# Uso de Características Temporais e Semânticas para Detectar Eventos em Vídeos de Violência Urbana

Saul Sousa da Rocha
saul.rocha2001@ufpi.edu.com
Universidade Federal do Piauí

Carlos Henrique Vale e Silva
carlosvale@ufpi.edu.com
Universidade Federal do Piauí

Mateus José da Silva
mateusjssilva@ufpi.edu.br
Universidade Federal do Piauí

Jose Rodrigues Torres Neto
jtorres@ufpi.edu.br
Universidade Federal do Piauí

Carlos Henrique G. Ferreira
carlos@ufmg.edu.br
Universidade Federal de Ouro Preto

Glauber Dias Gonçalves
ggoncalves@ufpi.edu.br
Universidade Federal do Piauí

## ABSTRACT

Videos published on platforms such as YouTube play a central role in covering urban violence events, but the high lexical similarity between different occurrences and the temporal proximity of postings hinder the automatic identification of which videos refer to the same event. Previous approaches predominantly explore semantic characteristics extracted from video, audio, and metadata, often relying on supervised techniques that require high computational costs and extensive annotations, making them impractical for continuous large-scale monitoring. Addressing this problem is of great importance for applications such as integrating multimedia records in investigations, fact-checking, assessing the impact of events, and building reliable historical archives. This work proposes two unsupervised and complementary heuristics: one based on temporal characteristics and anchor entities extracted via named entity recognition, and another based on multiple semantic attributes integrated into a similarity graph. We evaluate these approaches on a novel dataset of more than 1,400 manually annotated videos, collected between 2019 and 2024, covering both low- and high-impact events. Results show that the temporal heuristic consistently outperforms both the semantic heuristic and GPT-4 as a baseline, achieving accuracy up to 0.90 and NMI up to 0.98, while the semantic heuristic performs better in sparse-event scenarios. We also find that including full transcripts brings no substantial gains, indicating that titles and descriptions already contain the most relevant information for the task. These findings reinforce the potential of simple, low-cost, domain-adapted solutions to outperform generic approaches in challenging video clustering scenarios.

## KEYWORDS

video clustering, urban violence, event detection, unsupervised learning, YouTube analysis.

## 1 INTRODUÇÃO

A violência urbana no Brasil e em vários países do mundo continua sendo um dos principais desafios à paz e ao desenvolvimento social [20]. De acordo com o *Global Peace Index* de 2025, o Brasil ocupa a 130ª posição entre 163 países avaliados, destacando-se negativamente por suas altas taxas de criminalidade e homicídios, especialmente em áreas urbanas densamente povoadas, ficando na 9ª posição da América do Sul [2]. Segundo o Anuário Brasileiro de Segurança Pública, em 2023 ocorreram 6.393 mortes por intervenção policial [1]. Esse cenário evidencia a magnitude do problema, bem como a urgência de soluções tecnológicas que contribuam para o diagnóstico, monitoramento e prevenção da violência em ambientes urbanos.

Paralelamente ao agravamento desses indicadores, observamos uma crescente digitalização das manifestações da violência [7]. A produção massiva de conteúdo em vídeo nas plataformas de mídia social, com destaque para o YouTube, consolidou esse formato como meio dominante para comunicação, entretenimento e disseminação de informações [32, 33]. A popularidade e a acessibilidade da plataforma a tornaram um repositório valioso de dados sobre eventos sociais e de segurança pública, incluindo confrontos armados, operações policiais, protestos e ações emergenciais [21]. A partir da análise desse conteúdo, torna-se possível investigar a frequência, a dinâmica e até a repercussão de eventos violentos com alto nível de granularidade [29].

No entanto, transformar esse ecossistema em uma fonte estruturada de dados confiáveis impõe desafios significativos. Um dos principais obstáculos está na identificação automática de vídeos que se referem ao mesmo evento real. Vídeos sobre uma mesma operação policial podem ser publicados por diferentes canais, com títulos ambíguos, descrições parciais e datas que não refletem necessariamente o momento do ocorrido. Essa ausência de padronização, combinada com a informalidade e o ruído nos textos, dificulta a associação entre conteúdos correlatos. A literatura especializada tem abordado esse problema com técnicas multimodais, que integram informações visuais, auditivas e textuais para a detecção de eventos [9, 13, 30]. Apesar de promissoras, essas soluções demandam elevado poder computacional e infraestrutura sofisticada, características que limitam sua aplicabilidade em contextos de larga escala ou com recursos escassos, como é comum em iniciativas governamentais, projetos acadêmicos e iniciativas de monitoramento civil [19]. Por outro lado, abordagens baseadas exclusivamente em texto, como o GraphTMT [27], demonstram que transcrições de vídeos podem conter informações suficientes para a identificação de tópicos relevantes, sem a necessidade de processar sinais de áudio ou imagem. Essa constatação abre caminho para alternativas mais leves e escaláveis, centradas no Processamento de Linguagem Natural (PLN), especialmente quando se considera a quantidade massiva de dados disponíveis em plataformas como o YouTube para além das transcrições, tais como título, descrição do vídeo e data de publicação.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

Contudo, a aplicação dessas técnicas em contextos específicos traz desafios adicionais. Há uma enorme variabilidade na forma como os vídeos são publicados, com títulos subjetivos, vocabulário informal e expressões regionais, além de descrições frequentemente imprecisas ou ausentes [4]. O uso de gírias, abreviações, siglas de instituições locais e variações ortográficas são comuns, o que compromete diretamente a robustez de modelos treinados em contextos mais homogêneos ou com base em textos jornalísticos tradicionais [5]. Até onde sabemos, não há registros na literatura de trabalhos que abordem especificamente esses aspectos do ecossistema brasileiro de mídias sociais, tampouco que analisem tais mecanismos considerando suas particularidades linguísticas e sociais.

Superar essas limitações ampliaria substancialmente o potencial de aplicação prática dessas técnicas, tornando possível a identificação automática de vídeos relacionados a um mesmo evento em diversos cenários de relevância social, tais como: (i) apoio à cobertura jornalística, permitindo descobrir múltiplas perspectivas de um evento em tempo real; (ii) suporte à atuação de órgãos de segurança pública e defesa civil na triagem de ocorrências; (iii) estudos sobre percepção social de operações policiais e violência urbana; (iv) curadoria de acervos digitais e repositórios históricos de eventos marcantes; (v) monitoramento de desinformação, permitindo detectar conteúdos redundantes ou manipulados; e (vi) desenvolvimento de sistemas de alerta precoce baseados na detecção de grupos de conteúdo emergente.

Neste trabalho, propomos heurísticas de agrupamento de vídeos no contexto de violência urbana baseadas em atributos extraídos por técnicas de Processamento de Linguagem Natural (PLN) e em características temporais da postagem dos conteúdos. Em particular, nossa abordagem explora o reconhecimento de entidades nomeadas (*Named Entity Recognition - NER*) [16], como nomes de localidades, órgãos de segurança e datas, presentes nos textos associados aos vídeos (títulos, descrições e transcrições) desse contexto. Especificamente, investigamos a eficiência de duas heurísticas com estratégias distintas: (i) a similaridade semântica expressa nos textos (títulos, descrições e transcrições), e (ii) a coesão temporal e contextual de vídeos publicados em torno de um mesmo intervalo de tempo. Dessa forma, nosso trabalho visa oferecer uma solução eficiente e pragmática para agrupar vídeos que se referem a um mesmo evento de violência urbana (e.g., operação policial ou crimes), mesmo quando publicados por fontes distintas e com variações narrativas.

