risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 16/06/2025
Aceitação/Acceptance: 12/10/2025

Processamento de Linguagem Natural Aplicado à Identificação de Padrões Semânticos em Relatos de Mulheres Vítimas de Violência Doméstica e Familiar

Sabrina S. Vasconcellos[1], Deborah Q. G. Foroni[1], Peterson A. Belan[1]

sabrina.vasconcellos@uni9.edu.br; belan@uni9.pro.br,

[1] Universidade Nove de Julho (UNINOVE), Rua Vergueiro, 235/249 – São Paulo/SP – Brasil

DOI: 10.17013/risti.60.74–95

Resumo: A violência doméstica e familiar (VDF) contra a mulher é um problema persistente e subnotificado. Com o avanço das mídias sociais, relatos espontâneos de vítimas tornaram-se fontes relevantes de dados, embora seu volume e complexidade exijam técnicas automatizadas de análise. Este estudo busca identificar tópicos emergentes em relatos não estruturados de mulheres vítimas de VDF publicados no *YouTube*, aplicando Processamento de Linguagem Natural (PLN), aprendizado de máquina não supervisionado e modelagem de tópicos. A metodologia incluiu coleta via *API* do *YouTube*, pré-processamento textual, geração de *embeddings*, redução de dimensionalidade (*PCA* e *UMAP*), clusterização (*K-means* e *HDBSCAN*) e extração de tópicos com *BERTopic*. O melhor desempenho foi obtido com a combinação *UMAP* + *HDBSCAN* sem *stopwords*, revelando temas como ciclo da violência, ameaças e apoio familiar. Os resultados evidenciam a viabilidade do PLN e do *BERTopic* para análise automatizada e apoio à formulação de políticas públicas e pesquisas sociais.

Palavras-chave: Processamento de linguagem natural; Topic modeling; BERTopic; Violência doméstica e familiar.

Natural Language Processing Applied to the Identification of Semantic Patterns in Reports of Women Victims of Domestic and Family Violence

Abstract: Domestic and family violence (DFV) against women remains a persistent and underreported problem. With the rise of social media, spontaneous victim reports have become valuable data sources, although their volume and complexity require automated analytical techniques. This study aims to identify emerging topics in unstructured reports from women victims of DFV published on YouTube, using Natural Language Processing (NLP), unsupervised machine learning, and topic modeling. The methodology included data collection via the YouTube API, text preprocessing, embedding generation, dimensionality reduction (PCA and UMAP), clustering (K-means and HDBSCAN), and topic extraction using BERTopic. The best performance was achieved with the UMAP + HDBSCAN combination without stopwords, revealing themes such as the cycle of violence, threats, and family

support. The results highlight the feasibility of NLP and BERTopic for automated analysis and their potential to support public policy development and social research.

Keywords: Natural language processing; Topic modeling; BERTopic; Domestic and family violence.

1. Introdução

A violência doméstica e familiar contra a mulher é um fenômeno global que persiste em diferentes sociedades, sendo um problema social e de saúde pública. No Brasil, a Lei nº 11.340/2006, conhecida como Lei Maria da Penha, define a VDF como qualquer ação ou omissão baseada no gênero que resulte em morte, lesão, sofrimento físico, sexual ou psicológico, bem como dano moral ou patrimonial (BRASIL, 2006).

Em escala mundial, a Organização Mundial da Saúde (OMS) estima que uma em cada três mulheres já sofreu violência de gênero ao longo da vida, sendo que aproximadamente 27% das mulheres entre 15 e 49 anos relatam terem sido vítimas de violência física e/ou sexual praticada por parceiros (WHO, 2021).

As mídias sociais têm se consolidado como espaços de auto relatos de violência, oferecendo material importante para análise. Embora estudos sobre auto relatos não sejam novos (ESTROFF et al., 1994; LEVENDOSKY et al., 2004), a expansão dessas plataformas, que reúnem cerca de 4,7 bilhões de usuários no mundo (KEIPOS, 2022), ampliou a geração de dados públicos (GIGLIETTO; ROSSI; BENNATO, 2012). Nesse cenário, técnicas computacionais, como o Aprendizado de Máquina (AM), permitem o reconhecimento automático de padrões nesses relatos. Segundo Samuel (1959), o AM confere aos computadores a capacidade de aprender sem programação explícita, favorecendo análises mais robustas e reproduzíveis, inclusive em casos de VDF (AL-GARADI et al., 2022).

O uso de AM e PLN tem crescido em diversas áreas, incluindo saúde, marketing e biologia (CHEN et al., 2018), ampliando para as áreas de ciências sociais (MULLAH; ZAINON, 2021; NI et al., 2020). Enquanto ramo da Inteligência Artificial, o PLN aplica princípios linguísticos e métodos computacionais para construir modelos capazes de analisar e interpretar textos de forma automatizada (RUSSELL; NORVIG, 2004). Entre suas aplicações, destacam-se os métodos de *topic modeling*, que organizam grandes coleções de documentos a partir da identificação de temas latentes (BLEI, 2012; EGGER; YU, 2022).

Tradicionalmente, os dados sobre VDF são obtidos por meio de registros da saúde ou da segurança pública, como boletins de ocorrência (PREVIATTI; MILANI, 2016). No entanto, existem limitações na coleta, profissionais não se sentem preparados (ACOSTA et al., 2017; DE SOUZA et al., 2018; SOUZA; MARIN; RODRIGUES, 2021), além de protocolos oficiais simplificar os relatos, priorizando classificações legais (BORBUREMA et al., 2017) e deixando de capturar nuances contextuais, as quais podem ser resgatadas em entrevistas mais detalhadas (SCHRAIBER et al., 2007).

