica [Vairetti et al. 2023].

4.4. Processamento de Linguagem Natural

A inovação proposta consiste na transição de técnicas léxicas elementares para representações semânticas densas com o objetivo de atender à necessidade de adaptação de domínio no cenário administrativo e público brasileiro [Silva et al. 2024]. Adotamos embeddings densos utilizando o Serafim-335m, otimizado para gerar vetores que capturam a similaridade semântica global entre sentenças e permitem ao modelo identificar que expressões distintas pertencem ao mesmo contexto operacional [Gomes et al. 2024].

Para a análise de sentimento, utilizamos o léxico LeIA [Almeida 2018], amplamente validado para a língua portuguesa e baseado em regras gramaticais sólidas, cujo uso em detrimento de modelos puramente estatísticos é corroborado por Santos et al. [Santos et al. 2022]. Essa abordagem captura nuances afetivas do estado emocional do reclamante, oferecendo maior interpretabilidade e complementando as representações densas no cenário administrativo nacional [Blümel and Zaki 2022].

5. Experimentos e Análise dos Resultados

Esta seção avalia o desempenho da abordagem proposta. Os classificadores foram avaliados sob o protocolo de divisão cronológica (Seção 3.5), sobre o conjunto de teste de 2025, evitando o vazamento de informações temporais [Vaishnav et al. 2024]. Os resultados consolidados na Tabela 3 mostram que a inclusão de *embeddings* e análise de sentimento eleva o F1-Macro em todas as arquiteturas [Jondhale et al. 2024].

A coluna de ganho relativo quantifica, para cada algoritmo, a diferença de F1-Macro entre a variante tabular pura e a híbrida, isolando o efeito da incorporação de informação textual ao processo de classificação.

Tabela 3. Comparação de desempenho e ganho relativo

Algoritmo	Abordagem	Acurácia	F1-Macro	Ganho Relativo (%)
Regressão Logística	Tabular	0,763	0,563	-
	Híbrida	0,801	0,592	+5,27%
Random Forest	Tabular	0,828	0,508	-
	Híbrida	0,785	0,586	+15,34%
XGBoost	Tabular	0,838	0,468	-
	Híbrida	**0,775**	**0,593**	**+26,76%**

A abordagem híbrida com XGBoost alcançou o maior desempenho, com F1-Macro de 0,59 no conjunto de teste. Esse valor representa ganho de 26,76% sobre o XGBoost puramente tabular e de 5,4% sobre a melhor linha de base tabular do estudo (Regressão Logística, 0,56), refletindo maior capacidade de identificar a classe minoritária ("Não Resolvido") [Sujon et al. 2025].

O modelo híbrido supera as limitações estruturais ao integrar embeddings semânticos, capturando a subjetividade do consumidor para além dos limites das variáveis de cadastro [Vaishnav et al. 2024]. A riqueza do relato textual permite classificar corretamente desfechos de insucesso que modelos puramente estruturais negligenciam, conforme comprovado pelo incremento de 26,76% no F1-Macro [Jondhale et al. 2024].

5.1. Discussão e Impacto do Ganho Semântico

Para detalhar o comportamento do modelo híbrido completo além da métrica agregada, a Tabela 4 apresenta a precisão, a revocação e o F1 por classe, indicadores recomendados para cenários desbalanceados [Sujon et al. 2025].

Tabela 4. Desempenho por classe do modelo híbrido (teste, 2025)

Classe	Precisão	Revocação	F1
"Resolvido"	0,870	0,849	0,860
"Não Resolvido"	0,303	0,341	0,321
F1-Macro	–	–	**0,593**

Como esperado em uma base desbalanceada, a classe majoritária "Resolvido" apresenta F1 elevado (0,86), ao passo que a minoritária "Não Resolvido" permanece o principal desafio, com F1 de 0,32 e AUC-PR de 0,29, aproximadamente 1,8 vez acima da linha de base aleatória [Sujon et al. 2025]. A matriz de confusão correspondente registra 20.837 verdadeiros negativos, 3.699 falsos positivos, 3.104 falsos negativos e 1.606 verdadeiros positivos; dos 4.710 casos não resolvidos no conjunto de teste, o modelo recupera 1.606. Esse compromisso é desejável em um cenário de priorização, no qual deixar de sinalizar um caso de risco é mais oneroso do que revisar um caso adicional.

