

Caracterização em larga escala de golpes e fraudes financeiras a partir de reclamações bancárias do Reclame Aqui

Thalia B. M. de Almeida
thaliabarbosa@dcc.ufmg.br
Universidade Federal de Minas
Gerais
Belo Horizonte, Brasil

Karen Gomes
karengomes@dcc.ufmg.br
Universidade Federal de Minas
Gerais
Belo Horizonte, Brasil

Fabício Benevenuto
fabricio@dcc.ufmg.br
Universidade Federal de Minas
Gerais
Belo Horizonte, Brasil

Marcos André Gonçalves
mgoncalv@dcc.ufmg.br
Universidade Federal de Minas
Gerais
Belo Horizonte, Brasil

Antonio A. F. Loureiro
loureiro@dcc.ufmg.br
Universidade Federal de Minas
Gerais
Belo Horizonte, Brasil

ABSTRACT

Financial fraud has become a critical challenge in Brazil's digital banking ecosystem, yet large-scale evidence on how victims describe scams and unauthorized fraud in Portuguese remains scarce. This paper presents the first large-scale characterization of banking fraud narratives from Reclame Aqui, investigating recurring fraud mechanisms, social engineering strategies, and institutional responses. To enable this analysis, we develop a hybrid pipeline combining weak supervision, LLM-based annotation, and supervised fine-tuning of BERTimbau. Applied to a corpus of 221,519 complaints, the pipeline achieves a Macro F1 of 0.849 and identifies 20,127 candidate reports of scams and unauthorized fraud. BERTopic reveals 19 recurring fraud mechanisms, split between social engineering scams and unauthorized fraud. Pix-based social engineering emerges as the dominant pattern (33.9%), bringing together diverse impersonation schemes that converge on immediate, irreversible transfers. Further analyses show that trust and greed are the predominant psychological levers, primarily associated with phone calls and social media, respectively, while scams exhibit substantially lower resolution rates than unauthorized fraud (16.7% vs. 29.3%). Overall, our findings demonstrate that consumer complaint platforms constitute a valuable and underexplored resource for large-scale financial fraud characterization, providing empirical evidence on fraud mechanisms, social engineering strategies, and institutional responses that can support fraud prevention, regulatory analysis, and future research on financial crime in Brazilian Portuguese.

KEYWORDS

golpes, fraude, processamento de linguagem natural, BERTimbau, anotação por LLM, Reclame Aqui, modelagem de tópicos

1 INTRODUÇÃO

O sistema financeiro brasileiro passou a concentrar, nos últimos anos, uma parcela significativa das fraudes digitais [29]. Em muitos casos, o objetivo dos criminosos não é comprometer sistemas informatizados, mas induzir o próprio usuário a realizar operações financeiras. Para isso, utilizam estratégias como ligações em que se passam por funcionários de instituições financeiras ou lojas virtuais falsas destinadas à obtenção de pagamentos, principalmente via Pix [16]. Ao mesmo tempo, persistem fraudes praticadas sem a participação da vítima, como o uso de cartões clonados e abertura de contas. Em ambas as situações, os prejuízos costumam ser identificados apenas após o registro da movimentação financeira, levando o consumidor a procurar a instituição bancária para obter esclarecimentos.

A distinção entre golpe e fraude não autorizada, entretanto, não é estabelecida de forma uniforme. No Direito brasileiro, a noção de fraude aparece em diferentes dispositivos e contextos, onde o termo abarca condutas que incluem tanto a indução ao erro quanto o uso indevido de dados alheios, sem corresponder diretamente à separação usada em tarefas computacionais [6]. Na literatura internacional, golpes (*scams*) são usualmente tratados como uma modalidade dentro do campo mais amplo das fraudes [7]. Por isso, este artigo não propõe uma classificação jurídica. A distinção adotada é operacional, chamamos de golpe os relatos em que há indução, persuasão, personificação ou abuso de confiança para que a vítima participe da operação lesiva; chamamos de fraude não autorizada os relatos em que a operação é atribuída ao uso indevido de dados, cartões, contas ou documentos sem participação consciente da vítima. Essa separação orienta a anotação, a classificação e a análise dos mecanismos descritos nas reclamações.

Esse cenário ganha ainda mais relevância quando observado em termos de escala. O *Relatório de Identidade e Fraude 2025*, da Serasa Experian, apontou que 51% dos brasileiros declararam ter sido vítimas de algum tipo de fraude naquele ano, sendo que 54,2% dessas ocorrências resultaram em perdas financeiras efetivas [28]. A própria Febraban estima que aproximadamente 70% dos golpes bancários dependem de

engenharia social [13], evidenciando que a exploração da vítima, e não necessariamente uma vulnerabilidade técnica dos sistemas, constitui o principal vetor de ataque.

Há também uma motivação jurídico-econômica para esse mapeamento. A Súmula 479 do Superior Tribunal de Justiça (STJ) estabelece a responsabilidade objetiva dos bancos por danos decorrentes de fraudes vinculadas aos riscos inerentes à atividade bancária [31]. Em decisão recente, o STJ reconheceu a responsabilidade de instituições financeiras e de pagamento em casos do golpe da falsa central de atendimento quando falhas nos mecanismos de segurança contribuíram para a concretização da fraude [32]. Assim, compreender como golpes e fraudes se manifestam nas reclamações dos consumidores é relevante não apenas para prevenção, mas também para reduzir exposição a perdas, contestações e indenizações, além de avaliar se as respostas institucionais são compatíveis com os riscos associados à atividade bancária.

O Reclame Aqui é uma plataforma brasileira de comunicação entre consumidores e empresas, utilizada tanto para registrar reclamações quanto para consultar a reputação de fornecedores [24]. Nessas reclamações, vítimas descrevem em linguagem natural como os golpes ocorreram, quais estratégias foram utilizadas pelos criminosos, quais perdas sofreram e como as instituições financeiras responderam aos casos. Mais do que registrar conflitos entre consumidores e empresas, esse conjunto de relatos constitui um registro espontâneo dos mecanismos de fraude observados na prática, oferecendo uma oportunidade ainda pouco explorada para compreender esses fenômenos em larga escala. Um exemplo é o volume de casos registrado do golpe da falsa venda, que figurou entre as modalidades mais comunicadas às instituições associadas à Febraban no primeiro semestre de 2025. Foram registradas 174 mil ocorrências, número que representa crescimento de 314% em relação ao mesmo período de 2024 [14].

Entretanto, transformar esse grande volume de reclamações em evidências sobre o fenômeno exige primeiro identificar automaticamente quais relatos efetivamente descrevem golpes e fraudes. Até o nosso conhecimento, não existem conjuntos de dados rotulados em português que realizem essa distinção de forma sistemática. A anotação manual em larga escala é custosa, e abordagens baseadas apenas em palavras-chave produzem erros tanto por excesso quanto por omissão, já que relatos de fraude podem não empregar os termos mais óbvios. Além disso, a fronteira entre golpe praticado por terceiros e conduta atribuída à própria instituição nem sempre é nítida no texto, e a distinção entre golpe e fraude não autorizada tampouco aparece explicitamente na narrativa da vítima. Diante desse cenário, investigamos duas questões de pesquisa:

QP1 Como se caracterizam os relatos de golpes e fraudes não autorizadas presentes em reclamações bancárias brasileiras, em termos de tipologia, mecanismos de engenharia social e resposta institucional?