Avaliamos as heurísticas propostas em vídeos coletados do Youtube no contexto de monitoramento de violência urbana em três níveis de dificuldade de agrupamento: (1) eventos esparsos de menor repercussão majoritariamente associados a um único vídeo; (2) eventos de grande repercussão associados a vários vídeos; e (3) a junção de ambos. Ao todo, coletamos mais de 1400 vídeos e rotulamos manualmente 841 eventos de violência urbana nesses vídeos. Como referência para avaliação, implementamos uma solução para o agrupamento via *prompt* do GPT-4.0 [10], que é uma das ferramentas de LLM (*Large Language Model*) mais práticas e avançadas para tarefas de PLN atualmente. Então utilizamos métricas de classificação adaptadas para agrupamento (i.e., acurácia, precisão, revocação e f1-score) e métricas específicas de agrupamento (NMI e AMI) para avaliação de desempenho.

Os resultados mostram que a heurística de agrupamento temporal supera consistentemente tanto a heurística de agrupamento semântico quanto o GPT-4.0 como referência, alcançando precisão de até 0,90 e NMI de até 0,98, enquanto a heurística semântica apresenta melhor desempenho em cenários de eventos esparsos. Também constatamos que a inclusão de transcrições completas não traz ganhos substanciais, indicando que títulos e descrições já contêm as informações mais relevantes para a tarefa. Essas descobertas reforçam o potencial de soluções simples, de baixo custo e adaptadas ao contexto de interesse para superar abordagens genéricas em cenários desafiadores de agrupamento de vídeos.

O restante do artigo está organizado da seguinte forma: A seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta as abordagens desenvolvidas, detalhando o fluxo de processamento de texto e agrupamento. A Seção 4 descreve o conjunto de dados, métricas de avaliação e cenários avaliados. Os resultados experimentais são discutidos na Seção 5. Por fim, a Seção 6 conclui o trabalho e apresenta direções para estudos futuros.

## 2 TRABALHOS RELACIONADOS

A detecção automática de eventos em vídeos tem sido objeto de interesse crescente, impulsionada pelo aumento exponencial de conteúdo gerado em plataformas como o YouTube e pela inviabilidade da anotação manual em larga escala [31]. Diversos estudos propõem abordagens não supervisionadas ou com baixo grau de supervisão como alternativas aos métodos tradicionais, buscando formas escaláveis de associar vídeos a eventos, temas ou atividades de interesse público [8]. Grande parte da literatura recente tem se concentrado em abordagens multimodais, que integram informações visuais, auditivas e textuais para caracterizar eventos em vídeos. Em [30] e [9], por exemplo, os autores propõem estratégias de detecção imediata (*zero-shot*) de eventos com base em consultas textuais e atributos extraídos de múltiplas modalidades, como imagem, som e legenda. Métodos como o de [23] buscam aprender representações compartilhadas para reconhecer atividades nunca vistas, utilizando *autoencoders* e mecanismos de atenção temporal. Essas abordagens, embora eficazes em termos de precisão, apresentam alto custo computacional e forte dependência de dados anotados e infraestrutura especializada, o que compromete sua adoção em contextos com recursos limitados. Em [13], algoritmos de *clustering* não supervisionados são aplicados em larga escala para descobrir padrões acústicos recorrentes em vídeos do YouTube, evidenciando o potencial do áudio para detecção de eventos, mas também reforçando a complexidade inerente ao processamento de sinais. Em contraste, nosso trabalho adota uma abordagem exclusivamente textual, mais leve e adequada para cenários com dados ruidosos ou informais.

Paralelamente, outras pesquisas têm explorado atributos linguísticos como principal fonte informacional para a organização de vídeos. O GraphTMT, proposto por [27], modela tópicos a partir de grafos de palavras extraídas de transcrições automáticas, superando modelos tradicionais como LDA e K-Means em ambientes com ruído textual. Técnicas de caracterização temática também têm sido aplicadas em cenários multimodais: [32] investigam padrões em vídeos sobre o Mar da China Meridional, revelando reuso de conteúdo e mudanças no foco temático ao longo do tempo. Já o Tube2Vec, proposto por [6], constrói representações vetoriais de canais do YouTube a partir de dados sociais e recomendações. Apesar de avançadas, essas abordagens não se concentram na identificação

de vídeos relacionados a um mesmo evento específico, mas sim em estruturas temáticas, canais ou padrões de publicação.

No contexto de extração semântica textual, o reconhecimento de entidades nomeadas (NER) tem se consolidado como ferramenta central para a organização de conteúdo não estruturado. Em [3], os autores propõem um pipeline de NER com enriquecimento contextual para lidar com comentários de vídeos no YouTube. A proposta combina modelos como BERT e Sentence-BERT com técnicas de agrupamento (BERTopic e *Agglomerative Clustering*) para identificar diferentes formas de referenciar uma mesma entidade, abordando inconsistências como grafias ou abreviações diferentes, com o objetivo de mapear tendências emergentes. No entanto, esse tipo de aplicação visa análises de escopo mais amplo, sem foco na detecção precisa de eventos discretos. Avanços em escalabilidade também têm sido apresentados, como o PaDeLLM-NER, que acelera a inferência de entidades em modelos de linguagem por meio de decodificação paralela, mantendo a qualidade da extração mesmo em grandes volumes de dados [17]. Uma revisão abrangente sobre os desafios atuais do NER, especialmente em contextos com poucos recursos, linguagem ruidosa e necessidade de adaptação semântica, é apresentada em [14], reforçando a relevância dessa tecnologia para contextos como o brasileiro. Outras iniciativas exploram formas assistidas de anotação. Em [26], um framework baseado em *crowdsourcing* combina análise visual, textual e comportamental para enriquecer metadados de vídeos HTML5, permitindo a anotação automática de eventos durante a reprodução. Ainda que inovador, o foco está na melhoria da navegação e acessibilidade dos vídeos, e não na identificação automática de vídeos que relatam o mesmo evento a partir de atributos semânticos.

Em síntese, embora haja um corpo significativo de trabalhos voltados à organização, descrição e caracterização de vídeos em larga escala, poucos abordam diretamente o desafio de identificar vídeos que se referem ao mesmo evento real com base exclusivamente em dados textuais. Ademais, nenhum desses estudos trata das particularidades do contexto brasileiro, marcado por linguagem informal, variações regionais e formas pouco padronizadas de relatar acontecimentos. Nosso trabalho busca preencher essa lacuna ao empregar técnicas de NER combinadas a heurísticas de agrupamento eficientes e não supervisionadas, capazes de capturar relações semânticas e temporais mesmo em ambientes informacionais desestruturados.

