

Números da Sorte e Jogos de Azar: Caracterização das Bets Irregulares em Canais do Telegram no Brasil

Samuel Guimarães
samuelsg@ufmg.br
UFMG
Belo Horizonte, Brasil

Karen Gomes
karen.gomes@dcc.ufmg.br
UFMG
Belo Horizonte, Brasil

Nicolas Reis
nicolas.e.silva@ufv.br
UFV
Viçosa, Brasil

Filipe N. Ribeiro
filipe.ribeiro@ufop.edu.br
UFOP
João Monlevade, Brasil

Julio C. S. Reis
jreis@ufv.br
UFV
Viçosa, Brasil

Fabrizio Benevenuto
fabrizio@ufmg.br
UFMG
Belo Horizonte, Brasil

ABSTRACT

In Brazil, the online betting market outside legalized fixed-odds betting platforms is estimated to account for over 40% of the market share, generating billions in revenue. As unregulated platforms, they can offer odds that are unfeasible for regulated competitors, allow users who are otherwise prohibited from playing, and use banned game mechanics such as starting bonuses. To reach users, these platforms often use instant messaging apps like Telegram, which is already used for some illegal activities. This study investigates public Brazilian Telegram channels to identify and characterize unregulated betting websites by examining common domain characteristics, searching for evidence of coordinated action among creators, and analyzing channel content. In total, we find 6,523 websites, with evidence of coordination between domains based on Ad Analytics IDs, CSS and JavaScript files, and HTML templates, as well as evidence of phishing based on alerts from antivirus scanners. Finally, we find that channel content promises easy money through leaked bonus codes, supposed bugs in betting games, and raffles organized in the channels. Our results provide insights that we hope will improve understanding and help combat irregular betting.

KEYWORDS

Apostas de quota fixa, Bets, Golpes

1 INTRODUÇÃO

As plataformas de apostas online, popularmente conhecidas como *bets*, constituem um mercado em expansão no Brasil. A regulamentação do setor iniciou-se com a Lei n.º 13.756/2018 e foi consolidada pela Lei n.º 14.790/2023, que criou a Secretaria de Prêmios e Apostas do Ministério da Fazenda (SPA/MF) para autorizar e fiscalizar as operadoras. Desde 2025, apenas plataformas licenciadas podem atuar legalmente no país. Apesar desse avanço regulatório, o mercado de *bets*

irregulares permanece ativo, recorrendo a canais digitais para atrair usuários e contornar as ações de fiscalização.

Por operarem à margem da regulamentação, essas plataformas podem oferecer vantagens vedadas às operadoras autorizadas, como bônus promocionais, meios de pagamento proibidos e menor controle sobre a identificação dos usuários, comprometendo mecanismos de prevenção à lavagem de dinheiro e de proteção aos apostadores.

Segundo um relatório de auditoria do Tribunal de Contas da União (TCU) [20], com base em estudo elaborado pela LCA Consultoria Econômica para o Instituto Brasileiro de Jogo Responsável (IBJR) [14], estima-se que o mercado de *bets* irregulares represente entre 41% e 51% do mercado brasileiro de apostas, movimentando anualmente entre R\$ 26 bilhões e R\$ 40 bilhões.

Em relação ao comportamento dos usuários, pesquisa do Instituto Locomotiva, realizada em abril de 2025 com 2 mil apostadores adultos, mostrou que 61% dos entrevistados haviam apostado em plataformas irregulares e que 78% tinham dificuldade em distingui-las das operadoras autorizadas.¹

Como parte das ações de fiscalização, a Anatel e a Secretaria de Prêmios e Apostas do Ministério da Fazenda (SPA/MF) promoveram o bloqueio de milhares de domínios de plataformas irregulares. Segundo o TCU [20], até setembro de 2025 haviam sido encaminhadas 21.854 URLs para bloqueio, das quais 17.987 (82,3%) já estavam inacessíveis. O relatório também destaca que essas plataformas frequentemente contornam os bloqueios por meio do registro contínuo de novos domínios e da divulgação de novas URLs em redes sociais e aplicativos de mensagens.

Entre esses ambientes, o Telegram destaca-se por permitir grupos e canais públicos acessíveis por links permanentes, favorecendo a divulgação de plataformas de apostas. Apesar dessa relevância, ainda são escassos estudos que caracterizam sistematicamente a atuação das *bets* irregulares nesse ambiente.

Para preencher essa lacuna, realizamos um estudo em larga escala sobre a atuação de plataformas de apostas irregulares no Telegram, buscando responder às seguintes perguntas de pesquisa:

¹<https://ilocomotiva.com.br/estudos/mercado-ilegal-apostas-brasil-2025/>

QP1. Como as plataformas de apostas irregulares utilizam o Telegram para divulgar e manter o acesso aos seus serviços?

QP2. Existem evidências de coordenação entre diferentes plataformas de apostas irregulares no Telegram?

QP3. Há evidências da ocorrência de golpes ou outras práticas potencialmente fraudulentas associadas a essas plataformas?

2 TRABALHOS RELACIONADOS

Diversos trabalhos abordaram temas relacionados ao nosso estudo, incluindo a caracterização de comunidades no Telegram, análises do ecossistema de apostas online e métodos de identificação de domínios ou URLs suspeitas.

Entre os estudos de caracterização do ecossistema do Telegram, o TGDataset [12] introduziu um dos maiores conjuntos de dados da plataforma, coletando mais de 120 mil canais e 400 milhões de mensagens por meio da API Telethon. A partir desse conjunto de dados, os autores realizaram uma caracterização dos canais por meio da modelagem de tópicos, identificando, entre outros temas, comunidades associadas a atividades suspeitas, como mercados clandestinos. O TGDataset também serviu de base para estudos posteriores, incluindo investigações sobre contas falsas [13] e sobre a disseminação de teorias da conspiração no Telegram [10]. Entretanto, esses trabalhos têm caráter geral e não investigam comunidades dedicadas à divulgação de plataformas de apostas.

No contexto brasileiro, Kansan et al. [11] caracterizaram mais de 200 mil grupos públicos do WhatsApp, obtidos por meio de links de convite compartilhados em redes sociais e em sites agregadores. Utilizando modelagem de tópicos, os autores identificaram, entre outros assuntos, grupos associados a promessas de dinheiro fácil e a esquemas de apostas. No entanto, o estudo foi realizado antes da regulamentação do mercado de apostas e focou no WhatsApp, sem investigar a atuação de plataformas de apostas irregulares no Telegram.