QP2 Como identificar automaticamente esses relatos em um corpus ruidoso de reclamações bancárias em português?

Para responder essas questões, este trabalho apresenta três contribuições. Primeiro, propomos um pipeline híbrido que combina supervisão fraca, anotação por LLM e ajuste fino de BERTimbau para identificar relatos de golpes e fraudes em um grande corpus de reclamações bancárias. Em seguida, realizamos, até onde é de nosso conhecimento, a primeira caracterização em larga escala desses relatos em português, revelando uma tipologia de 19 mecanismos recorrentes e a centralidade do Pix como infraestrutura comum a diferentes golpes de engenharia social. Por fim, caracterizamos padrões de canais de abordagem, gatilhos psicológicos e diferenças nas taxas de resolução entre golpes e fraudes não autorizadas, mostrando que plataformas de reclamação constituem uma fonte valiosa e pouco explorada para compreender a dinâmica dos golpes financeiros, produzindo evidências que podem subsidiar pesquisas futuras, estratégias de prevenção e políticas de enfrentamento à fraude.

2 TRABALHOS RELACIONADOS

Os golpes financeiros envolvem linguagem em diferentes etapas do processo criminoso. Enquanto os golpistas utilizam mensagens para persuadir vítimas e induzi-las a realizar operações financeiras, as vítimas descrevem posteriormente essas experiências em reclamações, boletins de ocorrência e outras narrativas textuais. A literatura de Processamento de Linguagem Natural (PLN) aplicada a fraudes concentra-se predominantemente na primeira perspectiva. Salloum et al. [26], ao revisar 112 artigos sobre detecção de *phishing*, mostram que modelos baseados em *transformers* tendem a superar abordagens tradicionais quando o sinal principal está no corpo textual da mensagem. Resultado semelhante é observado por Boulieris et al. [3] no domínio mais amplo de fraude financeira, em que representações contextuais permitem distinguir textos fraudulentos e legítimos com maior precisão do que métodos puramente léxicos. Essa linha, porém, costuma tratar cada modalidade de ataque como um problema específico. Em uma revisão sistemática de 2.457 estudos, Papasavva et al. [21] indicam que a área permanece fragmentada por tipo de golpe, com pouca transferência entre cenários. Uma alternativa a essa fragmentação aparece em Tun and Birks [33], que vão além da classificação textual e utilizam PLN para reconstruir *crime scripts*, extraíndo sequências de ações, papéis e etapas envolvidas em golpes a partir de relatos escritos.

A segunda perspectiva, centrada nas narrativas produzidas pelas próprias vítimas, aparece com mais força em estudos baseados em plataformas de reclamação de consumidores. Bastani et al. [2] analisam reclamações do *Consumer Financial Protection Bureau* (CFPB) com LDA para identificar tópicos latentes e acompanhar sua evolução temporal como apoio à supervisão regulatória. No mesmo domínio, Sangaraju et al. [27] substituem LDA por BERTopic e mostram que *embeddings* contextuais produzem tópicos mais coerentes e semanticamente informativos [34–36]. Já Prenassi et al. [22] deslocam esse tipo de análise para um contexto brasileiro ao estudar transcrições de vídeos em português sobre o Fortune Tiger no YouTube, caracterizando padrões discursivos associados a conteúdo financeiramente nocivo. Esses trabalhos

evidenciam o potencial de fontes textuais produzidas fora de bases bancárias tradicionais, mas ainda exploram de forma limitada as reclamações bancárias brasileiras como relatos diretos de vitimização e fonte para compreender mecanismos de golpes financeiros.

Outro desafio recorrente nesse tipo de pesquisa é a construção de conjuntos de dados rotulados em domínios especializados. Ratner et al. [23] formalizam a supervisão fraca como uma forma de combinar funções heurísticas ruidosas para estimar rótulos latentes, reduzindo a dependência de anotação manual. Mais recentemente, LLMs passaram a ser utilizados como anotadores em tarefas de ciências sociais computacionais. Ziems et al. [38] mostram que esse uso pode ser promissor, mas depende de validação contra referências humanas, pois o desempenho varia conforme a tarefa, o domínio e o desenho da anotação.

Diante desse panorama, observa-se que diferentes vertentes da literatura abordam aspectos complementares do problema, como detecção automática de fraudes, modelagem de tópicos em reclamações de consumidores e construção de conjuntos de dados rotulados. Entretanto, esses estudos concentram-se predominantemente na identificação de modalidades específicas de fraude ou na avaliação de modelos de classificação. Ainda permanece pouco explorada a caracterização, em larga escala, de como vítimas descrevem golpes e fraudes em linguagem natural, quais mecanismos recorrentes emergem dessas narrativas e como esses eventos se relacionam às respostas das instituições financeiras. Este trabalho busca preencher essa lacuna ao combinar a identificação automática de relatos de golpes e fraudes com sua caracterização tipológica, a análise de estratégias de engenharia social e a investigação das respostas institucionais em um grande corpus de reclamações bancárias em português.

3 METODOLOGIA

O corpus deste trabalho foi construído a partir de reclamações coletadas por raspagem na plataforma Reclame Aqui, concentrando-se nos perfis de nove dos principais bancos e instituições de pagamento atuantes no Brasil [1, 11]. Além do texto da reclamação, foram coletados diferentes componentes informacionais da página [15], como instituição, categoria, data, resposta da empresa e indicador de resolução declarado pelo consumidor.

No pré-processamento, foram removidas marcações HTML e normalizado o Unicode, compondo um campo textual com título e corpo da reclamação. Dois padrões foram preservados intencionalmente. O primeiro é *********, usado pela plataforma para mascarar informações pessoais. O segundo é **[Editado pelo Reclame Aqui]**, inserido quando partes do conteúdo são ocultadas por envolverem termos sensíveis como “golpe”, “fraude” ou dados de terceiros. Extraímos também atributos sobre ocorrência, quantidade e posição desses marcadores, incluindo a presença de edição simultânea na reclamação e na resposta institucional.

A rotulagem seguiu um pipeline híbrido em quatro etapas. Na primeira, aplicamos supervisão fraca inspirada em [23],

com funções heurísticas que combinam densidade do marcador de edição, expressões explícitas de golpe e fraude, padrões de transação não reconhecida, indícios de invasão de conta e categorias de problema da plataforma. Cada reclamação recebeu um rótulo fraco em **{suspeito, sem_indicio, abstencao}**, obtido por maioria simples entre os votos; registros com pelo menos dois votos negativos e nenhum positivo foram descartados da fila de anotação.