# 3 ABORDAGENS NÃO SUPERVISIONADAS PARA IDENTIFICAÇÃO DE EVENTOS

A identificação automática de vídeos referentes a um mesmo evento em plataformas como o YouTube é desafiadora devido à heterogeneidade linguística, à variação na menção de entidades e à inconsistência nas datas de publicação. Mesmo conteúdos sobre um único episódio de violência urbana ou operação policial podem divergir quanto ao vocabulário, nível de detalhe e estilo narrativo, refletindo a multiplicidade de fontes e a informalidade dos relatos. Esses fatores dificultam a associação entre vídeos correlatos e demandam abordagens capazes de capturar simultaneamente: (i) a similaridade semântica presente nos textos (títulos, descrições e transcrições); e (ii) a coesão temporal e contextual das publicações.

Propomos duas heurísticas não supervisionadas e complementares para o agrupamento de vídeos: a primeira baseada em semelhança textual global, combinando representações vetoriais densas e medidas de similaridade contextual; a segunda fundamentada na proximidade temporal, utilizando janelas de tempo iterativas para identificar eventos comuns. Ambas exploram atributos extraídos automaticamente por reconhecimento de entidades nomeadas (*Named Entity Recognition* — NER), viabilizando aplicação em larga escala sobre acervos informais e não curados, com baixo custo computacional em comparação a trabalhos anteriores (ver Seção 2).

Ambas as heurísticas compartilham um fluxo inicial de processamento textual que converte o conteúdo multimídia do YouTube em informações estruturadas para análise automática. São extraídos metadados públicos (título, descrição) e, quando disponíveis, transcrições automáticas geradas pela biblioteca Vosk[1] com o modelo *vosk-model-small-pt-0.3*. Em seguida, aplica-se o reconhecimento de entidades nomeadas (*Named Entity Recognition* — NER) diretamente sobre o texto original, preservando capitalização e acentuação, utilizando o modelo *pt_core_news_lg* da biblioteca spaCy[2]. Essa etapa identifica, de forma automática, menções a localidades, datas, instituições e outras entidades relevantes.

Após a extração das entidades, cada heurística aplica uma estratégia distinta: a primeira explora similaridades semânticas entre vídeos por meio de representações vetoriais. A segunda organiza eventos considerando janelas temporais e a recorrência de entidades. A seguir, detalhamos cada abordagem proposta.

## 3.1 Agrupamento Semântico

A primeira heurística proposta explora múltiplas dimensões de similaridade para agrupar vídeos de um mesmo evento, combinando informações semânticas, entidades nomeadas, sinais quantitativos e contexto temático. A ideia central é que, embora dois vídeos sobre o mesmo incidente possam divergir amplamente em vocabulário ou estrutura narrativa, eles tendem a compartilhar pelo menos algumas dessas pistas. Ao integrar atributos heterogêneos, reduzimos a dependência de coincidências literais e aumentamos a robustez frente à informalidade e fragmentação dos relatos.

O processo, resumido no Algoritmo 1, é dividido em três etapas: (i) extração e representação de atributos, (ii) cálculo de similaridades par-a-par, e (iii) agrupamento por grafo.

*Etapa 1 — Extração e Representação.* Partimos da concatenação do título, descrição e, quando disponível, transcrição do vídeo. Em seguida, aplicamos reconhecimento de entidades nomeadas (NER) diretamente sobre o texto original, preservando capitalização e acentuação, de modo a maximizar a acurácia na detecção de localidades, pessoas, datas e instituições. A escolha do modelo *pt_core_news_lg* da biblioteca spaCy[3] se deve à sua cobertura e desempenho reportados para o português. Além disso, extraímos dois sinais estruturados: (a) número de vítimas, identificado por expressões regulares que buscam termos como "vítima(s)" ou "morte(s)"; e (b) nomes de operações policiais, detectados por padrões do tipo "operação [nome]". Para capturar a semântica global, normalizamos levemente o texto (remoção de ruído e conversão para minúsculas) e geramos

[1] https://alphacephei.com/vosk/models
[2] https://spacy.io/models/pt
[3] https://spacy.io/models/pt

**Algorithm 1:** Heurística de Agrupamento Semântico por Similaridades Múltiplas.

**Input:** Conjunto de vídeos $\mathcal{V}$; para cada $v \in \mathcal{V}$: título, descrição, (opcional) transcrição. Hiperparâmetros: pesos $(w_e, w_l, w_v, w_o, w_t)$; limiar $\tau$.
**Output:** Partição $\mathcal{C}$ de $\mathcal{V}$ em grupos (eventos).

**Etapa 1 — Extração e Representação**
**foreach** $v \in \mathcal{V}$ **do**
- $x_v \leftarrow$ concatenar {título, descrição, (transcrição)}
- `// (1) NER direto no texto original: preserva capitalização/acentos`
- $\text{ENT}_v \leftarrow \text{NER}(x_v)$ `// locais, pessoas, datas, instituições`
- `// (2) Sinais quantitativos/nomes estruturados`
- $D_v \leftarrow \text{VictimCount}(x_v)$ `// regex em "vítima(s)", "morte(s)"`
- $N_v \leftarrow \text{OpNames}(x_v)$ `// padrão "operação [nome]"`
- `// (3) Normalização leve só para matching/tópicos`
- $\tilde{x}_v \leftarrow \text{Normalize}(x_v)$ `// minúsculas; remover ruído`
- `// (4) Sinais de semântica global e tema`
- $e_v \leftarrow \text{Embed}(\tilde{x}_v)$ `// Sentence-Transformers`
- $T_v \leftarrow \text{Topics}(\tilde{x}_v)$ `// BERTopic`

**Etapa 2 — Similaridade Par-a-Par**
**foreach** *par não ordenado* $(i, j)$, $i \neq j$ **do**
- `// Semântica global (paráfrases/sinônimos)`
- $s_e(i,j) \leftarrow \text{cosine}(e_i, e_j)$
- `// Entidades explícitas (local/pessoa) com Jaccard aproximado`
- $L_i, P_i \leftarrow$ filtrar $\text{ENT}_i$ por {local,pessoa}
- $L_j, P_j \leftarrow$ filtrar $\text{ENT}_j$ por {local,pessoa}
- $s_l(i,j) \leftarrow \text{FuzzyJaccard}(L_i \cup P_i,\ L_j \cup P_j)$
- `// Magnitude factual (contagem de vítimas)`
- $$s_v(i,j) \leftarrow \begin{cases} 1, & D_i = D_j \\ 0, & \{D_i, D_j\} = \{0, >0\} \\ 1 - \frac{|D_i - D_j|}{\max(D_i, D_j)}, & \text{c.c.} \end{cases}$$
- `// Identificadores oficiais (nome de operação)`
- $s_o(i,j) \leftarrow \mathbf{1}(\exists\, n \in N_i,\ m \in N_j : \text{FuzzyMatch}(n,m) \geq 0.8)$
- `// Contexto temático (coocorrência de termos)`
- $s_t(i,j) \leftarrow \text{Jaccard}(T_i, T_j)$
- `// Combinação ponderada das pistas`
- $S(i,j) \leftarrow w_e s_e + w_l s_l + w_v s_v + w_o s_o + w_t s_t$

**Etapa 3 — Construção do Grafo e Agrupamento**
Crie grafo $G = (\mathcal{V}, \mathcal{E})$ com nós $\mathcal{V}$ e arestas $\mathcal{E} = \{(i,j) \mid S(i,j) \geq \tau\}$
$\mathcal{C} \leftarrow$ componentes conectados de $G$ (via BFS/DFS)
**return** $\mathcal{C}$

*embeddings* densos com o modelo *paraphrase-multilingual-mpnet-base-v2* [24] da biblioteca Sentence-Transformers[4]. O uso de modelos multilíngues e recursos textuais genéricos torna a abordagem facilmente adaptável a outros domínios e idiomas com mínima reconfiguração. Por fim, extraímos tópicos latentes com BERTopic[5], capazes de revelar padrões temáticos não explícitos no vocabulário.