Ainda no Telegram, Perlo et al. [17] analisaram 669 grupos do TGStat, identificando campanhas de divulgação em larga escala e compartilhamento repetitivo de URLs, inclusive em comunidades relacionadas a apostas esportivas e criptomoe-das. Embora o estudo evidencie a presença de comunidades relacionadas a apostas no Telegram, seu objetivo foi caracterizar os tópicos do TGStat, sem investigar especificamente a atuação de plataformas de apostas irregulares.

No contexto das plataformas de apostas, Vaidya et al. [21] caracterizaram o ecossistema de *bets* na Índia a partir de anúncios na Meta, publicações no Instagram e avaliações do Google Play, identificando estratégias recorrentes de marketing baseadas em urgência e escassez. Embora realizado em um contexto regulatório distinto, o estudo identificou plataformas também presentes em nossa coleta e padrões de comunicação semelhantes aos observados neste trabalho.

No contexto brasileiro, Prenassi et al. [18] e Costa et al. [5] investigaram conteúdos relacionados ao *Fortune Tiger* no YouTube, analisando, respectivamente, as narrativas dos vídeos e o posicionamento dos usuários nos comentários. Diferentemente desses estudos, nosso trabalho caracteriza plataformas de apostas irregulares no Telegram, considerando

conjuntamente as narrativas, as estratégias de divulgação e a infraestrutura dos domínios.

Além da análise do conteúdo compartilhado, trabalhos anteriores também exploraram informações de infraestrutura para caracterizar domínios e URLs associados a atividades suspeitas. Couto et al. [6], por exemplo, investigaram diferenças entre sites brasileiros de baixa e alta credibilidade com auxílio de atributos extraídos de protocolos de rede, como idade do domínio, informações de registro, certificados SSL e localização da hospedagem. Os resultados mostraram que essas características permitem distinguir diferentes tipos de sites e identificar padrões associados a domínios de desinformação.

Em outra linha de pesquisa, Acharya et al. [1] investigaram campanhas fraudulentas de arrecadação de doações divulgadas em plataformas como Instagram, Facebook, Telegram e YouTube. Além da análise do conteúdo publicado, os autores utilizaram informações dos domínios associados às campanhas, incluindo verificações do VirusTotal, para identificar URLs potencialmente maliciosas e caracterizar os golpes. Inspirados por essas abordagens, neste trabalho utilizamos atributos de infraestrutura dos domínios e informações do VirusTotal para analisar plataformas de apostas irregulares, identificando padrões comuns e possíveis evidências de coordenação entre diferentes operadores.

3 METODOLOGIA

Nesta seção, descrevemos a metodologia aplicada, que incluiu a busca por canais no Telegram, a análise de conteúdo dos mesmos, a extração de domínios de sites de apostas irregulares das mensagens, o agrupamento dos sites por similaridade, e a coleta de dados sobre esses domínios.²

3.1 Coleta de dados

A coleta focou em canais públicos do Telegram, motivada tanto pela existência de uma API oficial para acesso às mensagens quanto por evidências na literatura de que a plataforma é amplamente utilizada para divulgação de atividades ilícitas, incluindo plataformas de apostas [10, 12, 17].

Como ponto de partida, utilizamos o Telemetr.io³, um serviço que indexa grupos e canais públicos do Telegram juntamente com metadados como categoria, número de membros e volume de mensagens. A partir dos 100 maiores canais brasileiros listados pela plataforma, identificamos um conjunto inicial de 20 canais relacionados a apostas, dos quais apenas um se associou a uma plataforma legalizada.

Em seguida, analisamos os padrões presentes nos nomes desses canais, observando o uso recorrente do domínio da plataforma de apostas combinado com termos como *Canal*, *Chat*, *Comunidade*, *Oficial* e *VIP*. Esses padrões foram utilizados para construir consultas em mecanismos de busca (Google e DuckDuckGo), restringindo os resultados a links públicos do Telegram por meio do filtro “site:t.me” seguido das palavras-chave citadas e das palavras “aposta” e “bet”.

²<https://github.com/samuelSg1/webmedia26>

³Também acessível pelo link <https://tlmtr.io/>

Tabela 1: Descrição das Técnicas de medição de Similaridade empregadas

Técnica	Processo de Detecção de Similaridade
Hash estrutural	Serializa a árvore DOM do HTML (nomes de tags + classes CSS + IDs, sem incluir texto) e calcula o hash SHA-256, fazendo sites com o mesmo template ter o mesmo hash
Hash de conteúdo	Hash de similaridade do texto visível + tokens CSS, sensível à localidade. Conteúdo é considerado similar com a distância de Hamming ≤ 6 bits
Hash de pacotes JS	Hash SHA-256 de cada arquivo .js baixado, onde o mesmo hash indica o código JS idêntico, independentemente do domínio
Hash de arquivos CSS	SHA-256 de cada arquivo .css baixado, onde o mesmo hash indica o mesmo CSS
IP do servidor	Resolve DNS e obtém o IP da hospedagem ou operador. Mesmo IP indica a mesma hospedagem.
Certificado SSL/TLS	Lê os Nomes Alternativos do Assunto (SAN) do certificado TLS e SSL. O mesmo certificado em vários domínios indica o mesmo operador.
IDs de rastreamento de anúncios	Extraí IDs do Google Analytics (UA-/G-), GTM, Pixel do Facebook, Hotjar, Clarity e Pixel do TikTok. Mesmo ID representa a mesma conta nesses serviços e um mesmo criador.

Esse processo permitiu expandir a coleta para 125 canais públicos, totalizando 1.447.707 mensagens.

Por fim, extraímos todas as URLs presentes nas mensagens, utilizando os metadados fornecidos pela API do Telegram e obtivemos 7.164 domínios distintos. Os domínios foram inicialmente filtrados por palavras-chave frequentemente associadas a plataformas de apostas (e.g., *bet*, *win*, *slot*, *luck*) e, posteriormente, os casos restantes foram inspecionados manualmente. Ao final desse processo, identificamos 6.523 domínios associados a plataformas de apostas irregulares.