Na segunda etapa, construímos um conjunto de referência por anotação manual de 300 reclamações selecionadas de forma estratificada a partir dos rótulos fracos, em linha com trabalhos que exploram seleção eficiente de instâncias [9, 10], com amostragem aleatória usando **seed=42** dentro de cada estrato. A anotação foi conduzida por um dos autores, seguindo diretrizes previamente definidas. As diretrizes classificaram como **suspeito** as reclamações cujo assunto principal era um golpe ou fraude com mecanismo descrito ou fortemente indicado, e como **sem_indicio** os demais casos, incluindo cobranças, falhas operacionais e mau atendimento. A anotação foi forçada para a decisão binária, sem opção de abstenção.

Na terceira etapa, os 300 casos anotados serviram para selecionar um anotador LLM entre três modelos avaliados em esquema *few-shot*, adotando como critério o maior Cohen’s κ contra o anotador humano [38]. O modelo selecionado ampliou o conjunto rotulado.

Na quarta etapa, treinamos e comparamos classificadores binários sobre esse conjunto, avaliando três famílias de modelos em validação cruzada estratificada, com e sem concatenação da resposta da empresa ao texto da reclamação. O classificador escolhido foi retreinado no conjunto completo e aplicado ao corpus inteiro em uma análise retrospectiva, na qual tanto a reclamação quanto a resposta da empresa já estavam disponíveis. Essa configuração não representa triagem em tempo real: para uso antes da resposta institucional, deve-se usar a versão treinada apenas com o texto da reclamação. O limiar operacional foi definido pela curva Precision–Recall para equilibrar precisão e recall da classe positiva.

Sobre as reclamações acima do limiar, aplicamos modelagem de tópicos com BERTopic [17] para descobrir agrupamentos recorrentes de mecanismos relatados. Comparamos três modelos de embeddings e selecionamos o de maior coerência. A clusterização utilizou UMAP seguido de HDBSCAN, com *outliers* reatribuídos ao tópico mais próximo. O número de tópicos foi fixado após comparação de métricas de coerência e inspeção qualitativa. Por fim, cada tópico foi rotulado com apoio de LLM e inspeção de palavras-chave e exemplos representativos, produzindo uma tipologia que separa **golpe**, **fraude** e demais casos que agrupa reclamações sobre falhas operacionais ou práticas abusivas da própria instituição no qual classificador não distinguiu com confiança suficiente do escopo de interesse e foram excluídas das análises tipológicas principais.

4 PROJETO EXPERIMENTAL

Esta seção apresenta as escolhas experimentais associadas à seleção do anotador LLM, do classificador final, do limiar de inferência e da modelagem de tópicos.

4.1 Seleção do Anotador LLM

A Tabela 1 compara os três LLMs candidatos em relação ao conjunto de referência humano. Todos obtiveram concordância quase perfeita ($\kappa > 0,80$); o **llama-3.3-70b** alcançou o maior valor ($\kappa = 0,8223$) e foi adotado para ampliar o conjunto rotulado de 300 para 2.000 exemplos.

Tabela 1: Concordância entre os rótulos manuais e as predições dos LLMs na amostra de validação ($n = 300$).

Modelo	Cohen's κ vs. humano
llama-3.3-70b	0,8223
nemotron-super-49b	0,8108
llama-4-maverick	0,8036
Fleiss' κ (4 anotadores)	0,8482

4.2 Seleção do Classificador

Com 2.000 exemplos rotulados, dos quais 290 pertencem à classe **suspeito** (14,5%), comparamos três famílias de classificadores em validação cruzada estratificada (Tabela 2). O BERTimbau ajustado com reclamação e resposta institucional obteve o maior Macro F1 (0,849) e foi adotado como modelo final. Como essa configuração usa a resposta da empresa, seus resultados devem ser interpretados como análise retrospectiva; para triagem no momento da publicação da reclamação, a configuração sem resposta é a referência operacional.

Tabela 2: Macro F1 médio dos modelos de classificação

Modelo	Sem resposta	Com resposta
BERT Embeddings		
MLP	0,774 $\pm$ 0,024	0,760 $\pm$ 0,046
Logistic Regression	0,744 $\pm$ 0,018	0,740 $\pm$ 0,053
SVM (RBF)	0,747 $\pm$ 0,046	0,724 $\pm$ 0,034
XGBoost	0,703 $\pm$ 0,035	0,718 $\pm$ 0,037
LightGBM	0,704 $\pm$ 0,054	0,713 $\pm$ 0,047
Random Forest	0,547 $\pm$ 0,036	0,536 $\pm$ 0,035
TF-IDF		
LinearSVC	0,815 $\pm$ 0,017	0,806 $\pm$ 0,020
BERTimbau Fine-Tuning		
BERTimbau	0,812 $\pm$ 0,052	0,849 $\pm$ 0,028

Embora o ganho sobre o baseline TF-IDF+LinearSVC seja pequeno em Macro F1 (0,849 vs. 0,815) e não significativo no teste de McNemar ($p = 0,711$), mantivemos o BERTimbau pelo maior recall da classe positiva com resposta institucional (0,803 vs. 0,597). A ablação com/sem resposta confirma que parte do ganho vem desse sinal retrospectivo: o recall passa de 0,703 para 0,803 ($p = 0,0159$). Esse efeito provavelmente decorre não de reconhecimento explícito do golpe, mas de padrões linguísticos indiretos, como menções ao MED ou vocabulário de contestação, que tendem a ocorrer sistematicamente em respostas a casos suspeitos.

O limiar operacional foi definido pela curva Precision-Recall. Adotamos probabilidade $\geq 0,70$, com precisão aproximada de 68% e recall de 72% para a classe **suspeito**. Aplicado

ao corpus completo de 221.519 reclamações, o modelo identificou **20.127 casos** candidatos, usados na etapa posterior de caracterização tipológica.

4.3 Modelagem de Tópicos

Como o BERTopic depende da representação semântica inicial dos documentos, a escolha dos *embeddings* afeta diretamente a formação dos tópicos. Esse ponto é especialmente relevante em reclamações bancárias, nas quais mecanismos distintos podem compartilhar termos como “Pix”, “conta” e “cartão”, enquanto casos semelhantes podem ser narrados com vocabulário diferente.

Comparamos três modelos de *embeddings*, mantendo fixas as demais etapas do pipeline, e avaliamos os resultados por C_v [25] e NPMI [4], métricas em que valores mais altos indicam tópicos mais coesos. O BGE-M3 apresentou o melhor desempenho ($C_v = 0,680$; NPMI = 0,167), superando o **mpnet** ($C_v = 0,632$; NPMI = 0,108) e o **multilingual-e5** ($C_v = 0,625$; NPMI = 0,103), sendo adotado na modelagem final.