A análise dos resultados revela que a introdução de representações semânticas densas permitiu uma reorganização do espaço de características, sendo o XGBoost o algoritmo com maior ganho relativo de F1-Macro (+0,1252) ao adotar a abordagem híbrida. Esse resultado sugere que a combinação dos embeddings densos do Serafim-335m [Gomes et al. 2024], dos indicadores de sentimento [Almeida 2018] e do contexto temático [An et al. 2024] gera um espaço de alta dimensão no qual o boosting sequencial captura padrões não lineares entre o sentimento do relato e os metadados operacionais, nuances que seriam ignoradas por modelos lineares ou por metadados puros [Baldo et al. 2022, Jondhale et al. 2024]. Esse comportamento reforça que, para a natureza ruidosa dos dados governamentais, a capacidade de aprendizado iterativo do modelo é mais determinante para o sucesso do que a simplicidade estatística [Baldo et al. 2022].

Em termos práticos, a transição de uma triagem reativa para uma priorização baseada em risco pode reduzir custos, dado que a literatura aponta a mediação extrajudicial como a via mais econômica no cenário brasileiro, frente ao custo do litígio [Carneiro et al. 2024, Vairetti et al. 2023].

5.2. Análise Temática e Preditores de Insucesso

A análise qualitativa das macrocategorias geradas pelo Qwen2-7B [An et al. 2024] e sua influência no modelo XGBoost revela uma dinâmica distinta entre o peso estatístico e a severidade observada de cada tipo de demanda. A rotulagem das sete macrocategorias foi conduzida sobre uma amostra de 11.073 reclamações extraída do conjunto de teste de 2025, suficiente para caracterizar a distribuição temática e os preditores de insucesso sem exigir a inferência do LLM sobre a totalidade da base. Nessa amostra, o bloco categórico foi dominado pelos eixos Atendimento e Financeiro e Cobrança, que juntos concentraram mais de 93% da importância relativa da variável categoria.

Embora o grupo Atendimento concentre a maior importância relativa, seu impacto predominante está associado à classe “Resolvido”, com taxa de não resolução inferior ao *baseline* [Senacon 2024]. Em contraste, Financeiro e Cobrança é o principal eixo com contribuição líquida positiva para “Não Resolvido”, sugerindo que disputas de cobrança, contestações de valores e sanções contratuais adicionam sinal preditivo de maior complexidade [Sousa et al. 2020].

Tabela 5. Distribuição e importância das macrocategorias (amostra rotulada, 2025)

Macrocategoria	N	Taxa Não Resolvido	Lift vs Baseline	Peso Bloco Categoria	Impacto no Modelo
Atendimento	571	12,74%	0,79	48,61%	Resolvido
Financeiro e Cobrança	7.731	14,59%	0,91	45,02%	Não Resolvido
Planos e Ofertas	1.111	17,21%	1,07	6,37%	Resolvido
Linha, Chip e Portabilidade	1.074	23,33%	1,45	0,00%	Variável
Qualidade e Falha Técnica	459	23,10%	1,43	0,00%	Variável
Cadastro e Aparelho	104	29,20%	1,81	0,00%	Variável
Jurídico e Regulatório	23	18,33%	1,14	0,00%	Variável

Os grupos Linha, Chip e Portabilidade e Qualidade e Falha Técnica, apesar de taxas de não resolução superiores ao *baseline*, exibiram contribuição desprezível no bloco categórico (valores arredondados para zero). Esse padrão indica que o XGBoost não dependeu da etiqueta categórica isolada para representar esse risco, capturando-o prioritariamente por meio dos *embeddings* textuais [Gomes et al. 2024] e dos metadados estruturados. Nesse sentido, embora a categorização via Qwen2 [An et al. 2024] não produza ganho mensurável em F1-Macro (decorrendo o ganho de 26,76% da integração sinérgica dos demais atributos do modelo [Sujon et al. 2025]), sua contribuição é interpretativa: as macrocategorias oferecem uma leitura temática dos preditores de insucesso, útil como ferramenta de apoio à decisão e que ajuda a explicar por que a abordagem híbrida superou a versão puramente tabular.