3.2 Análise do conteúdo dos canais

Em seguida, analisamos o conteúdo das mensagens enviadas nos canais. Para isso, empregamos quatro algoritmos distintos para modelagem de tópicos: Alocação Latente de Dirichlet (LDA) [3], BERTopic [9] e Fatoração de Matriz Não Negativa (NMF) [15] com duas funções de perda, a Divergência de Kullback-Leibler (NMF-KL), e a Norma de Frobenius (NMF-F). Aplicamos os algoritmos a todas as mensagens com links de *bets* para avaliar como as plataformas são descritas e oferecidas aos membros. Como pré-processamento, o texto é padronizado em letras minúsculas, os emojis são removidos, os links foram substituídos pela tag “<link>”, e as mensagens repetidas são contadas e removidas. Devido a uma característica das propagandas de *bets*, os horários também foram agrupados e substituídos pela tag “<horario>”. Após executar os algoritmos, realizamos uma análise dos Top 10 tópicos baseado na coerência e diversidade dos tópicos medidas pelas métricas de Coerência por Informação Mútua Pontual Normalizada (NPMI) [2] e a média da distância de Jaccard entre pares de tópicos [19], respectivamente. Utilizamos os tópicos do melhor algoritmo em nossas análises.

3.3 Agregação dos Domínios por Similaridade

A partir dos links extraídos das mensagens dos canais, analisamos a similaridade entre os domínios de apostas irregulares para identificar evidências de coordenação entre diferentes plataformas (QP2). Para cada domínio, coletamos atributos relacionados ao conteúdo da página (texto, estrutura HTML e arquivos JavaScript) e à infraestrutura do serviço web (endereço IP, certificado SSL/TLS e identificadores de rastreamento, como o Meta Pixel). A Tabela 1 resume as técnicas empregadas.

Para detectar páginas quase duplicadas, utilizamos *fingerprints* de similaridade baseados em SimHash [16]. Após avaliação empírica de diferentes limiares, adotamos distância de Hamming $d \leq 6$.

Por fim, construímos um grafo para cada métrica de similaridade, no qual os vértices representam domínios e as arestas indicam características em comum. Também construímos dois grafos agregados: um considerando todas as métricas e outro excluindo a similaridade de conteúdo e estrutura HTML, por serem mais suscetíveis a conectar plataformas distintas que compartilham apenas características comuns de sites de apostas.

3.4 Análise dos Domínios

Após coletar e agregar todos os domínios de *bets* irregulares presentes nos grupos, analisamos os domínios a partir de uma série de atributos usados em trabalhos anteriores [6, 7] para avaliar sites de baixa credibilidade. Além disso, também avaliamos os domínios quanto à presença de malwares e ameaças de segurança utilizando a API de um serviço de escaneamento antivírus online.

Atributos de Domínio. O registro e a configuração de serviços de DNS (Domain Name System) disponibilizam informações públicas sobre as características de um domínio como a duração total da compra de um nome de domínio, o tempo de criação do domínio e, quando não censurado, as informações dos donos do website. Dessa forma, utilizamos os protocolos WHOIS [8], RDAP [8] e DIG para obter informações relevantes de domínios encontrados durante a coleta.

Atributos de Geolocalização. A localização dos servidores pode fornecer indícios sobre a infraestrutura utilizada pelas *bets*. Para extrair atributos de geolocalização, utilizamos a API IPStack⁴, obtendo informações como endereço IP, localização geográfica e Sistema Autônomo (AS) associado aos domínios.

Escaneamento de Vírus e Malwares Online. Por fim, utilizamos a API do VirusTotal⁵ para verificar os domínios em 91 mecanismos de detecção de ameaças. Para cada domínio, coletamos o resultado consolidado da análise, incluindo classificações como *harmless*, *suspicious*, *malicious* e *undetected*, utilizadas posteriormente na caracterização dos sites.

⁴<https://ipstack.com/>

⁵<https://docs.virustotal.com/docs/api-overview>

4 RESULTADOS

Nesta seção, apresentamos e discutimos os resultados das análises realizadas que visam caracterizar o ecossistema de *bets* irregulares no Telegram Brasileiro.

4.1 Caracterização dos Canais de Telegram

Primeiramente, realizamos uma caracterização geral dos canais que compõem nosso conjunto de dados, considerando seu tamanho e distribuição de mensagens. A Figura 1 apresenta a função de distribuição acumulada do número de membros dos canais coletados. Cerca de 90% dos canais possuem mais de mil membros, enquanto 70% superam 10 mil e aproximadamente 40% possuem mais de 100 mil membros.

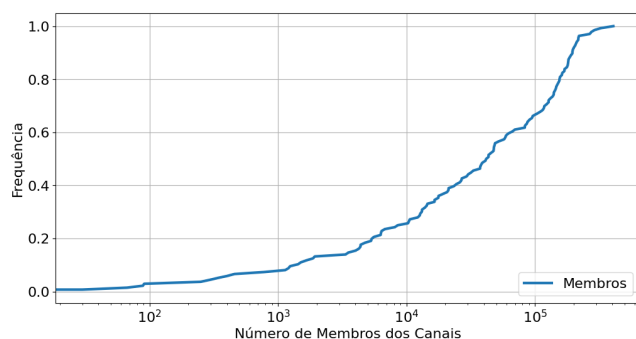


Figura 1: CDF do Número de Membros dos Canais

A escala dessas comunidades é particularmente relevante quando comparada à de canais associados a *bets* legalizadas. Segundo o Telemetr.io, o maior canal associado a uma *bet* legalizada (*blaze.bet.br*) possuía 469.553 membros no momento da coleta. Entre as *bets* irregulares, o maior canal identificado possuía 407.471 membros. Entretanto, como o canal associado à *blaze.bet.br* é privado, considerando apenas canais públicos, o maior canal de uma *bet* legalizada (*betapp.bet.br*) possuía 166.003 membros. Esses números mostram que canais associados a *bets* irregulares podem reunir centenas de milhares de usuários, evidenciando a escala que esse ecossistema pode alcançar no Telegram.

Além da distribuição do número de membros, analisamos a evolução temporal das mensagens que continham links para *bets* irregulares. A Figura 2 apresenta essa distribuição ao longo do período analisado. Observa-se um primeiro pico em 9 de novembro de 2023, após um período de baixa atividade, seguido de uma queda abrupta e de um aumento mais persistente no volume de mensagens a partir de 2024. O pico de novembro de 2023 está associado principalmente à divulgação repetitiva, em um mesmo canal, de uma nova plataforma de apostas que oferecia bônus de lançamento por tempo limitado. As diferentes versões da mensagem enfatizavam a disponibilidade do novo serviço e condições como depósito e saque mínimos, além da ausência de taxas.