Na configuração final, o UMAP foi utilizado com as configurações: **n_neighbors=10**, **n_components=5**, **min_dist=0,0** e métrica cosseno. Já o HDBSCAN utilizou os parâmetros **min_cluster_size = max(20, 0,3% N)**, **min_samples=5** e seleção EOM. A configuração final foi selecionada com base na comparação entre métricas de coerência e inspeção qualitativa dos agrupamentos. A solução resultante representou o melhor compromisso entre coerência semântica e interpretabilidade, evitando tanto a fragmentação excessiva quanto o agrupamento de mecanismos distintos em um mesmo tópico. Os *outliers* foram reatribuídos quando a similaridade com o tópico mais próximo superava o percentil 50 das similaridades internas. Por fim, os tópicos foram rotulados por LLM com auto-consistência [37], usando três execuções do mesmo *prompt* e votação majoritária.

4.4 Anotação de Canal e Gatilho

Após a definição da tipologia, os tópicos classificados como golpe foram analisados quanto ao canal de abordagem e ao gatilho psicológico explorado. As categorias de gatilho derivam dos princípios de persuasão descritos por Cialdini [8] e da caracterização de vítimas de golpes de Stajano e Wilson [30]. A anotação foi feita pelo **llama-3.3-70b** usando o *prompt* da Listagem 1, em uma amostra estratificada dos golpes, composta pelos 100 casos mais representativos de cada tópico, definidos pela similaridade ao centróide. Em tópicos com menos de 100 casos, todos os exemplos disponíveis foram incluídos.

A validação contra codificação humana independente, em uma amostra cega de 100 casos, obteve Cohen's $\kappa = 0,780$ para canal e $\kappa = 0,625$ para gatilho, com concordâncias simples de 0,829 e 0,686. Segundo Landis e Koch [18], ambos atingem concordância substancial. Como gatilho é mais ambíguo, seus cruzamentos são tratados como padrões indicativos sujeitos a maior incerteza.

¹Prompts e código disponíveis em <https://github.com/thaliaalmeida/webmedia2026-golpes-fraudes>.

```

1 Você é um analista de golpes e fraudes financeiras no Brasil.
2 Leia UMA reclamação e identifique, do ponto de vista do
3   CRIMINOSO,
4   dois atributos: canal e gatilho.
5
6 DEFINIÇÕES
7 - canal = o MEIO pelo qual o golpista fez o primeiro contato.
8   Pix, cartão e boleto NÃO são canal.
9 - gatilho = a ALAVANCA PSICOLÓGICA explorada para induzir a ví-
10  tima.
11
12 USE "não_aplicável" quando não houver golpista abordando a ví-
13 tima
14 (fraude sem engano, falha operacional).
15
16 USE "outro" se houve golpe, mas o relato não deixa claro o canal/
17 gatilho.
18
19 VALORES PERMITIDOS
20 canal:  ligação | WhatsApp | SMS | app | site | redes_sociais
21         | e-mail | presencial | correspondência
22         | não_aplicável | outro
23 gatilho: urgência | medo | ganância | confiança | autoridade
24         | não_aplicável | outro
25
26 RESPONDA APENAS um JSON em uma única linha, sem texto antes ou
27 depois:
28 {"canal": "<valor>", "gatilho": "<valor>", "motivo": "<até 8 palavras>"}

```

Listing 1: Prompt para anotação de canal e gatilho.¹

5 RESULTADOS

Nesta seção, apresentamos os resultados do corpus final e discutimos sua relação com as questões de pesquisa.

5.1 Dataset

A coleta contemplou perfis de bancos e instituições de pagamento do Brasil na plataforma Reclame Aqui, resultando em 221.519 reclamações brutas. Devido às restrições de paginação da plataforma, os registros concentram-se nos anos mais recentes, com sub-representação de períodos anteriores a 2022. Assim, volumes absolutos, comparações entre instituições e tendências temporais não são interpretados como medidas de risco ou prevalência; as proporções apresentadas descrevem apenas a composição interna do corpus coletado.

5.2 Análise de Erros

A Tabela 3 apresenta exemplos representativos de um verdadeiro positivo (VP), um falso positivo (FP) e um falso negativo (FN). O VP caracteriza um caso prototípico de golpe, com personificação de funcionário bancário, indução à transferência via Pix e narrativa sequencial com vocabulário discriminativo. O FP ilustra uma fonte recorrente de confusão: a presença do token [Editado pelo Reclame Aqui] e de construções linguísticas típicas de fraudes aproxima o relato da classe positiva, embora a reclamação descreva apenas uma disputa com a própria instituição. Já o FN corresponde a uma cobrança não reconhecida narrada de forma sucinta, sem vocabulário discriminativo ou sequência de eventos, tornando-o lexicalmente semelhante a reclamações operacionais.

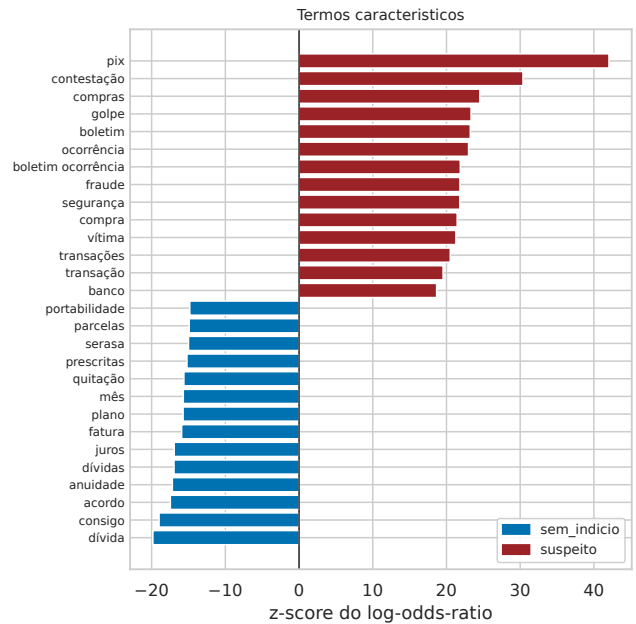


Figura 1: Termos mais característicos das classes suspeito e sem_indicio.

5.3 Padrões dos Relatos

5.3.1 Marcador de Censura da Plataforma. O token [Editado pelo Reclame Aqui] é inserido pela plataforma quando partes do conteúdo envolvem termos sensíveis, como “golpe”, “fraude”, palavras ou dados de terceiros. Sua ocorrência é substancialmente maior em casos suspeito (33,8%) do que em sem_indicio (7,1%), refletindo o uso mais frequente desse vocabulário nas reclamações de interesse. Entretanto, o marcador, isoladamente, não constitui evidência de fraude, funcionando apenas como um sinal auxiliar explorado pelo classificador em conjunto com o restante do contexto. Esse comportamento também ajuda a explicar parte dos falsos positivos discutidos na Seção 5.2.