*Etapa 2 — Similaridade Par-a-Par.* Para cada par de vídeos, calculamos cinco métricas complementares:

(1) **Similaridade de *embedding* ($s_e$)**: mede proximidade no espaço semântico, sendo eficaz para capturar paráfrases e sinônimos mesmo com redação divergente.
(2) **Similaridade de entidades de localização e pessoa ($s_l$)**: obtida via variação do índice de Jaccard [12] com correspondência aproximada (*Difflib SequenceMatcher*[6]), útil para unir vídeos que compartilham cenário geográfico ou protagonistas.
(3) **Similaridade no número de vítimas ($s_v$)**: captura magnitude factual do evento, com similaridade máxima quando as contagens coincidem ou ambos omitem a informação.
(4) **Similaridade de nomes de operação ($s_o$)**: funciona como identificador quase-único; atribuímos similaridade 1,0 se ao menos um par de nomes extraídos tem correspondência ≥ 0,8.
(5) **Similaridade de tópicos latentes ($s_t$)**: medida pela sobreposição de tópicos extraídos com BERTopic, indicada pela similaridade de Jaccard, revelando afinidade temática mesmo sem entidades ou termos coincidentes.

Essas métricas são combinadas por média ponderada, última equação do algoritmo, onde pesos $(w_e, w_l, w_v, w_o, w_t)$ calibram a contribuição de cada atributo.

*Etapa 3 — Agrupamento por Grafo.* Construímos um grafo onde cada vídeo é um nó e há uma aresta entre dois nós se a similaridade final $S$ excede o limiar $\tau$. Os grupos são formados pelos componentes conectados desse grafo, obtidos via busca em largura (BFS). Esse modelo garante que conexões indiretas (*e.g.*, $A$ similar a $B$ e $B$ similar a $C$) resultem no agrupamento conjunto de $A$, $B$ e $C$. O valor de $\tau$ controla a coesão de forma que valores altos produzem grupos menores e mais homogêneos, enquanto valores baixos favorecem maior abrangência.

## 3.2 Agrupamento Temporal

A heurística temporal explora a proximidade no tempo como principal evidência de que dois vídeos tratam do mesmo evento, apoiando-se em entidades nomeadas para identificar e rastrear incidentes ao longo da linha do tempo. A motivação é que, em plataformas como o YouTube, repercussões de eventos de violência urbana e operações policiais concentram-se em um intervalo relativamente curto após a ocorrência. Assim, mesmo que o vocabulário varie entre vídeos, a coesão temporal e a recorrência de certas entidades, como nomes de localidades ou operações, podem servir como âncoras robustas para agrupamento.

O processo é dividido em três etapas principais, resumidas no Algoritmo 2:

*Etapa 1 — Extração de entidades e definição de janelas.* Para cada vídeo, extraímos entidades com NER (Seção 3), aplicando normalização com conversão para minúsculas, remoção de acentos e exclusão de caracteres especiais. Definimos então uma *janela temporal* centrada na data de publicação do vídeo, com tamanho $2w + 1$ dias, sendo $w$ um parâmetro fixo (e.g., $w = 15$ dias). Essa janela reflete a duração típica da repercussão de eventos desse tipo na plataforma.

*2 — Seleção da entidade candidata.* Dentro da janela do vídeo corrente, calculamos a frequência de cada entidade única extraída dos vídeos presentes nesse intervalo. A *entidade candidata* é definida como a mais frequente na janela *e* presente no vídeo analisado. Essa escolha se justifica por duas razões: (i) entidades frequentes no intervalo tendem a representar o evento central; e (ii) exigir que a entidade esteja no vídeo evita associações por coincidência temporal.

*3 — Associação a grupos existentes ou criação de novos grupos.* Se não existir um grupo com o nome da entidade candidata, criamos um novo grupo, atribuindo a ele o vídeo e registrando sua abrangência temporal inicial com base na janela do vídeo. Se já existir grupo(s)

[4]https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
[5]https://maartengr.github.io/BERTopic
[6]https://docs.python.org/3/library/difflib.html

**Algorithm 2:** Heurística de Agrupamento Temporal.

**Input:** Conjunto de vídeos $\mathcal{V}$; para cada $v \in \mathcal{V}$: data de publicação, título, descrição e (opcional) transcrição.
Parâmetro: meia-janela $w$ (em dias).
**Output:** Partição $\mathcal{C}$ de $\mathcal{V}$ em grupos (eventos).

**Etapa 1 — Extração e preparação de entidades**
**foreach** $v \in \mathcal{V}$ **do**
  $x_v \leftarrow$ concatenar {título, descrição, (transcrição)}
  Aplicar normalização: minúsculas, remoção de acentos e caracteres especiais
  $\text{ENT}_v \leftarrow \text{NER}(x_v)$ `// locais, operações, datas`

**Etapa 2 — Iteração por ordem cronológica**
Ordenar $\mathcal{V}$ por data de publicação
**foreach** $v \in \mathcal{V}$ **do**
  $\mathcal{W}_v \leftarrow$ vídeos com data $\in [\text{data}(v) - w,\ \text{data}(v) + w]$
  cand $\leftarrow$ entidade mais frequente em $\mathcal{W}_v$ presente em $\text{ENT}_v$
  **if** cand *não é identificador de grupo existente* **then**
    Criar novo grupo $G$ com nome cand e abrangência $\mathcal{W}_v$
    Adicionar $v$ a $G$
  **else if** $v$ *dentro da abrangência de algum grupo* $G$ *com nome* cand **then**
    Adicionar $v$ a $G$
    Expandir abrangência de $G$ para incluir $\mathcal{W}_v$
  **else**
    Criar novo grupo $G'$ com nome cand (com sufixo, se necessário) e abrangência $\mathcal{W}_v$
    Adicionar $v$ a $G'$
**return** $\mathcal{C}$

com esse nome, verificamos se o vídeo está dentro da abrangência temporal de algum deles:

- Caso positivo, o vídeo é adicionado ao grupo correspondente, e a abrangência temporal desse grupo é atualizada para englobar todo o intervalo combinado — desde a data mais antiga até a mais recente entre seus vídeos. Esse ajuste dinâmico garante que a janela de um grupo possa se expandir progressivamente.
- Caso negativo, interpretamos como um evento distinto ou fase posterior, criando um novo grupo com o mesmo identificador (e.g., "-2", "-3" para fases subsequentes).

## 4 AVALIAÇÃO EXPERIMENTAL

Nesta seção, apresentamos a metodologia de avaliação da abordagem proposta, contemplando as heurísticas de agrupamento semântico e temporal. Inicialmente, descrevemos os conjuntos de dados de vídeos coletados da plataforma YouTube para a avaliação. Em seguida, apresentamos as métricas de avaliação e as configurações empregadas em cada heurística.