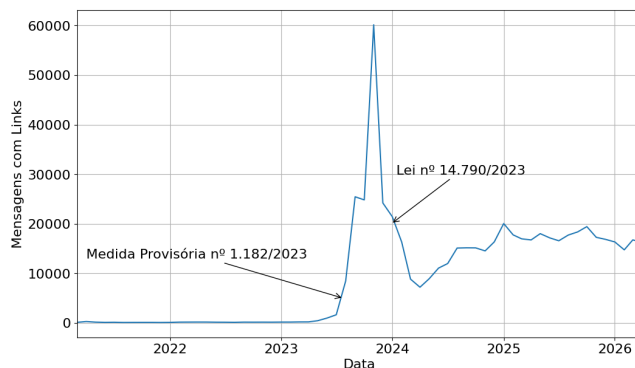


Figura 2: Linha do Tempo das Mensagens com Links

A recorrência de mensagens promocionais e o crescimento sustentado de sua frequência indicam que os canais do Telegram constituem um importante meio de divulgação e distribuição de links para plataformas de apostas irregulares. Em particular, a persistência dessa atividade após o início da regulamentação, que ocorreu ao longo de 2023 com a medida provisória n.º 1.182 e a lei n.º 14.790/2023, sugere que as medidas de bloqueio não eliminaram a presença dessas plataformas no ambiente analisado.

4.2 Análise do conteúdo das mensagens dos canais de Telegram

Em seguida, analisamos como os canais do Telegram promovem *bets* irregulares. Conforme descrito na Seção 3.2, aplicamos diferentes algoritmos de modelagem de tópicos às mensagens contendo links para essas plataformas. Após avaliar a coerência e diversidade dos tópicos, o NMF-KL apresentou o melhor desempenho, com os seguintes hiperparâmetros: $\text{beta_loss} = \text{"kullback-leibler"}$, $\text{init} = \text{"nndsvda"}$, $\text{solver} = \text{"mu"}$, $\text{max_iter} = 1000$, $\text{l1_ratio} = 0,5$, $\text{alpha_W} = 0,00005$, $\text{alpha_H} = 0,00005$. A Tabela 2 resume os dez tópicos identificados pelo NMF-KL e sua frequência.

Tabela 2: Tópicos encontrados pelo NMF-KL

Tópico	Descrição	Principais palavras-chave	%Mensagens Únicas
3	Links com recompensa	link, bônus, recompensas, VIP	24,61%
1	Melhores horários para jogar	horário, slots, fortune, PG	14,18%
7	Grandes prêmios para os vencedores	fortune, vencedor, prêmio, sucesso	13,29%
2	Código de resgate de bônus	resgate, código, login, oficial	11,15%
5	Lançamento de plataforma	pix, cadastro, lançamento, link	10,05%
6	Facilidades de depósito, saque e valor de taxas	saque, depósito, mínimo, taxa	7,70%
10	Resgate de bônus por atividades diárias	resgate, código, login, atividades	6,54%
8	Baixe o app para jogar	dinheiro, app, plataforma, bônus	5,35%
4	Vantagens exclusivas com cadastro na plataforma	membro, registre, exclusivo, contato	3,88%
9	Convide amigos e receba uma comissão	ganhar, comissão, amigos, prêmio	3,26%

Os tópicos revelam que a principal estratégia de divulgação consiste na oferta de benefícios proibidos para operadores licenciados. Os temas mais frequentes envolvem bônus, códigos de resgate, recompensas por atividades e promoções de lançamento, indicando que incentivos financeiros são amplamente utilizados para atrair usuários. Também são recorrentes mensagens que prometem maiores chances de ganho, supostos horários mais favoráveis para apostar que incluem até falhas nas plataformas que permitiriam lucros acima da média. Além disso, os canais enfatizam vantagens competitivas, como cashback, taxas reduzidas, pagamentos por cartão de crédito e programas de indicação com comissões por novos cadastros. Também são divulgados aplicativos das próprias plataformas e sorteios destinados a reforçar a percepção de legitimidade e estimular o engajamento dos usuários.

Em resumo, o que encontramos é que os canais vendem como promessas de vantagens desleais em comparação com o mercado legal, aproveitando a capacidade de utilizar funcionalidades proibidas pelas empresas legalizadas.

4.3 Similaridade dos Domínios

Os resultados anteriores indicam que os canais do Telegram são utilizados para divulgar plataformas de apostas irregulares. As mensagens coletadas permitem, portanto, identificar as URLs associadas a essas plataformas. Para isso, extraímos os links indicados nos metadados das mensagens por meio dos campos `MessageEntityUrl` e `MessageEntityTypeUrl`. A partir dos domínios identificados, analisamos a similaridade entre as plataformas com base em diferentes características de conteúdo e infraestrutura. Os agrupamentos resultantes fornecem diferentes níveis de evidência sobre possíveis relações entre os domínios e são utilizados para investigar a segunda questão de pesquisa (QP2), referente à possível coordenação entre plataformas de apostas irregulares.

A similaridade que demonstra com mais força uma coordenação é a presença de IDs iguais de serviços de rastreamento de anúncios como o Pixel da Meta. Para dois sites terem esse mesmo ID, é necessário saber as informações da conta atrelada, e portanto as páginas seriam feitas pelos mesmos indivíduos. Por outro lado, sites que compartilham exatamente os mesmos arquivos de JavaScript ou CSS podem indicar o mesmo desenvolvedor, porém a similaridade também pode ser causada pelo uso de um mesmo template ou framework de desenvolvimento web. Ainda mais circunstancial é a similaridade por informações de hospedagem e registro, como os certificados SSL/TLS ou o IP do servidor de hospedagem. Como os serviços de hospedagem podem agrupar muitos sites em um mesmo servidor, a correspondência de IP entre sites tem chance de ser uma coincidência. Por fim, a similaridade do conteúdo textual ou da estrutura do site pode ser causada meramente pelo fato de que se tratarem de *bets*, podendo se tratar de duas *bets* irregulares distintas buscando imitar um mesmo site de aposta legalizado.