5.3.2 Vocabulário Discriminativo. A Figura 1 contrasta os termos mais característicos de cada classe por *log-odds* com prior informativo, expresso como z-score. O vocabulário da classe suspeito é dominado por “pix”, “golpe”, “clonado”, “transferência” e “não reconhecida”; o vocabulário de sem_indicio, por “cobrança”, “cancelamento”, “atendimento” e “tarifa”.

5.4 Tipologia dos Casos

Dos 20.127 casos classificados como suspeitos, o BERTopic identificou 29 tópicos. Após a rotulagem e a inspeção dos exemplos representativos, 14.098 reclamações (≈70%) foram mapeadas para 19 tipos de golpes e fraudes. Os ≈30% restantes reúnem 1.815 reclamações (9,0%) em nove tópicos coerentes mas fora do escopo, como depósito não creditado, conta bloqueada por suspeita e cartão emitido sem solicitação, majoritariamente falsos positivos do classificador (Tabela 3), e 4.214 outliers (20,9%), relatos curtos ou idiossincráticos

Tabela 3: Exemplos de acertos e erros do classificador. Excertos truncados e anonimizados ([...] = trecho omitido).

	Excerto da reclamação	Ref.	Pred.
VP	“Recebi ligações do número do banco dizendo que tinha empréstimo em meu nome e que ia ser cancelado. No final da conversa muito convincente tenho um pix enviado pra alguém que não conheço de 1.900,00 e um empréstimo pessoal de 1.900,00. Fui ao banco cancelar [...] e disseram que não podem.”	suspeito	suspeito
FP	“[...] venho por meio dessa renomada plataforma expressar a minha revolta e insatisfação com a [Editado pelo Reclame Aqui] do PicPay que não realiza meu estorno e continua a me enrolar. [...] A quantidade que me subtraída é de um valor alto.”	sem_indicio	suspeito
FN	“[...] recebi uma notificação de uma cobrança de uber de 61,84. [...] não fiz essa compra!! Abri contestação e só recebi um aviso de que poderia demorar até 5 dias.”	suspeito	sem_indicio

sem a sequência narrativa dos tópicos identificados. Nenhum mecanismo de golpe emergente foi observado entre os tópicos excluídos.

A distribuição final divide-se entre *golpes* (até 47,8%) e *fraudes não autorizadas* (ao menos 52,2%), distinguindo casos baseados em engenharia social daqueles associados ao uso indevido de dados, cartões, contas ou contratos. Essa distinção é analiticamente relevante porque os dois grupos implicam mecanismos diferentes. Classificações setoriais costumam agrupar ambos sob a rubrica de “fraude bancária”, o que obscurece essa diferença e dificulta tanto a prevenção quanto a definição de critérios de ressarcimento [13].

A Tabela 4 mostra que a tipologia confirma modalidades já descritas na literatura e em relatórios setoriais, como golpes de engenharia social, falso funcionário, falsa venda, clonagem de cartão e uso indevido de identidade [13]. No entanto, os resultados também mostram que essas categorias aparecem de forma mais granular nos relatos das vítimas. Fraudes envolvendo cartão, por exemplo, não formam um bloco único, elas se dividem em compra não reconhecida, cartão furtado ou roubado, cartão clonado e cartão não entregue. Da mesma forma, o uso indevido de identidade aparece em diferentes formas contratuais, como conta aberta, empréstimo consignado, boleto, dívida desconhecida e financiamento de veículo em nome da vítima.

Dentre os golpes observados, destaca-se o *golpe do falso funcionário*, no qual o criminoso se apresenta como representante da instituição financeira para explorar relações de confiança e autoridade, padrão compatível com modalidades de engenharia social já descritas em alertas setoriais e no debate jurídico sobre falsa central de atendimento [14]. Tipos menores, como *ligações de tentativa de golpe*, *SMS de compra falsa* e *compra em site falso*, também remetem a vetores conhecidos de ataque, associados respectivamente a variações de *vishing*, *smishing* e *phishing*, isto é, tentativas de engano conduzidas por voz, mensagem curta ou páginas fraudulentas [12, 20]. Já o *golpe de premiação no TikTok*, embora tenha menor frequência, sugere a adaptação de golpes de falsa recompensa ao ambiente das plataformas sociais, nas quais promessas de ganho, perfis falsos e dinâmicas de engajamento podem ampliar a exposição das vítimas [19]. Assim, a tipologia confirma modalidades tradicionais de engenharia social, mas também indica a migração desses roteiros para canais digitais de sociabilidade e entretenimento.

Tabela 4: Tipologia de golpes e fraudes não autorizadas (n = 14.098).

Tópico	Tipo	Classe	%
1	Golpe via Pix por engenharia social	golpe	33,9
2	Compra não reconhecida no cartão	fraude	11,6
3	Golpe do falso funcionário	golpe	8,3
4	Cobrança indevida de seguro	fraude	7,7
5	Cartão furtado/roubado usado por terceiro	fraude	6,6
6	Empréstimo consignado não autorizado	fraude	4,6
7	Cartão clonado	fraude	4,1
8	Cartão não entregue	fraude	3,7
9	Boleto emitido em nome da vítima	fraude	3,5
10	Conta aberta em nome da vítima	fraude	3,2
11	Produto não entregue	fraude	2,7
12	Ligações de tentativa de golpe	golpe	1,8
13	Golpe de premiação no TikTok	golpe	1,7
14	Desaparecimento de saldo	fraude	1,7
15	Cobrança de dívida desconhecida	fraude	1,2
16	SMS de compra falsa	golpe	1,1
17	Golpe de compra em site falso	golpe	1,1
18	Financiamento de veículo em nome da vítima	fraude	1,0
19	Cobrança não autorizada (Uber)	fraude	0,5

O Tópico 1 merece atenção especial. Como ilustra a Tabela 5, o Pix aparece como etapa final comum de mecanismos distintos: personificação de autoridade ou contato confiável, tomada de conta, falso vendedor, falso advogado, falso parente, promoção enganosa e tarefas remuneradas. O tópico agrupa esses casos porque todos narram uma transferência imediata e difícil de reverter, embora os roteiros de engano sejam diferentes. Juridicamente, isso agrava a situação da vítima, pois o Mecanismo Especial de Devolução (MED) opera com prazos curtos e depende da cooperação da instituição destinatária.

5.5 Canal de Abordagem e Gatilhos

Gatilho e canal aplicam-se aos *golpes*; nas fraudes por uso indevido, não há manipulação psicológica direta da vítima. As proporções a seguir vêm de uma amostra de casos representativos por tópico e devem ser lidas como sinais de padrões recorrentes, não como estimativas populacionais.

Entre os golpes com canal identificável, os meios mais frequentes são ligação telefônica (34%), redes sociais (27%), SMS (17%) e WhatsApp (12%). Entre os gatilhos, predominam confiança (36%) e ganância (incentivo de ganho financeiro, 34%), seguidos de autoridade (17%), medo (9%) e urgência (3%). A manipulação observada apoia-se, portanto, mais em vínculo de confiança e promessa de ganho do que em pressão

Tabela 5: Exemplos representativos explorados no Tópico 1.