### 4.1 Bases de Dados

Utilizamos duas bases de dados distintas para avaliar a proposta, contemplando diferentes perfis de eventos de violência urbana reportados por usuários no YouTube. Optamos pelo YouTube como fonte devido à sua ampla popularidade e penetração no Brasil, bem como à disponibilidade gratuita de dados por meio da API do YouTube (versão 3).

A primeira base, denominada Eventos esparsos, representa um cenário de monitoramento contínuo da violência urbana com alta variabilidade de eventos. Nesse contexto, a coleta parte de buscas por palavras-chave genéricas como "vítimas", "morte", "roubo", "furto" e termos relacionados a "operação policial", de modo a identificar vídeos potencialmente relevantes. Dessa forma, a maior parte dos eventos capturados é de menor repercussão, geralmente com poucos vídeos, embora alguns casos de grande impacto social também estejam presentes.

**Tabela 1: Principais características dos conjuntos de dados.**

| Métrica | Ev. esparsos | Al. repercussão | Total |
|---|---|---|---|
| Total de vídeos | 960 | 465 | 1 425 |
| Total de eventos distintos | 834 | 7 | 841 |
| Ev. com apenas 1 vídeo | 769 | 0 | 769 |
| Eventos com $\geq 2$ vídeos | 65 | 7 | 72 |
| Vídeos em ev. múltiplos | 191 | 465 | 656 |
| Média de vídeos por ev. | 2,9 | 66,4 | 9,1 |

Para compor essa base, seguimos a metodologia de monitoramento descrita em [22], coletando vídeos relacionados a operações policiais no período de 2019 a 2023. As buscas semanais são realizadas de forma consecutiva, totalizando 104 unidades de tempo (semanas) ao longo dos anos cobertos. Cada busca é limitada a 20 resultados, conforme indicado por [22] como um bom compromisso entre relevância e diversidade de vídeos retornados pela API.

Após a coleta, aplicamos um filtro para reter apenas vídeos diretamente relacionados a operações policiais. Primeiramente, selecionamos vídeos classificados pela API na categoria "Notícias e Política", a fim de excluir blogs e documentários policiais. Em seguida, mantemos apenas aqueles cujo título contém termos como "pm", "pms", "policia", "policial", "policias", "policiais", "operacao" ou "operacoes", considerando textos transformados para minúsculas e com remoção de acentos e sinais ortográficos. Esse processo resulta em 960 vídeos, correspondentes a 834 eventos distintos — dos quais 65 possuem múltiplos vídeos (191 no total) e 769 são representados por apenas um vídeo.

A segunda base, denominada Alta repercussão, concentra-se em eventos de grande repercussão nas mídias sociais e na imprensa, contemplando exclusivamente casos reportados em múltiplos vídeos. Para sua construção, identificamos sete incidentes de violência urbana que resultaram em operações policiais de ampla divulgação no Brasil em 2024, por meio de busca manual em mídias sociais e veículos jornalísticos. Em seguida, utilizamos palavras-chave específicas de cada incidente para coletar vídeos via API do YouTube. Após curadoria manual, selecionamos 465 vídeos, todos pertencentes a eventos com múltiplos registros.

A Tabela 1 resume as principais características de ambas as bases. Observa-se que Eventos esparsos é majoritariamente composta por eventos com apenas um vídeo (média de 2,9 vídeos/evento), enquanto a Alta repercussão reúne eventos com elevado número de vídeos por evento (média de 66,4 vídeos/evento), configurando cenários de complexidade distinta para as heurísticas.

A Figura 1 mostra a distribuição da duração dos eventos em dias. Nota-se que a maioria concentra-se em apenas um dia, com queda acentuada após dois ou três dias. Menos de 5% das operações se estendem por mais de sete dias, sugerindo que eventos prolongados são exceção e possivelmente requerem atenção especial das heurísticas para evitar fusões indevidas com outros eventos temporalmente próximos.

A Figura 2 apresenta as vinte operações com maior número de vídeos. Observa-se um padrão concentrado, no qual poucos eventos concentram grande volume de material (por exemplo, a

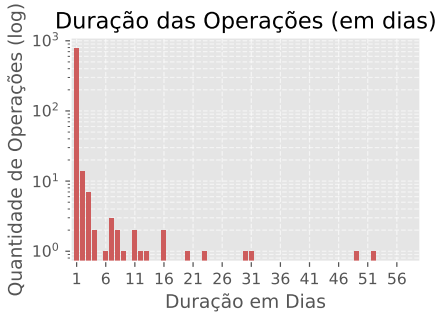


**Figura 1: Distribuição da duração dos eventos em dias.**

operação 890 com 225 vídeos), enquanto a maioria possui cobertura substancialmente menor. Essa distribuição em "degrau" implica que, nos eventos de maior repercussão, as heurísticas devem lidar com grande diversidade narrativa e de perspectivas sobre o mesmo fato; já nos demais, a tarefa exige identificar correspondências a partir de sinais mais sutis.

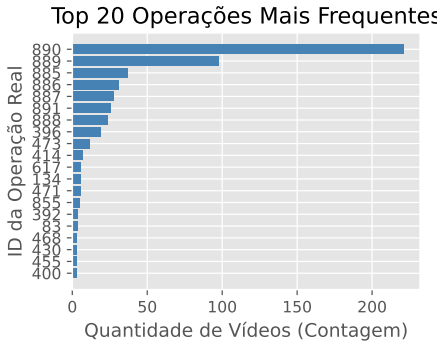


**Figura 2: Vinte operações com maior número de vídeos, ordenadas de forma decrescente.**

Para avaliar a eficácia das heurísticas de agrupamento, construímos um conjunto de referência com rótulos confiáveis para cada evento. Esse *ground truth* é obtido por três voluntários, que associam manualmente cada vídeo a um identificador numérico de evento, permitindo mapear múltiplos vídeos para um mesmo evento real. O processo de rotulação considera título, descrição, transcrição e data de postagem; quando necessário, realizamos a visualização completa do vídeo. Os voluntários seguem um protocolo padronizado: (i) leitura integral do título e descrição e, quando aplicável, visualização do vídeo; (ii) comparação com vídeos já rotulados, a fim de evitar duplicidades ou erros; e (iii) atribuição de um novo rótulo caso nenhuma correspondência seja encontrada. Esse procedimento resulta em um conjunto de dados consistente e confiável, adequado ao desenvolvimento e validação das heurísticas propostas.

A Tabela 2 apresenta cinco exemplos de vídeos rotulados manualmente, ilustrando a heterogeneidade de formatos e conteúdos. Há títulos objetivos e descritivos, bem como descrições incompletas ou pouco informativas. Tal cenário reforça a necessidade de heurísticas robustas, capazes de lidar com lacunas textuais e variações linguísticas na identificação de vídeos referentes ao mesmo evento.

## 4.2 Métricas e Configurações

Avaliamos o desempenho da abordagem proposta — heurísticas de agrupamento semântico e temporal — utilizando métricas de acurácia, precisão, revocação e F1-score adaptadas para avaliação de agrupamentos [18], bem como métricas específicas para comparação de partições: *normalized mutual information* (NMI) e *adjusted mutual information* (AMI) [25, 28]. Para referência de desempenho, aplicamos os mesmos agrupamentos com o modelo de linguagem GPT-4 [11] como linha de base. Os parâmetros das heurísticas seguem as definições da Seção 3.