Ainda assim, mesmo com a possibilidade de serem coincidências, as diversas semelhanças encontradas podem ser

combinadas e produzirem conjuntos de sites muito semelhantes. A figura 3 apresenta duas *bets* cujos arquivos CSS, arquivos JS, estrutura HTML e conteúdo são semelhantes. Nota-se que são basicamente clones um do outro, apenas com alguns dígitos do domínio e algumas cores alteradas.



Figura 3: Exemplo de bets com o mesmo hash de CSS

Para cada técnica de similaridade, construímos um grafo no qual cada vértice representa um domínio e uma aresta indica uma correspondência segundo a respectiva métrica. Além dos grafos individuais, construímos dois grafos agregados: um considerando todas as métricas e outro excluindo as duas métricas menos precisas, o Hash Estrutural e o Hash de Conteúdo. A Tabela 3 apresenta métricas relevantes dos grafos obtidos para cada técnica, enquanto a Figura 4 ilustra o grafo agregado que considera todas as métricas, exceto Hash Estrutural e Hash de Conteúdo.

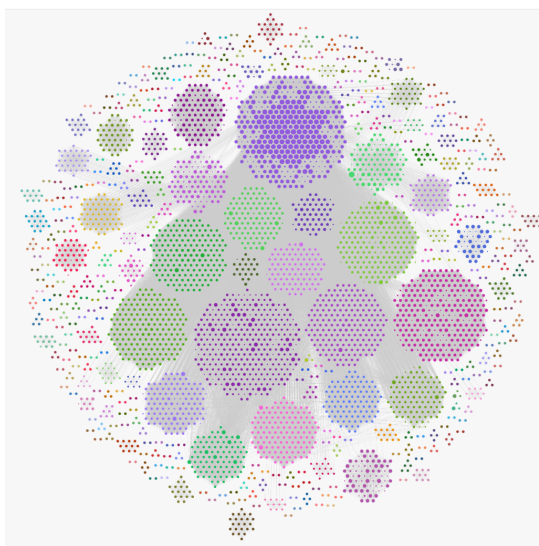


Figura 4: Grafo completo de todas as métricas de similaridade exceto estrutura e conteúdo

Tabela 3: Métricas dos grafos dos domínios

Técnica	#Nós	#Arestas	Grau Médio	#Comp. Conect.	Maior Comp.	% Maior Comp.	Coef. Agrup.	Coef. Assor.	Densidade
<i>Hash estrutural</i>	5.029	1.920.178	764	163	1.384	27,52%	0,9797	1	0,1519
<i>Hash de conteúdo</i>	4.250	1.139.724	536	190	1.435	33,76%	0,9722	0,9999	0,1262
<i>IP do servidor</i>	3.890	129.660	67	605	197	5,06%	0,8093	0,9965	0,0171
<i>Hash de pacotes JS</i>	1.212	71.313	118	50	401	33,09%	0,9485	0,9893	0,0972
<i>Hash de arquivos CSS</i>	1.071	56.001	105	86	269	25,12%	0,9233	0,9984	0,0977
<i>Certificado SSL/TLS</i>	400	771	4	134	16	4,00%	0,5150	0,9999	0,0097
<i>IDs de rastreamento de anúncios</i>	193	1.568	16	29	40	20,72%	0,8477	0,9999	0,0846
<i>Grafo completo exceto estrutura e conteúdo</i>	4.440	200.780	90	375	3.030	68,24%	0,8324	0,6290	0,0204
<i>Grafo completo</i>	5.411	2.266.459	838	63	4.764	88,04%	0,8973	0,7100	0,1548

Dos 6.523 domínios identificados, 5.411 apresentaram alguma similaridade no grafo completo. Como esperado, métricas menos restritivas, como similaridade de conteúdo e estrutura HTML, conectam um número maior de domínios, enquanto métricas mais conservadoras, como certificados SSL/TLS e IDs de rastreamento de anúncios, produzem agrupamentos menores (Tabela 3).

As métricas estruturais também revelam diferenças na conectividade dos grafos. Enquanto hashes de HTML, conteúdo, JS e CSS formam componentes maiores, a similaridade baseada em IPs gera mais componentes conectados devido à distribuição dos sites entre diferentes provedores. Em particular, os hashes de JS e CSS apresentam elevado grau médio, indicando o uso recorrente de templates ou código compartilhado entre plataformas, como ilustrado na Figura 3.

O maior componente baseado em IDs de rastreamento de anúncios reúne 40 domínios. Embora menor que os componentes obtidos por métricas de conteúdo, essa representa a evidência mais forte de coordenação, pois o compartilhamento desses identificadores pressupõe acesso à mesma conta de rastreamento.

Em conjunto, os resultados indicam que parte do ecossistema de *bets* irregulares não é composta por sites independentes, mas por grupos de domínios administrados de forma coordenada.

Tendo analisado as métricas individualmente, também avaliamos uma combinação de todas as similaridades. Considerando todas as métricas, 5.411 domínios (82,95%) foram conectados, formando um maior componente com 4.764 sites. Entretanto, como similaridade estrutural e de conteúdo podem conectar plataformas distintas apenas por compartilharem características comuns de sites de apostas, também construímos um grafo excluindo essas duas métricas. Nesse caso, 3.030 domínios permaneceram conectados, distribuídos em 375 componentes. A Figura 4 apresenta esse grafo, com os nós coloridos segundo as comunidades detectadas pelo algoritmo de Louvain [4].

Os coeficientes de agrupamento (0,8324 e 0,8973) e de assortatividade (0,6290 e 0,7100) indicam elevada sobreposição entre as diferentes métricas de similaridade, sugerindo que os agrupamentos não são resultado de evidências isoladas, mas da concordância entre múltiplos atributos. Por outro lado, a densidade do grafo reduz de 15,48% para 2,04% quando similaridade estrutural e de conteúdo são removidas, indicando

que essas métricas introduzem um número significativo de conexões potencialmente espúrias.

Em conjunto, esses resultados reforçam a existência de evidências consistentes de coordenação entre parte das *bets* irregulares, mesmo quando consideradas apenas as métricas mais conservadoras.

4.4 Domínios das Bets Irregulares

Após agrupar os domínios por similaridade, aplicamos técnicas de caracterização de sites baseadas em atributos de domínio, geolocalização e detecção de ameaças, seguindo a abordagem de Couto et al. [6]. A Tabela 4 apresenta os atributos que se mostraram mais relevantes para caracterizar as *bets* irregulares.