ID	Padrão observado	Exemplo de relato
1	Falso funcionário do banco	“recebi uma ligação do banco [INST] informando que estavam tentando fazer uma retirada em minha conta e por isso precisava trocar minha senha [...] fizeram um pix enquanto conversamos”
2	Falso advogado	“se passando por minha advogada, fez ligação de vídeo, conseguiu o acesso do app, conseguiu enviar um pix no valor de R\$ 800,00 [...] o [INST] deu estorno negado”
3	Falso parente	“uma pessoa se passou pela minha filha solicitando um Pix eu fiz o pagamento via Banco [...]”
4	Falso gerente	“Fui vítima de [EDITADO] via pix [...] na ocasião em que os interlocutores se identificaram como gerente do Banco [INST]. Durante o contato, solicitaram a confirmação de seus dados pessoais [...]”
5	Invasão do app/celular	“percebi meu celular em constante processo de atualização, tentei desligar, mas não consegui [...] foram feitas duas transferências pix da minha conta sem minha autorização”
6	Vendedor falso	“Realizei a compra de um produto pela internet e efetuei o pagamento via Pix. Após a confirmação do pagamento, a suposta vendedora deixou de responder minhas mensagens [...] não havendo qualquer entrega do produto”
7	Questionário falso em nome do banco	“o site prometia ajudar financeiramente se respondesse um questionário sobre marcas, e o site era referente ao nome do [INST] e tudo era muito bem feito [...] já tinha feito dois pix, o primeiro de 33,50 e o segundo de 29,90”

por urgência. Isto é, a vítima é frequentemente levada a agir porque acredita estar interagindo com uma fonte confiável, uma instituição legítima ou uma oportunidade vantajosa.

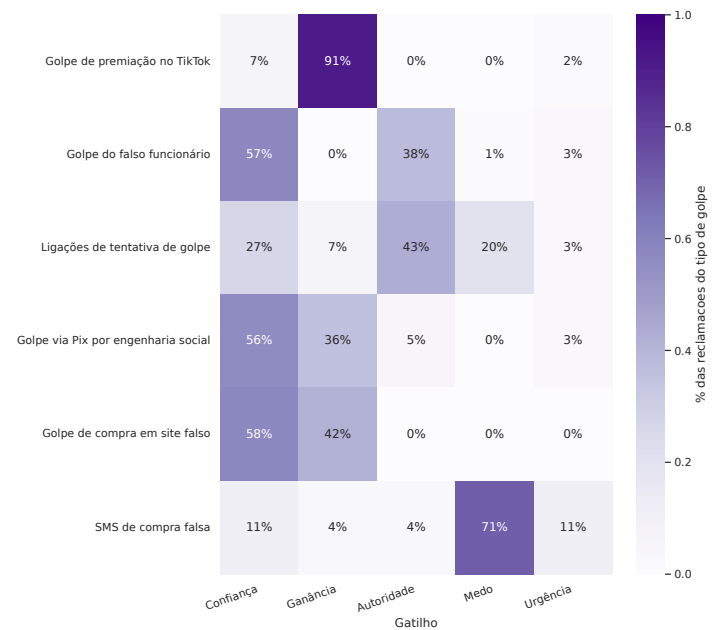
As Figuras 2 e 3 indicam que canal e gatilho não operam de forma independente. A confiança concentra-se no *golpe do falso funcionário* e no *golpe via Pix por engenharia social*, nos quais o criminoso simula proximidade institucional ou relação pessoal para induzir a transferência. A autoridade aparece sobretudo em ligações telefônicas, padrão compatível com golpes de falsa central ou falso atendente, nos quais a voz ao vivo facilita a encenação de legitimidade e reduz o tempo de verificação da vítima [32].

A ganância está associada às redes sociais, especialmente ao *golpe de premiação no TikTok*, no qual a promessa de ganhos fáceis constitui o principal mecanismo de persuasão. Nesses casos, a oferta de prêmios ou recompensas financeiras atua como isca para induzir a vítima a realizar ações que culminam em pagamentos aos golpistas. Já o medo aparece de forma mais clara no *SMS de compra falsa*, em que a mensagem simula uma transação suspeita e induz a vítima a reagir rapidamente. Assim, cada canal favorece um repertório específico de manipulação: confiança e autoridade pelo telefone, ganância nas redes sociais e medo por SMS.

Essa associação reforça que medidas de prevenção não devem tratar os canais apenas como vetores de contato, mas como ambientes sociotécnicos que favorecem certos roteiros de golpe. Os resultados indicam a necessidade de mensagens preventivas específicas por canal, como validação ativa de identidade em ligações, cautela com promessas de premiação em redes sociais, verificação independente de cobranças recebidas por SMS e atenção a pedidos de transferência via Pix em interações mediadas por WhatsApp ou telefone.

5.6 Resposta dos bancos

Dos 20.127 casos suspeitos, 18.382 (91,3%) possuem resposta da empresa. A modelagem dessas respostas com BERTopic revela três padrões principais: redirecionamento para canal privado ($n = 925$), que reduz a transparência pública do desfecho; respostas centradas no Mecanismo Especial de Devolução do Pix (MED, $n = 763$), que procedimentalizam o

**Figura 2: Distribuição de cada gatilho por tipo de golpe.**

caso sem reconhecer responsabilidade direta; e mensagens empáticas, porém genéricas ($n = 596$). Em conjunto, sugerem uma resposta institucional cautelosa, mais voltada a encaminhar o caso do que a reconhecer publicamente o golpe.

Além disso, existe um contraste expressivo na taxa de resolução entre golpe e fraude. Essa taxa é definida pelo campo **resolvida** da plataforma Reclame Aqui, no qual o próprio consumidor informa se considera o problema resolvido. Trata-se de uma avaliação autodeclarada, não de verificação independente. Com essa métrica, os golpes são marcados como resolvidos a quase metade da taxa das fraudes (16,7% vs. 29,3%), diferença altamente significativa ($\chi^2_1 = 314,0$, $p < 0,001$). A Figura 4 mostra a taxa de resolução para golpes e fraudes não autorizadas. A diferença sugere que fraudes não

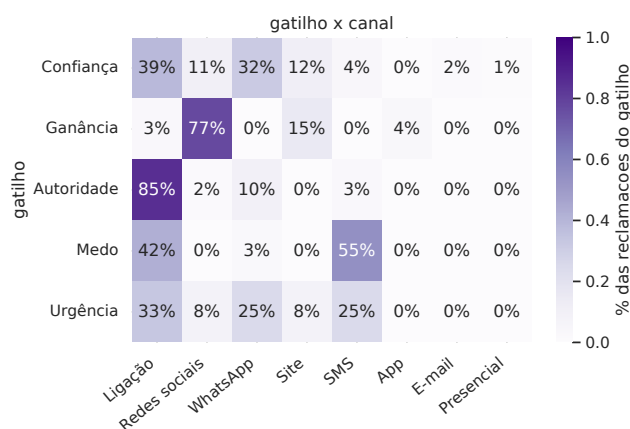


Figura 3: Distribuição de cada gatilho por canal de abordagem.