As métricas de acurácia, precisão, revocação e F1-score foram adaptadas para o contexto de agrupamento, utilizando os rótulos manuais como referência e mapeando grupos previstos para grupos reais de forma ótima, a fim de maximizar o número de acertos. Para isso, empregamos o método Húngaro [15] sobre a matriz de confusão gerada pelos agrupamentos, obtendo a correspondência entre rótulos prevista e real. A partir desse mapeamento, a acurácia representa a fração de vídeos corretamente atribuídos; a precisão é a proporção de pares agrupados que realmente pertencem ao mesmo evento; a revocação mede a capacidade do método de recuperar todos os pares corretos; e o F1-score é a média harmônica entre precisão e revocação, equilibrando ambos. Tal abordagem, mapeamento ótimo seguido de métricas macro, é padrão em avaliação de algoritmos de clustering [18].

O NMI (*Normalized Mutual Information*) mede a quantidade de informação compartilhada entre os agrupamentos previstos e os rótulos reais, normalizada pela média das entropias de ambas as partições, assumindo valores entre 0 (nenhuma relação) e 1 (correspondência perfeita) [28]. Já o AMI (*Adjusted Mutual Information*) corrige o valor da informação mútua pela expectativa de acordo ao acaso, mantendo o intervalo $[-1, 1]$ e permitindo comparações mais justas [25]. Por essa razão, o AMI é menos sensível a flutuações aleatórias e ao número de grupos, penalizando situações como granularidade excessiva (muitos clusters minúsculos) ou superfragmentação de grupos grandes em vários clusters pequenos — aspectos observados em alguns cenários e discutidos na Seção 5.

No agrupamento semântico, os pesos atribuídos a cada atributo foram $w_e = 0{,}07$ para *embeddings*, $w_l = 0{,}39$ para entidades de localização e pessoa, $w_v = 0{,}25$ para número de vítimas, $w_o = 0{,}18$ para nomes de operações policiais e $w_t = 0{,}11$ para tópicos. A similaridade final (S) é a média ponderada desses atributos, e a formação de grupos ocorre conectando vídeos com $S \geq \tau = 0,55$, valor ajustado para privilegiar grupos menores e mais homogêneos. No agrupamento temporal, adotamos uma janela de $w = 15$ dias (total de 30 dias de abrangência), cobrindo 15 dias antes e depois da data de cada vídeo. Essa configuração foi definida a partir da análise exploratória das bases, que indicou que a maioria dos eventos de violência urbana repercute no YouTube por cerca de duas a três semanas.

Por fim, o GPT-4 foi avaliado a partir de um *prompt* estruturado, fornecendo a lista de vídeos em formato CSV com título, descrição, data e transcrição (quando disponível). O modelo foi instruído a

**Tabela 2: Exemplos de eventos rotulados manualmente.**

| Rótulo | Trecho do Título | Trecho da Descrição |
|---|---|---|
| 2 | "PM realiza operação em aglomerado contra tráfico..." | "Militares ocuparam durante toda a manhã um aglomerado ..." |
| 11 | "Número de mortos na operação policial..." | "O número de mortos na operação policial..." |
| 11 | "Operação policial em Jacarezinho deixa..." | "Uma operação policial realizada nesta quinta-feira..." |
| 11 | "#OABRJDebate a operação policial no Jacarezinho" | "Neste programa, nossos convidados avaliam a operação..." |
| 88 | "Operação policial no Complexo do Alemão..." | "Uma operação conjunta realizada entre..." |

usar o conteúdo textual e a proximidade temporal para atribuir rótulos numéricos sequenciais a cada evento, garantindo que vídeos do mesmo evento recebessem o mesmo rótulo. O objetivo foi testar a capacidade do LLM em realizar agrupamento com base em interpretação semântica e contextual, sem ajustes finos ou treinamento adicional.

Por fim, elaboramos um *prompt* fornecido ao modelo de linguagem GPT-4.0, com o objetivo de rotular vídeos do YouTube relacionados a operações policiais. O *prompt* especifica as entradas no formato CSV, delimitadas por ponto e vírgula, contendo as colunas título, descrição, data de postagem e transcrição (quando disponível). O modelo foi instruído a utilizar o conteúdo textual, complementado pela proximidade temporal, para determinar se diferentes vídeos correspondiam a um mesmo evento. Atribuir um identificador numérico sequencial, iniciando em 1, para cada evento distinto; e garantir que todos os vídeos de um mesmo evento recebessem o mesmo rótulo, retornando como saída um CSV contendo a coluna "evento" na mesma ordem das entradas originais.

*Prompt* Utilizado

Você é um assistente de IA especialista em identificar vídeos do YouTube sobre operações policiais.
Sua tarefa é ler um CSV (separador ';') com colunas: título, descrição, data de postagem e transcrição (quando disponível).
Considerar todos os vídeos, mesmo com campos ausentes.
Utilizar o conteúdo textual (título, descrição, transcrição) e a proximidade temporal.
Atribuir rótulo numérico único (iniciando em 1) para cada evento, garantindo que vídeos iguais tenham o mesmo rótulo. Retornar um CSV apenas com a coluna `evento` na ordem original.

As bases de dados rotuladas e o código-fonte para processamento estão disponíveis em repositório público.[7]

## 5 RESULTADOS

Nesta seção analisamos o desempenho das abordagens avaliadas — Heurística Temporal (HT), Heurística Semântica (HS) e GPT-4 — em seis variações de cenários, considerando a presença ou ausência de transcrições, consolidados na Tabela 3. Foram avaliadas as métricas acurácia (Acu), precisão (Pre), revocação (Rev), F1-score (F1), Normalized Mutual Information (NMI) e Adjusted Mutual Information (AMI). O melhor resultado em cada métrica dentro de cada subcenário está destacado em **negrito**.

Observa-se que, em praticamente todos os cenários, a **HT** apresentou desempenho superior, alcançando as maiores acurácias, F1-scores e medidas de informação mútua. No cenário **todos os eventos**, tanto com quanto sem transcrição, a HT atingiu **0.87** de acurácia e **0.84** de F1, superando a HS (0.57 Acu, 0.68 F1) e o GPT (0.49 Acu, 0.59 F1). Isso confirma a importância do critério temporal na identificação de eventos, dado que a alta similaridade lexical entre diferentes eventos relacionados à violência urbana e operações policiais compromete a eficácia de métodos puramente semânticos.

No cenário de **eventos esparsos**, a inclusão das transcrições também não alterou o padrão geral de desempenho. Contudo, HS mostrou um ganho expressivo nesse cenário (0.74 Acu, 0.80 F1) com e sem transcrição. Esse comportamento sugere que, em bases esparsas, atributos textuais extraídos de títulos e descrições são suficientes para melhorar a discriminação de eventos, enquanto o excesso de detalhes da transcrição não traz informações novas e discriminativas para agrupar eventos. Ainda assim, a HT manteve o melhor desempenho (até **0.90** Acu e **0.89** F1 sem transcrição), embora o AMI tenha caído para 0.47, refletindo a predominância de grupos unitários. Essa queda no AMI está relacionada à estrutura muito fina da partição (834 grupos reais × 884 previstos) na base de dados de eventos esparsos, onde predominam eventos únicos e há fragmentação de alguns grupos maiores. O ajuste por acaso do AMI penaliza mais nesses casos, pois até partições aleatórias produzem MI elevado, resultando em um valor ajustado baixo.