5.3. Análise de Ablação

Para quantificar a contribuição isolada de cada componente, conduzimos um estudo de ablação incremental adicionando os blocos de atributos sucessivamente ao XGBoost sob o mesmo protocolo cronológico (Tabela 6).

Tabela 6. Contribuição incremental dos componentes (F1-Macro, 2025)

Configuração	F1-Macro	Δ
Tabular (linha de base)	0,468	–
+ Serafim	0,585	+0,117
+ Serafim + LeIA	0,592	+0,007
+ Serafim + LeIA + Qwen2 (completo)	**0,593**	+0,001

Os resultados evidenciam que os *embeddings* densos do Serafim-335m [Gomes et al. 2024] concentram praticamente todo o ganho sobre a linha de base tabular (incremento de $\sim$0,11 no F1-Macro, $\sim$95% da melhoria total), enquanto a análise de sentimento via LeIA [Almeida 2018] contribui marginalmente e a categorização via Qwen2 [An et al. 2024] não altera o desempenho preditivo, atuando sobretudo como recurso interpretativo (Seção 5.2).

Tabela 7. Representação densa vs. clássica (XGBoost, 2025)

Representação	Acurácia	F1-Macro
Baseline tabular	0,838	0,468
TF-IDF (texto puro)	0,747	0,577
TF-IDF + metadados	0,812	0,594
Híbrido com Serafim	0,775	0,593

Avaliamos a significância das diferenças com o teste de McNemar e um procedimento de *bootstrap* (5.000 reamostragens) sobre a diferença de F1-Macro, em consonância com as recomendações de avaliação para cenários desbalanceados

[Sujon et al. 2025]. Tanto o Serafim quanto o TF-IDF apresentam desempenho estatisticamente superior ao do *baseline* tabular (diferença de F1-Macro próxima de 0,11; p ¡ 0,001), o que confirma que a incorporação do texto é o fator determinante do desempenho.

Embora o TF-IDF combinado a metadados alcance F1-Macro agregado equivalente (diferença não significativa, p = 0,28), na classe minoritária "Não Resolvido", alvo prioritário deste trabalho, a representação densa do Serafim-335m é significativamente superior à clássica (diferença de F1 de +0,023; IC 95% [+0,010; +0,036]; p ¡ 0,001), o que justifica sua adoção como representação textual da abordagem proposta.

6. Conclusão e trabalhos futuros

Este trabalho apresentou um protocolo experimental para a predição de desfechos no portal Consumidor.gov.br, comparando modelos puramente tabulares com uma abordagem híbrida que integra *embeddings* semânticos. Os resultados demonstram que a acurácia isolada é uma métrica insuficiente para conjuntos desbalanceados, pois mascara a incapacidade de identificar casos de insucesso, fenômeno conhecido como o paradoxo da acurácia [Sujon et al. 2025].

O XGBoost híbrido sugere capacidade de distinguir nuances presentes em reclamações complexas, como falhas técnicas e cobranças indevidas, que modelos puramente tabulares tendem a negligenciar [Jondhale et al. 2024]. No recorte avaliado, a abordagem pode apoiar a identificação precoce de casos com maior risco de não resolução [Vairetti et al. 2023], e a leitura das macrocategorias de insatisfação pode subsidiar a priorização de intervenções [Senacon 2024].