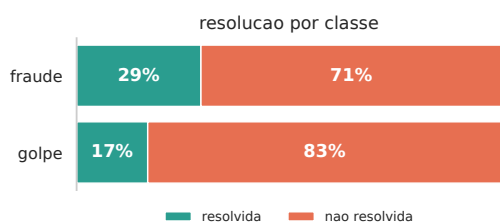


Figura 4: Taxa de resolução das reclamações por classe (golpe e fraude não autorizada).

autorizadas se ajustam melhor a fluxos tradicionais de contestação e estorno, pois a não participação da vítima tende a ser mais objetiva. Nos golpes, a vítima muitas vezes realiza a operação sob manipulação, deslocando a disputa para falha de segurança, transação atípica e culpa concorrente. Assim, fraudes tendem a gerar resolução mais clara, enquanto golpes produzem respostas procedimentais e menos conclusivas [5, 32].

5.7 Limitações

Este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos resultados. A caracterização foi construída sobre as reclamações identificadas pelo pipeline, cuja configuração final concatena a resposta institucional ao texto da reclamação. As análises da Seção 5.6 herdaram, assim, uma dependência entre seleção e desfecho, de magnitude limitada, já que o MED cobre 4,2% das respostas e a separação entre golpe e fraude decorre do texto da vítima. Como o Tópico 1 agrega roteiros heterogêneos e dois tópicos excluídos tangenciam o uso indevido por terceiros, a fatia de 47,8% atribuída a golpes é um limite superior, e as distribuições apresentadas são estimativas agregadas, não classificações exatas.

O corpus reúne apenas reclamações espontaneamente publicadas, sujeito a vieses de autorrelato e cobertura, de modo que os resultados descrevem esse ecossistema, não a prevalência nacional. Por fim, o conjunto de referência foi anotado por um único autor, e a concordância dos LLMs com essa

referência não substitui concordância inter-anotadores humana; a anotação de canal e gatilho, feita sobre os casos mais próximos dos centróides, favorece relatos prototípicos e tende a exibir associações mais nítidas do que na população.

6 CONCLUSÃO

Este trabalho investigou como reclamações de consumidores podem ser utilizadas para caracterizar golpes e fraudes bancárias em larga escala. Para isso, propusemos um pipeline híbrido que combina supervisão fraca, anotação por LLM e ajuste fino de BERTimbau para identificar automaticamente relatos de interesse em um grande corpus, viabilizando uma caracterização inédita desse fenômeno em português.

A análise de mais de 20 mil relatos revelou uma tipologia de 19 mecanismos recorrentes, evidenciando a centralidade do Pix em diferentes golpes de engenharia social, padrões consistentes de canais de abordagem e gatilhos psicológicos, além de diferenças significativas entre golpes e fraudes não autorizadas nas respostas institucionais e na percepção de resolução. Esses resultados mostram que golpes financeiros apresentam padrões recorrentes que podem ser identificados sistematicamente nas narrativas das próprias vítimas.

Entre as limitações do estudo estão o viés temporal do corpus, a ausência de uma verdade jurídica para os casos analisados, a construção do conjunto de referência por um único anotador humano e o uso de LLMs em diferentes etapas do pipeline. Trabalhos futuros incluem ampliar a cobertura temporal e institucional, incorporar validação por especialistas e investigar a evolução temporal das modalidades de golpe.

Mais amplamente, este trabalho demonstra o potencial de plataformas de reclamação como observatórios em larga escala da dinâmica dos golpes financeiros. Ao transformar relatos espontâneos de consumidores em evidências sobre mecanismos de fraude, estratégias de engenharia social e respostas institucionais, essa abordagem abre possibilidades para pesquisas em PLN, prevenção a fraudes e formulação de políticas públicas.

7 CUIDADOS ÉTICOS

Esta pesquisa utilizou exclusivamente dados públicos do Reclame Aqui, sem interação com participantes. A coleta foi sequencial, com intervalos de 1,5 e 0,75 segundo entre requisições de listagem e de reclamação, para não sobrecarregar a plataforma. Conforme a Resolução CNS nº 510/2016, os dados foram usados apenas para fins científicos, analisados de forma agregada e, quando reproduzidos, anonimizados e truncados, observando os princípios da Lei nº 13.709/2018 (LGPD).

AGRADECIMENTOS

Este trabalho foi apoiado pelo Instituto Kunumi, CAPES, CNPq, Fapemig, AWS, NVIDIA, INCT-IACiber (processo nº 408432/2024-1) e INCT-TILD-IAR (processo nº 408490/2024-1). O modelo GPT 5.5 foi empregado como apoio à revisão textual. As sugestões geradas foram avaliadas pelos autores, que permanecem responsáveis pelas decisões metodológicas, pela interpretação dos resultados e pelas conclusões do estudo.