Já no cenário de **alta repercussão**, independentemente da presença de transcrições, a HT manteve **0.87** de acurácia e **0.84** de AMI, mas com quedas acentuadas em precisão e F1 (0.17 e 0.15, respectivamente). Isso indica que, apesar de preservar a coerência temporal dos agrupamentos, a sobreposição semântica e temporal entre eventos de grande cobertura midiática gera confusões residuais. A HS e o GPT apresentaram desempenho muito baixo (acurácia em torno de 0.25 e F1 praticamente nulo), evidenciando a limitação de métodos sem componente temporal explícito nesse contexto. Com a base de dados Alta repercussão, essa queda de precisão, revocação e F1 macro ocorre devido à superfragmentação de grupos grandes em muitos rótulos previstos. Por exemplo, um grupo com 221 vídeos foi dividido em 20 subgrupos, mas apenas um é mapeado corretamente via Hungarian. Os demais contam como erros, derrubando as médias macro. O AMI permanece alto (0.84) porque os grupos resultantes são internamente puros (*homogeneity* = 1.0) e o número de grupos reais é pequeno.

De forma geral, os resultados permitem concluir que: (i) a proximidade temporal é o discriminador mais eficaz para este contexto; (ii) atributos semânticos ganham relevância em bases esparsas, mas não sustentam desempenho em cenários densos; (iii) mesmo modelos de linguagem de última geração, como o GPT-4, não superam

[7] https://github.com/LABPAAD/urban_events_identification

**Tabela 3: Resultados consolidados em todos os subcenários (com e sem transcrição).**

| Cenário | Técnica | Acu | Pre | Rev | F1 | NMI | AMI |
|---|---|---|---|---|---|---|---|
| Todos eventos (com Transcr.) | HS | 0.57 | 0.69 | 0.74 | 0.68 | 0.86 | 0.22 |
| | **HT** | **0.87** | **0.85** | **0.84** | **0.84** | **0.97** | **0.86** |
| | GPT | 0.49 | 0.60 | 0.63 | 0.59 | 0.82 | 0.31 |
| Todos eventos (sem Transcr.) | HS | 0.57 | 0.69 | 0.74 | 0.68 | 0.86 | 0.20 |
| | **HT** | **0.87** | **0.85** | **0.84** | **0.84** | **0.97** | **0.86** |
| | GPT | 0.49 | 0.60 | 0.63 | 0.59 | 0.82 | 0.31 |
| Eventos esparsos (com Transcr.) | HS | 0.74 | 0.81 | 0.80 | 0.80 | 0.93 | 0.05 |
| | **HT** | **0.90** | **0.91** | **0.90** | **0.90** | **0.98** | **0.48** |
| | GPT | 0.59 | 0.58 | 0.62 | 0.58 | 0.90 | 0.10 |
| Eventos esparsos (sem Transcr.) | HS | 0.74 | 0.81 | 0.80 | 0.80 | 0.93 | 0.05 |
| | **HT** | **0.90** | **0.91** | **0.89** | **0.89** | **0.98** | **0.47** |
| | GPT | 0.59 | 0.58 | 0.62 | 0.58 | 0.90 | 0.10 |
| Alta repercussão (com Transcr.) | HS | 0.25 | 0.02 | 0.01 | 0.01 | 0.45 | 0.18 |
| | **HT** | **0.87** | **0.17** | **0.14** | **0.15** | **0.85** | **0.84** |
| | GPT | 0.24 | 0.03 | 0.00 | 0.01 | 0.35 | 0.13 |
| Alta repercussão (sem Transcr.) | HS | 0.25 | 0.02 | 0.01 | 0.01 | 0.45 | 0.18 |
| | **HT** | **0.87** | **0.17** | **0.14** | **0.15** | **0.85** | **0.84** |
| | GPT | 0.24 | 0.03 | 0.00 | 0.01 | 0.35 | 0.13 |

heurísticas simples, porém ajustadas ao contexto, quando o problema exige sensibilidade temporal e resiliência à alta sobreposição lexical; e (iv) métricas como o AMI e o F1 macro podem apresentar quedas expressivas não por baixa pureza dos agrupamentos, mas por efeitos estruturais, como granularidade excessiva e fragmentação de eventos em muitos grupos pequenos, o que eleva o linha de base aleatório de MI ou penaliza fortemente a média macro.

## 6 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho abordou o problema de identificação automática de vídeos que descrevem o mesmo evento de violência urbana no YouTube, um cenário desafiador pela alta similaridade lexical entre diferentes ocorrências e pela concentração temporal típica dessas coberturas. Nesse contexto, o desafio central está em distinguir eventos distintos que compartilham vocabulário, entidades e narrativas semelhantes, especialmente quando ocorrem em períodos temporais próximos ou se estendem por vários dias.

Para enfrentar esse desafio, foram propostas duas heurísticas não supervisionadas e complementares. A heurística temporal, baseada na coesão temporal dos vídeos e no uso de entidades mais frequentes como âncoras de agrupamento, e a heurística semântica, que combina múltiplos atributos textuais, incluindo embeddings, entidades nomeadas, número de vítimas, nomes de operações policiais e tópicos, em um grafo de similaridade. Também foi avaliada, como linha de base, uma abordagem baseada em GPT-4, a fim de verificar o desempenho de um modelo de linguagem de grande porte sem ajustes específicos para a tarefa.

O diferencial do estudo está em combinar pistas temporais e semânticas no contexto específico de violência urbana, um contexto em que a proximidade temporal dos fatos se mostrou mais eficaz do que a análise semântica textual isolada, mesmo quando enriquecida com transcrições completas. Além disso, foi apresentado um conjunto de dados inédito, abrangendo mais de 1.400 vídeos coletados ao longo de cinco anos e manualmente anotados para referência. Os resultados mostraram de forma consistente que a heurística temporal superou as demais abordagens em praticamente todos os cenários e métricas, atingindo acurácia próxima de 0.90 e NMI próximo de 0.98 nos melhores casos. A heurística semântica apresentou desempenho relevante em bases esparsas, enquanto o GPT-4, embora tenha obtido ganhos pontuais, não superou as heurísticas especializadas neste contexto. Outro achado relevante foi a constatação de que a inclusão de transcrições não trouxe ganhos significativos, indicando que títulos e descrições concentram a maior parte das informações úteis para o agrupamento.

Como trabalhos futuros, propõe-se a integração adaptativa das heurísticas temporal e semântica, ponderando dinamicamente a contribuição de cada uma conforme a densidade e a sobreposição lexical do conjunto de vídeos; a incorporação de atributos visuais, explorando descritores de imagem e vídeo como fatores adicionais de similaridade; o aprimoramento por meio de aprendizado supervisionado ou semi-supervisionado para otimizar a combinação de atributos; a aplicação e comparação em outros contextos e plataformas; análise temporal expandida, investigando técnicas mais avançadas de modelagem, como séries temporais multiatributo ou representações baseadas em redes temporais; e comparar as heurísticas propostas com técnicas clássicas de agrupamento, como K-Means e HDBSCAN, para complementar a análise experimental.