Atributos de Domínio. Quanto aos atributos de domínio, ao todo, 7,37% dos domínios continham hífen e 73,10% continham algum dígito; apenas seis e 31 desses casos, respectivamente, correspondiam a subdomínios. Esse padrão sugere o uso recorrente de pequenas variações nos domínios como estratégia para contornar ações de fiscalização.

Tabela 4: Atributos coletados para análise dos domínios

Categoria	Tipo	Atributo	Descrição
Domínio	Bool	subdomain-hifen	Subdomínio contém hífen
		subdomain-digit	Subdomínio contém dígitos
	Número	domain-hifen	Domínio contém hífen
		domain-digit	Domínio contém dígitos
		tld-br-or-com	TLD é .br, .com ou .com.br
Geolocalização	Lista	whois-privacy	Whois privado
		domain-age	Idade (dias) do registro inicial
		domain-expiry	Dias até expiração
	String	domain-update	Dias desde a última alteração
		contact-persons	Pessoas listadas como Contatos
	Bool	contact-emails	Email dos Contatos
		owner-document	Documento do Proprietário
Malware	String	owner-name	Nome do Proprietário
		ip-in-brazil	IP aponta para o Brasil
		lat-ip	Coordenada da Latitude do IP
	Número	long-ip	Coordenada de Longitude do IP
		isp	Provedor de Hospedagem
	String	ip-cc	Código do País das coord. do IP
		city-ip	Cidade do datacenter do IP final
	Número	total_risks	Nº de ameaças detectadas
	Lista	status_vt	Status retornado dos Antivírus
		threat_details	Detalhes das ameaças detectadas

Por outro lado, 63,69% dos domínios utilizavam os domínios de nível superior (*Top-Level Domains* or TLD) “.br”, “.com” ou “.com.br”. Como operadoras autorizadas devem utilizar o domínio “.bet.br”, o TLD torna-se um indicativo

relevante de irregularidade. Em particular, 1.743 (26,72%) domínios combinavam a palavra *bet* com o TLD “.com”, padrão comum antes da regulamentação e que pode transmitir aparência de legitimidade. Outros 457 domínios (7%) utilizavam o TLD “.bet”, sem o sufixo “.br”.

Outro atributo relevante é a idade do registro dos domínios, bem como o tempo desde sua última atualização e até sua expiração (*domain-age*, *domain-update* e *domain-expiry*). A Figura 5 apresenta a distribuição desses atributos.

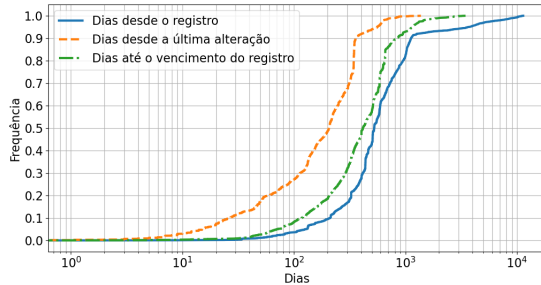


Figura 5: CDF dos atributos *domain-age*, *domain-update*, *domain-expiry*

A maioria dos domínios possui registros relativamente antigos: apenas 10% foram registrados nos 200 dias anteriores à coleta. Ainda assim, cerca de 80% dos registros ocorreram após a aprovação da Lei n.º 14.790/2023 (900 dias antes da coleta de dados), e praticamente todas as alterações cadastrais também foram realizadas nesse período. As poucas exceções correspondem, em sua maioria, a domínios originalmente registrados para outras finalidades e posteriormente reaproveitados como plataformas de apostas irregulares, como o domínio *zzzz.com*, registrado em 1995 e convertido em um site de apostas no final de 2023.

Por fim, para investigar possíveis evidências de coordenação entre essas plataformas (QP2), analisamos os campos *owner-name* e *whois-privacy*. Observamos que 77,57% dos sites utilizavam mecanismos de privacidade para ocultar as informações de seus proprietários. No entanto, 236 domínios ainda possuíam registros públicos, permitindo identificar 89 proprietários distintos, dos quais 39 estavam associados a mais de um domínio.

Atributos de Geolocalização. Os atributos de geolocalização e hospedagem permitem caracterizar a infraestrutura utilizada pelas *bets* irregulares. A partir dos atributos *ip-cc*, *city-ip*, *lat-ip* e *long-ip*, construímos a Figura 6, que apresenta a distribuição geográfica dos sites.

A maioria das *bets* irregulares estava hospedada no Canadá (2.961), Brasil (1.820) e Estados Unidos (570). Outros 221 domínios (3,39%) estavam distribuídos entre diferentes países e 952 (14,59%) não possuíam informações claras de localização. Essa distribuição é consistente com padrões observados em estudos sobre sites de baixa credibilidade [6]. Em relação aos provedores de hospedagem (*isp*), predominam Cloudflare

(2.758), Amazon (1.721), Google (242) e Microsoft Azure (204). Apesar disso, apenas 27,90% dos sites estavam hospedados no Brasil, indicando preferência por infraestrutura no exterior, possivelmente para dificultar ações de fiscalização e remoção. O predomínio de hospedagens no Canadá parece ter grande relação com o uso da Cloudflare.

Escaneamento de Vírus e Malwares Online. Por fim, para responder a terceira pergunta de pesquisa (RQ3), avaliamos os domínios quanto à presença de riscos de segurança como malwares. A API do serviço VirusTotal foi utilizada para medir todos os domínios, e os resultados estão apresentados na Tabela 5.

Tabela 5: Detalhes das ameaças detectadas nos Domínios Perigosos (Maliciosos+Suspeitos)

Status	Domínios	%Sem Ameaça	%Total
Inofensivo	5.896	99,48%	90,38%
Não Encontrado	31	12,97%	0,48%
Total Sem Ameaças	5.927	100,00%	90,86%
Ameaça	Domínios	%Perigoso	%Total
Phishing	239	40,10%	3,66%
Suspeito	135	22,65%	2,07%
Malicioso	132	22,15%	2,02%
Malware	46	7,72%	0,71%
Malware + Phishing	16	2,68%	0,25%
Não Recomendado	16	2,68%	0,25%
Spam	7	1,17%	0,11%
Malware + Spam	4	0,67%	0,06%
Phishing + Malware + Spam	1	0,17%	0,02%
Total Perigoso	596	100,00%	9,14%
Total	6.523	-	100,00%

Avaliando os resultados, observa-se que, para 90,38% dos sites, a maioria dos mecanismos do VirusTotal os classificou como *Inofensivos*. Além disso, 31 sites não receberam resposta de nenhum dos 91 mecanismos, possivelmente por não estarem mais disponíveis. Ao todo, 596 domínios receberam pelo menos um alerta de ameaça, com 158 tendo mais de três alertas. Apenas 28 dos 91 mecanismos retornaram alguma classificação diferente de *Inofensivo* ou *Não Detectado*, com uma média de 2,55 mecanismos sinalizando ameaça por domínio identificado. Entre as classificações, *Phishing* foi a mais frequente, seguida por *Suspeito* e *Malicioso*. Após consolidar classificações sobrepostas, priorizando os rótulos mais específicos, *Phishing* esteve presente em 42,95% dos domínios classificados como perigosos, seguido por *Malware* (11,24%) e *Spam* (2,01%).