REFERÊNCIAS

- [1] Banco Central do Brasil. 2025. IF.data: Dados Seleccionados de Entidades Supervisionadas. <https://www3.bcb.gov.br/ifdata/>. Data-base: 12/2025. Acesso em: 19/08/2026.
- [2] Kaveh Bastani, Hamed Namavari, and Jeffry Shaffer. 2019. Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints. *Expert Systems with Applications* 127 (2019), 256–271. arXiv:1807.07468 [cs.IR] doi:10.1016/j.eswa.2019.03.001
- [3] Petros Boulrieris, John Pavlopoulos, et al. 2023. Fraud Detection with Natural Language Processing. *Machine Learning* (2023). doi:10.1007/s10994-023-06354-5
- [4] Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of the Biennial GSCCL Conference*. 31–40.
- [5] Jo Braithwaite. 2024. “Authorized Push Payment” Bank Fraud: What Does an Effective Regulatory Response Look Like? *Journal of Financial Regulation* 10, 2 (2024), 174–193. doi:10.1093/jfr/fjae006
- [6] Brasil. 1940. Decreto-Lei n. 2.848, de 7 de dezembro de 1940. Código Penal. https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm Acesso em: 28 jun. 2026.
- [7] Mark Button, Chris Lewis, and Jacki Tapley. 2009. *Fraud Typologies and the Victims of Fraud: Literature Review*. Technical Report. National Fraud Authority, London. <https://researchportal.port.ac.uk/en/publications/fraud-typologies-and-the-victims-of-fraud-literature-review> Acesso em: 28 jun. 2026.
- [8] Robert B. Cialdini. 2006. *Influence: The Psychology of Persuasion* (revised ed.). Harper Business.
- [9] Washington Cunha, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2023. An Effective, Efficient, and Scalable Confidence-Based Instance Selection Framework for Transformer-Based Text Classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 665–674. doi:10.1145/3539618.3591638
- [10] Washington Cunha, Felipe Viegas, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2025. A Noise-Oriented and Redundancy-Aware Instance Selection Framework. *ACM Transactions on Information Systems* 43, 1 (2025). doi:10.1145/3701769
- [11] Dados Bancários. 2026. *Ranking de Instituições por Ativo Total – Maiores Bancos do Brasil*. <https://dadosbancarios.com.br/ranking/ativo-total/> Dados compilados a partir do sistema IF.data do Banco Central do Brasil.
- [12] European Union Agency for Cybersecurity. s.d.. Threat Landscape. <https://www.enisa.europa.eu/topics/cyber-threats/threat-landscape>. Acesso em: 29 jun. 2026.
- [13] Febraban. 2024. *Comitê Executivo da Febraban lança Selo de Prevenção a Fraudes das Instituições Financeiras*. Federação Brasileira de Bancos. <https://portal.febraban.org.br/noticia/3996/pt-br/> Inclui estimativa de que cerca de 70% dos golpes são de engenharia social.
- [14] Febraban. 2025. *Golpe da falsa venda é o mais aplicado contra clientes bancários no 1º semestre do ano*. Federação Brasileira de Bancos. <https://portal.febraban.org.br/noticia/4381/pt-br/>
- [15] David Fernandes, Edleno Silva de Moura, Berthier A. Ribeiro Neto, Altigran Soares da Silva, and Marcos André Gonçalves. 2007. Computing Block Importance for Searching on Web Sites. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*. ACM, 165–174. doi:10.1145/1321440.1321466
- [16] Maria Beatriz Giusti. 2025. Mais de 24 milhões de pessoas foram vítimas de golpes pelo Pix em um ano. Senado Notícias. <https://www12.senado.leg.br/noticias/audios/2025/08/mais-de-24-milhoes-de-pessoas-foram-vitimas-de-golpes-pelo-pix-em-um-ano> Acesso em: 28 jun. 2026.
- [17] Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *arXiv preprint arXiv:2203.05794* (2022).
- [18] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. doi:10.2307/2529310
- [19] Terri Miller. 2023. Slow your scroll: Spot and avoid social media giveaway scams. <https://consumer.ftc.gov/consumer-alerts/2023/11/slow-your-scroll-spot-avoid-social-media-giveaway-scams>. Federal Trade Commission. Acesso em: 29 jun. 2026.
- [20] National Institute of Standards and Technology. s.d.. Phishing. <https://csrc.nist.gov/glossary/term/phishing>. NIST Computer Security Resource Center Glossary. Acesso em: 29 jun. 2026.
- [21] Antonis Papasavva, Max Aliapoulos, et al. 2024. Applications of AI-Based Models for Online Fraud Detection and Analysis. *Crime Science* (2024). Systematic Literature Review.
- [22] Gabriel Prenassi, Ana Machado, Ester Souza, Mateus Brito, Davi Reis, Jessica Costa, Geovana Oliveira, Carlos Ferreira, and Leonardo Rocha. 2025. Narrativas de Jogos de Azar em Plataformas de Vídeo: Um Estudo Linguístico-Temático sobre o jogo do Tigrinho no YouTube. In *Anais do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web*. Sociedade Brasileira de Computação, Porto Alegre, RS, Brasil, 366–375. doi:10.5753/webmedia.2025.15565
- [23] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment* 11, 3 (2017), 269–282. doi:10.14778/3157794.3157797
- [24] Reclame AQUI. 2026. Como funciona. <https://www.reclameaqui.com.br/como-funciona/> Acesso em: 28 jun. 2026.
- [25] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of WSDM*. 399–408.
- [26] Said Salloum, Tarek Gaber, Sunil Vadera, and Khaled Shaalan. 2022. A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques. *IEEE Access* 10 (2022), 65703–65727. doi:10.1109/ACCESS.2022.3183083
- [27] Vasudeva Raju Sangaraju, Bharath Kumar Bolla, Deepak Kumar Nayak, and Jyothsna Kh. 2022. Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings. arXiv:2205.07259 [cs.LG] doi:10.48550/arXiv.2205.07259
- [28] Serasa Experian. 2025. *Mais da metade dos brasileiros já foi vítima de fraude e 20% deles perderam até R\$ 5 mil, revela estudo inédito da Serasa Experian*. Serasa Experian. <https://www.serasaexperian.com.br/sala-de-imprensa/prevencao-a-fraude/mais-da-metade-dos-brasileiros-ja-foi-vitima-de-fraude-e-20-deles-perderam-ate-rdollar-5-mil-revela-estudo-inedito-da-serasa-experian/>
- [29] Serasa Experian. 2025. Relatório de Identidade Digital e Fraude 2025. <https://www.serasaexperian.com.br/images-cms/wp-content/uploads/2025/03/Relatorio-de-Identidade-Digital-e-Fraude-2025-Serasa-Experian.pdf>
- [30] Frank Stajano and Paul Wilson. 2011. Understanding Scam Victims: Seven Principles for Systems Security. *Commun. ACM* 54, 3 (2011), 70–75. doi:10.1145/1897852.1897872
- [31] Superior Tribunal de Justiça. 2012. Súmula 479. <https://processo.stj.jus.br/SCON/sumstj/doc.jsp?b=SUM&i=1&l=10&operador=E&ordenacao=%40NUM&p=false&sumula=479.num> As instituições financeiras respondem objetivamente pelos danos gerados por fortuito interno relativo a fraudes e delitos praticados por terceiros no âmbito de operações bancárias. Acesso em: 28 jun. 2026.
- [32] Superior Tribunal de Justiça. 2025. Bancos e instituições de pagamento devem indenizar clientes por falhas que viabilizam golpe da falsa central. <https://www.stj.jus.br/sites/portallp/Paginas/Comunicacao/Noticias/2025/21102025-Bancos-e-instituicoes-depagamento-devem-indenizar-clientes-por-falhas-que-viabilizam-golpe-da-falsa-central.aspx> Notícia institucional referente aos REsp 2.222.059 e 2.229.519. Acesso em: 28 jun. 2026.
- [33] Zeya Lwin Tun and Daniel Birs. 2023. Supporting Crime Script Analyses of Scams with Natural Language Processing. *Crime Science* 12, 1 (2023). doi:10.1186/s40163-023-00177-4
- [34] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Cunha, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 753–761. doi:10.1145/3289600.3291032
- [35] Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos André Gonçalves. 2020. CluHTM: Semantic Hierarchical Topic Modeling Based on CluWords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 8138–8150. doi:10.18653/v1/2020.acl-main.724

- [36] Felipe Viegas, Washington Cunha, Christian Gomes, Leonardo Rocha, and Marcos André Gonçalves. 2025. Exploiting Contextual Embeddings in Hierarchical Topic Modeling and Investigating the Limits of the Current Evaluation Metrics. *Computational Linguistics* (2025).
- [37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of ICLR*.
- [38] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (2024), 237–291. doi:10.1162/coli_a_00502