Em síntese, o trabalho demonstrou que heurísticas simples, mas ajustadas ao contexto, podem superar abordagens genéricas de última geração quando a tarefa exige sensibilidade temporal e tolerância à alta sobreposição lexical. Os resultados obtidos indicam caminhos promissores para soluções mais robustas, híbridas e multimodais voltadas à detecção automática de eventos em grandes acervos de vídeos.

## AGRADECIMENTOS

Pesquisa apoiada por FAPEPI (procs. 00110.000474/2023-17; 00110.00 0473/2023-64), PIBIC/UFPI, Projeto Ilíada executado pela RNP e CPQD no âmbito do MCTI PPI-Softex recursos da Lei 8248/1991, CNPq, FAPEMIG, CAPES e INCT-TILD-IAR.

# REFERÊNCIAS


[1] 2025. Atlas da Violência 2025. https://www.ipea.gov.br/atlasviolencia Acesso em: julho de 2025.

[2] 2025. Global Peace Index 2025: Measuring Peace in a Complex World. https://www.visionofhumanity.org Acesso em: julho de 2025.

[3] Ziyad Amer and Michelle D Davies. 2025. Context-Enriched Named Entity Recognition (NER) for Identifying Emerging Trends in Video Comments. *University of California, Berkeley* (2025).

[4] Murali Raghu Babu Balusu, Taha Merghani, and Jacob Eisenstein. 2018. Stylistic Variation in Social Media Part-of-Speech Tagging. arXiv:1804.07331 [cs.CL] https://arxiv.org/abs/1804.07331

[5] Thales Felipe Costa Bertaglia and Maria das Graças Volpe Nunes. 2017. Exploring word embeddings for unsupervised textual user-generated content normalization. *arXiv preprint arXiv:1704.02963* (2017).

[6] Léopaul Boesinger, Manoel Horta Ribeiro, Veniamin Veselovsky, and Robert West. 2024. Tube2Vec: Social and Semantic Embeddings of YouTube Channels. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 2084–2090.

[7] Celia Chen, Scotty Beland, Ingo Burghardt, Jill Byczek, William J. Conway, Eric Cotugno, Sadaf Davre, Megan Fletcher, Rajesh Kumar Gnanasekaran, Kristin Hamilton, Jordan Heustis, Tanaya Jha, Emily Klein, Hayden Kramer, Alex Leitch, Jessica Perkins, Casi Sherman, Celia Sterrn, Logan Stevens, Rebecca Zarrella, and Jennifer Golbeck. 2025. Cross-Platform Violence Detection on Social Media: A Dataset and Analysis. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25)*. Association for Computing Machinery, New York, NY, USA, 494–498. https://doi.org/10.1145/3717867.3717877

[8] Bruno Degardin and Hugo Proença. 2021. Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognition Letters* 145 (2021), 50–57.

[9] Mohamed Elhoseiny, Jingen Liu, Hui Cheng, Harpreet Sawhney, and Ahmed Elgammal. 2016. Zero-shot event detection by multimodal distributional semantic embedding of videos. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[10] Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Hongxia Ma, Li Zhang, Boxing Chen, Hao Yang, et al. 2024. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. *arXiv preprint arXiv:2402.18191* (2024).

[11] Yizheng Huang and Jimmy X. Huang. 2024. Exploring ChatGPT for next-generation information retrieval: Opportunities and challenges. *Web Intelligence* 22, 1 (2024), 31–44. https://doi.org/10.3233/WEB-230363 arXiv:https://journals.sagepub.com/doi/pdf/10.3233/WEB-230363

[12] Paul Jaccard. 1912. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE. *New Phytologist* 11, 2 (1912), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x arXiv:https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x

[13] Aren Jansen, Jort F Gemmeke, Daniel PW Ellis, Xiaofeng Liu, Wade Lawrence, and Dylan Freedman. 2017. Large-scale audio event discovery in one million youtube videos. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 786–790.

[14] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825* (2024).

[15] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

[16] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering* 34, 1 (Jan. 2022), 50–70. https://doi.org/10.1109/tkde.2020.2981314

[17] Jinghui Lu, Ziwei Yang, Yanjie Wang, Xuejing Liu, and Can Huang. 2024. Padellmner: Parallel decoding in large language models for named entity recognition. arXiv e-prints, pages arXiv–2402. (2024).

[18] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.

[19] Manuel Mondal, Mourad Khayati, Hông-Ân Sandlin, and Philippe Cudré-Mauroux. 2025. A survey of multimodal event detection based on data fusion. *The VLDB Journal* 34, 1 (2025), 9.

[20] LUCAS M. NOVAES. 2024. The Violence of Law-and-Order Politics: The Case of Law Enforcement Candidates in Brazil. *American Political Science Review* 118, 1 (2024), 1–20. https://doi.org/10.1017/S0003055423000540

[21] Raphael Ottoni, Evandro Cunha, Gabriel Magno, Pedro Bernardina, Wagner Meira Jr, and Virgílio Almeida. 2018. Analyzing right-wing youtube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM conference on web science*. 323–332.

[22] Omitido para revisão as cegas. 2024. Omitido para revisão as cegas.. In *Omitido*.

[23] AJ Piergiovanni and Michael Ryoo. 2020. Learning multimodal representations for unseen activities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 517–526.

[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[25] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. 2016. Adjusting for Chance Clustering Comparison Measures. *Journal of Machine Learning Research* 17, 134 (2016), 1–32. http://jmlr.org/papers/v17/15-627.html

[26] Thomas Steiner, Ruben Verborgh, Rik Van de Walle, Michael Hausenblas, and Joaquim Gabarró Vallès. 2011. Crowdsourcing event detection in YouTube videos 58-67. In *DeRiVE 2011 Detection, Representation, and Exploitation of Events in the Semantic Web*. CEUR Workshop Proceedings, 58–67.

[27] Jason Thies, Lukas Stappen, Gerhard Hagerer, Björn W Schuller, and Georg Groh. 2021. GraphTMT: unsupervised graph-based topic modeling from video transcripts. In *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*. IEEE, 1–8.

[28] Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary?. In *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, Quebec, Canada) *(ICML '09)*. Association for Computing Machinery, New York, NY, USA, 1073–1080. https://doi.org/10.1145/1553374.1553511

[29] Andrew J Weaver, Asta Zelenkauskaite, and Lelia Samson. 2012. The (non) violent world of YouTube: Content trends in web video. *Journal of Communication* 62, 6 (2012), 1065–1083.

[30] Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. 2014. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2665–2672.

[31] Kanwal Yousaf and Tabassam Nawaz. 2022. A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos. *IEEE Access* 10 (2022), 16283–16298. https://doi.org/10.1109/ACCESS.2022.3147519

[32] Niloofar Yousefi, Mainuddin Shaik, and Nitin Agarwal. 2024. Characterizing multimedia information environment through multi-modal clustering of youtube videos. In *International Conference on Smart Multimedia*. Springer, 295–309.

[33] Érica Pereira, Philipe Melo, Manoel Júnior, Vitor Mafra, Julio Reis, and Fabrício Benevenuto. 2022. Analyzing YouTube Videos Shared on WhatsApp and Telegram Political Public Groups. In *Proceedings of the 28th Brazilian Symposium on Multimedia and the Web* (Curitiba). SBC, Porto Alegre, RS, Brasil, 29–38. https://sol.sbc.org.br/index.php/webmedia/article/view/22088