Embora apenas 9,14% dos domínios tenha recebido algum alerta de segurança, eles foram divulgados em mais de 85 mil mensagens dos canais analisados, indicando que conteúdos potencialmente maliciosos ainda podem alcançar um número expressivo de usuários. Mesmo se existirem possíveis falsos positivos, considerando apenas os 158 com mais de 3 alertas, temos, no mínimo, 2,42% de sites com risco.

5 CONCLUSÃO

Neste trabalho buscamos caracterizar os canais do Telegram sendo usados para propagandear e coordenar sites de apostas irregulares no contexto da legalização das *bets* no Brasil.

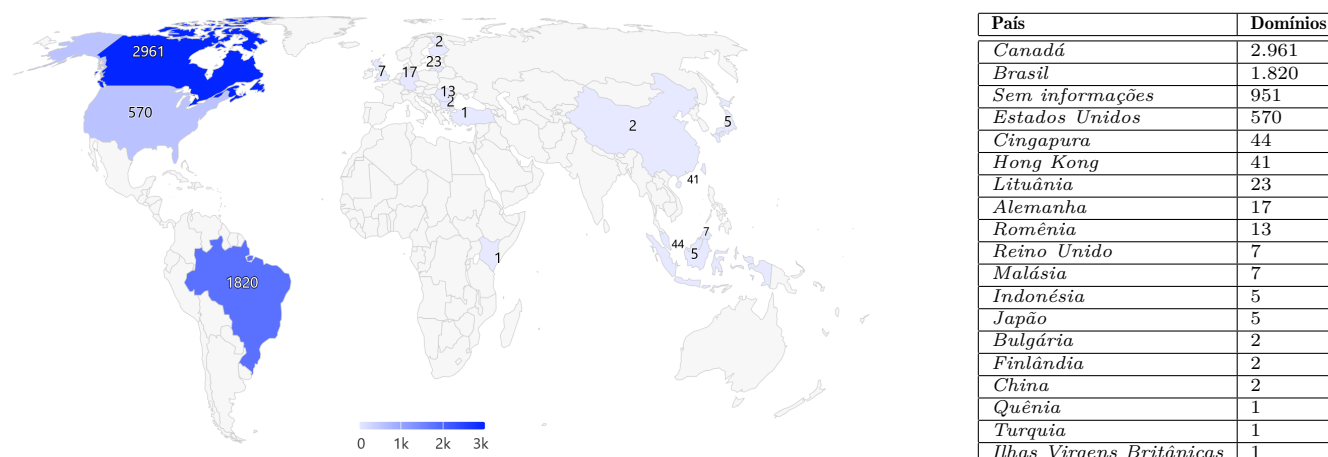


Figura 6: Localização dos IPs das bets irregulares e os vinte países com maior número de domínios hospedados.

Para isso, coletamos 125 canais públicos brasileiros obtidos a partir do site de agregação de links do Telegram chamado Telemetr.io e de pesquisas em mecanismos de busca. Após coletar todas as 1.447.707 mensagens enviadas pelos canais, foi possível extrair todos os links enviados, encontrando 6.523 domínios de *bets* irregulares.

Ao analisar o conteúdo das mensagens onde esses links foram encontrados através de um algoritmo de modelagem de tópicos, o NMF com a função objetivo da divergência de Kullback-Leibler, foi possível determinar quais as narrativas usadas nos canais para a propaganda dos sites de apostas irregulares. Entre os tópicos encontrados, destacam-se ofertas de bônus proibidos, facilidade de uso da plataforma com meios de pagamentos banidos, sugestões de horários e bugs para ganhos acima da média, e a possibilidade de receber comissões a partir do número de pessoas convidadas para o cadastro nas plataformas.

Em seguida, a partir dos domínios coletados realizamos o agrupamento dos sites pela similaridade de suas características. A partir disso, encontramos evidências para responder a segunda pergunta de pesquisa (QP2). Considerando as evidências de maior força, 193 *bets* tem os mesmos IDs de rastreamento de anúncios, reunidas em 29 agrupamentos. Ao expandir os grafos com múltiplas técnicas de similaridade, ao considerar todas as similaridades exceto o hash estrutural e de conteúdo, é possível revelar 375 conjuntos de páginas, com 3.030 sites de aposta no total, 68,24% dos domínios medidos pela técnica e 46,44% de todas as *bets* encontradas.

Por fim, analisamos os domínios a partir de uma série de atributos usados para avaliar sites de baixa credibilidade [6], buscando descrever os domínios e responder aspectos de todas as três perguntas de pesquisa. Quanto a primeira pergunta, as *bets* utilizam uma série de domínios similares, com dígitos ou um hífen adicionados, combinando a palavra bet com o TLD “.com” ou com o TLD “.bet” para transmitir legitimidade se espelhando no TLD “.bet.br” das *bets* legalizadas. 80% desses domínios foram registrados desde a legalização das

bets, em sua maioria hospedada no exterior, com apenas 27,90% hospedadas no Brasil.

Considerando a segunda pergunta sobre a coordenação entre as *bets*, os dados do registro dos domínios são privados para 77,57% dos sites, mas em 236 é possível encontrar nome atrelados, com 39 aparecendo no registro de múltiplas plataformas. Em relação a terceira pergunta sobre possíveis golpes, encontramos 596 domínios, 9,14% do total, que foram identificados como suspeitos ou maliciosos. Dentre eles, 42,95% foram detectados como Phishing, 11,24% como contendo Malware e 2,01% foram acusados de serem Spam.