Como trabalhos futuros, pretende-se explorar a adaptação de modelos de linguagem de maior escala para o domínio jurídico-administrativo brasileiro, visando reduzir ainda mais a esparsidade dos dados textuais. Além disso, planeja-se o desenvolvimento de um sistema de fila de priorização de reclamações, no qual o modelo preditivo ordena automaticamente as demandas conforme o risco estimado de não resolução, permitindo que as empresas concentrem seus esforços de mediação nos casos com maior probabilidade de insucesso e, consequentemente, aumentem as taxas de resolução antes da judicialização [Senacon 2024].

Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2623–2631.

Alarifi, G., Rahman, M. F., and Hossain, M. S. (2023). Prediction and analysis of customer complaints using machine learning techniques. *International Journal of E-Business Research*, 19(1):1.

Almeida, R. J. A. (2018). LeIA – léxico para inferência adaptada. `https://github.com/rafjaa/LeIA`. Repositório GitHub.

An, Y., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., et al. (2024). Qwen2 technical report. arXiv preprint arXiv:2407.10671.

Baldo, F., Grando, J., Weege, K. M., and Bonassa, G. M. (2022). Adaptive fast XGBoost for binary classification. In *Anais do XXXVII Simpósio Brasileiro de Banco de Dados (SBBD)*.

Blümel, J. and Zaki, M. (2022). Comparative analysis of classical and deep learning-based natural language processing for prioritizing customer complaints. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, pages 1–10.

Carneiro, C. D., Duzert, Y., and Almeida, R. A. d. (2024). The economic benefits of business mediation in the Brazilian scenario. *Revista de Administração de Empresas*, 64(3):e2023–0145.

Gomes, L., Branco, A., Silva, J., Rodrigues, J., and Santos, R. (2024). Open sentence embeddings for Portuguese with the Serafim PT* encoders family. In *Progress in Artificial Intelligence – 23rd EPIA Conference on Artificial Intelligence (EPIA 2024)*, pages 267–279, Cham. Springer Nature Switzerland.

Jondhale, R., Patil, S., Shinde, A., Ajalkar, D., and Biradar, S. (2024). Predicting consumer complaint disputes in finance using machine learning. In *2024 Second International Conference on Advances in Information Technology (ICAIT)*, pages 385–391.

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

Santos, B. L., Ferreira, G. E., Ó, M. T. d., Braz, R. R., and Digiampietri, L. A. (2022). Comparison of natural language processing techniques in social bot detection on Twitter during Brazilian presidential elections. In *Proceedings of iSys – Revista Brasileira de Sistemas de Informação*.

Senacon (2024). Boletim consumidor.gov.br 2023. Boletim, Secretaria Nacional do Consumidor, Ministério da Justiça e Segurança Pública.

Silva, M. O., Oliveira, G. P., Costa, L. G. L., and Pappa, G. L. (2024). Evaluating domain-adapted language models for governmental text classification tasks in Portuguese. In *Anais do XXXIX Simpósio Brasileiro de Banco de Dados (SBBD)*, page 247.

Sousa, G. N. d., Guimarães, I. d. S., Viana, J., Reinhold, O., Fernando, A., and Lobato, F. M. F. (2020). Análise do setor de telecomunicação brasileiro: Uma visão sobre reclamações. *RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação*, 37:31–48.

Sujon, K. M., Hassan, R., Choi, K., and Samad, M. A. (2025). Accuracy, precision, recall, F1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, 12(1).

Vairetti, C., Aránguiz, I., Maldonado, S., Karmy, J. P., and Leal, A. (2023). Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making. *European Journal of Operational Research*, 312(3):1108–1122.

Vaishnav, D., Woo, J., et al. (2024). Predictive analysis of CFPB consumer complaints using machine learning. arXiv preprint arXiv:2407.06399.

Yang, Y., Xu, D., Yang, J., and Chen, Y. (2018). An evidential reasoning-based decision support system for handling customer complaints in mobile telecommunications. *Knowledge-Based Systems*, 162:202–214.