Como trabalhos futuros, pretendemos expandir a análise de domínios de *bets* para outras fontes de dados, além de contrastar os resultados brasileiros com plataformas em países em situações distintas das etapas de legalização dos sites de apostas. Esperamos que esses resultados contribuam para uma maior compreensão do ambiente das *bets* irregulares, trazendo informações que permitam trabalhos subsequentes e possivelmente auxiliando no combate às plataformas irregulares.

6 LIMITAÇÕES

Nossa estratégia de descoberta de canais, baseada no Telemetr.io e em mecanismos de busca, pode introduzir vieses de cobertura. Embora tenhamos avaliado outros agregadores, como WGrupos, TGStat e Telegrupos, e observado grande sobreposição nos canais e nas palavras-chave identificadas, canais não indexados ou que não se apresentam explicitamente como comunidades de apostas podem ter ficado fora da amostra. Também restringimos a coleta a canais públicos, excluindo canais privados por questões éticas relacionadas à coleta de dados em espaços de acesso restrito. Por fim, o estudo concentra-se no Telegram, embora o WhatsApp seja o aplicativo de mensagens mais utilizado no Brasil; essa limitação decorre da maior disponibilidade de dados históricos e da possibilidade de coleta em larga escala no Telegram.

7 CUIDADOS ÉTICOS

Esta pesquisa atende à Resolução CNS nº 510/2016. O estudo utilizou exclusivamente dados provenientes de fontes públicas, incluindo o website Telemetr.io, canais públicos do Telegram e informações disponibilizadas pelo serviço VirusTotal, sem interação direta com participantes humanos. Os dados foram utilizados exclusivamente para fins científicos e analisados de forma agregada. Em conformidade com a Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018), o artigo não divulga dados pessoais ou outros identificadores de usuários, apresentando apenas exemplos estritamente necessários para ilustrar os resultados.

AGRADECIMENTOS

Este trabalho foi financiado e apoiado pelo Instituto Kunumi, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial para Cibersegurança (INCT-IACiber), processo no. 408432/2024-1, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR), processo nº. 408490/2024-1, além de bolsas individuais do CNPq, Capes e Fapemig. Os modelos Gemini 2.5 Flash e o GPT 5.5 foram empregados para resumir e revisar o texto. As decisões metodológicas, a interpretação dos resultados e as conclusões do estudo são de responsabilidade dos autores.

REFERÊNCIAS

- [1] Bhupendra Acharya, Dario Lazzaro, Antonio Emanuele Cinà, and Thorsten Holz. 2025. Pirates of charity: Exploring donation-based abuses in social media platforms. In *Proceedings of the ACM on Web Conference 2025*. 3968–3981.
- [2] Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*. 13–22.
- [3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
- [4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
- [5] Jessica Costa, Geovana Oliveira, Guilherme Fonseca, Davi Reis, Giancarlo Oliveira Teixeira, Washington Cunha, Leonardo Rocha, and Carlos HG Ferreira. 2025. Characterizing YouTube's Role in Online Gambling Promotion: A Case Study of Fortune Tiger in Brazil. In *Proceedings of the 17th ACM Web Science Conference 2025*. 42–51.
- [6] João MM Couto, Julio CS Reis, and Fabrício Benevenuto. 2024. Can computer network attributes be useful for identifying low-credibility websites? a case study in brazil. *Social Network Analysis and Mining* 14, 1 (2024), 153.
- [7] João MM Couto, Julio CS Reis, Ítalo Cunha, Leandro Araújo, and Fabrício Benevenuto. 2022. Caracterizando websites de baixa credibilidade no brasil. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. 503–516.
- [8] Leslie Daigle. 2004. *WHOIS protocol specification*. Technical Report.
- [9] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).
- [10] Vincenzo Imperati, Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Francesco Sassi. 2025. The Conspiracy Money Machine: Uncovering Telegram's Conspiracy Channels and their Profit Model. In *34th USENIX Security Symposium (USENIX Security 25)*. USENIX Association, Berkeley, CA, USA, 5229–5246.
- [11] Daniel Pimentel Kansao, Philippe De Freitas Melo, and Fabrício Benevenuto. 2022. “Click Here to Join”: A Large-Scale Analysis of Topics Discussed by Brazilian Public Groups on WhatsApp. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*. 55–65.
- [12] Massimo La Morgia, Alessandro Mei, and Alberto Maria Mongardini. 2025. Tgdataset: Collecting and exploring the largest telegram channels dataset. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Vol. 1. Association for Computing Machinery, New York, NY, USA, 2325–2334.
- [13] Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Jie Wu. 2025. Pretending to be a VIP! Characterization and Detection of Fake and Clone Channels on Telegram. *ACM Transactions on the Web* 19, 2 (2025), 1–24.
- [14] LCA Consultoria Econômica. 2025. *Dimensionamento do Mercado Ilegal de Apostas no Brasil*. Technical Report. LCA Consultoria Econômica. https://ibjr.org.br/wp-content/uploads/2025/09/LCA_IBJR_Dimensionamento_Mercado_ilegal.pdf Estudo elaborado para o Instituto Brasileiro de Jogo Responsável (IBJR). Acesso em: 2 jun. 2026.
- [15] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *nature* 401, 6755 (1999), 788–791.
- [16] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on World Wide Web*. 141–150.
- [17] Alessandro Perlo, Giordano Paoletti, Nikhil Jha, Luca Vassio, Jussara Almeida, and Marco Mellia. 2025. Topic-wise Exploration of the Telegram Group-verse. In *Companion Proceedings of the ACM on Web Conference 2025*. 1792–1801.
- [18] 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 Tigri- nho no YouTube. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC, 366–375.
- [19] Nam Khanh Tran, Sergej Zerr, Kerstin Bischoff, Claudia Niederée, and Ralf Krestel. 2013. Topic cropping: Leveraging latent topics for the analysis of small corpora. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 297–308.
- [20] Tribunal de Contas da União. 2025. *Relatório de Auditoria: Fiscalização das Apostas de Quota Fixa*. Technical Report. Tribunal de Contas da União (TCU). <https://pesquisa.apps.tcu.gov.br/redireciona/acordao-completo/ACORDAO-COMPLETO-2761620> Acesso em: 2 jun. 2026.
- [21] Aatman Vaidya and Kiran Garimella. 2026. Quantifying the Illicit Ecosystem of Betting Apps in India. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 20. 2312–2329.