Assim, diante da ampla geração de dados não estruturados nas mídias sociais como textos, vídeos, imagens e áudios (ELSAYED; ABDELWAHAB; AHDELKADER, 2019),

essas plataformas se apresentam como fontes complementares de pesquisa (BOYD; CRAWFORD, 2012; GHANI et al., 2019). O volume de informações ultrapassa a capacidade de processamento humano (MUSTAK et al., 2021), o que torna as técnicas de AM, PLN e *topic modeling* fundamentais para identificar estruturas semânticas e padrões em relatos públicos de mulheres vítimas de violência doméstica e familiar.

Dessa forma, o presente estudo tem como objetivo explorar o potencial das mídias sociais como fonte complementar de dados sobre violência doméstica e familiar contra a mulher, aplicando técnicas de Aprendizado de Máquina e PLN para identificar padrões, estruturas semânticas e temas recorrentes nos relatos públicos.

2. Referencial Teórico

2.1. Violência Doméstica e Familiar

A violência doméstica e familiar contra a mulher é um problema global e recorrente, intensificado durante a pandemia da SARS-CoV-2, quando fatores como isolamento social, estresse econômico e convivência elevaram os índices de agressões (VIERIA; GARCIA; MACIEL, 2020; FORNARI et al., 2021).

No Brasil, instrumentos legais como a Convenção de Belém do Pará (1994) e a Lei Maria da Penha (Lei nº 11.340/2006) estabeleceram avanços no enfrentamento da VDF, sendo a Lei Maria da Penha considerada uma das legislações mais abrangentes do mundo (BRASIL, 2006; LISBOA; ZUCCO, 2022). Entretanto, desafios persistem quanto à efetividade das políticas públicas e à destinação de recursos para sua implementação (LISBOA; ZUCCO, 2022).

Do ponto de vista psicológico, Walker (1979) os episódios de violência seguem um ciclo repetitivo de três fases, o aumento da tensão, ato de violência e arrependimento, representados na Figura 1. A compreensão desse ciclo ajuda identificar padrões de reincidência e subsidiar tanto políticas públicas quanto intervenções preventivas e de proteção às vítimas.

A primeira fase, chamada “aumento da tensão”, é marcada pela irritabilidade do agressor diante de situações banais, enquanto a vítima busca evitar conflitos, muitas vezes assumindo a culpa para não “provocá-lo”.

Na segunda fase, ocorre a “explosão”, caracterizada pela perda de controle do agressor e a materialização da violência física, psicológica, verbal, moral ou patrimonial, gerando na vítima sentimentos de medo, vergonha, solidão e dor, além de possíveis reações como afastamento, denúncia ou até ideação suicida.

Já a terceira fase, denominada “arrependimento», envolve a tentativa do agressor de reconciliação por meio de promessas de mudança, o que pode confundir a vítima e dificultar a ruptura do ciclo, especialmente em relações com filhos.

Esse processo cíclico contribui para a perpetuação da violência e para o silêncio das vítimas, motivado por vergonha, medo e constrangimento, fatores que resultam na subnotificação de casos e na discrepância entre registros oficiais e a real dimensão do problema (FERREIRA; MORAES, 2020).

Figura 1 – Ciclo da Violência Doméstica Contra a Mulher

2.2. Aprendizado de Máquina

Desde o surgimento dos computadores, pesquisadores têm buscado desenvolver sistemas capazes de aprender a partir de dados, o que deu origem ao campo do Aprendizado de Máquina (AM) (MICHALSKI; CARBONELL; MITCHELL, 1983). O conceito foi inicialmente proposto por Arthur Samuel (1959), que o definiu como a capacidade dos computadores de aprenderem sem serem explicitamente programados. Posteriormente, Mitchell (1997) apresentou uma definição formal, afirmando que um programa "aprende a partir de experiências e com respeito a uma tarefa T é medida de desempenho P, se o desempenho em T, medido por P, melhora com a experiência E".

O AM abrange três principais abordagens: supervisionado, não supervisionado e semi-supervisionado (SARAVANAN; SUJATHA, 2018). No aprendizado supervisionado, o modelo aprende a partir de pares de entrada e saída para realizar previsões sobre novos exemplos (RUSSELL; NORVIG, 2004; XIE et al., 2018). Já o aprendizado não supervisionado busca identificar padrões em dados não rotulados, agrupando informações de forma autônoma (ALLOGHANI et al., 2020; RUSSELL; NORVIG, 2004). Por fim, o aprendizado semi-supervisionado combina dados rotulados e não rotulados para aprimorar o desempenho em cenários com escassez de rótulos (LIU et al., 2015; VAN DER MAATEN; HINTON, 2008).

Neste estudo, foi adotado algoritmos de aprendizado não supervisionado aplicados a dados textuais, com o objetivo de identificar padrões semânticos e estruturas latentes por meio da combinação das técnicas *Uniform Manifold Approximation and Projection (UMAP)* e *Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)*.

A manipulação de dados textuais gera representações vetoriais de alta dimensionalidade, o que impõe desafios aos algoritmos de agrupamento e visualização. Para contornar esse problema, utilizam-se técnicas de redução de dimensionalidade, que simplificam o espaço de características mantendo a estrutura relevante dos dados.

Entre as técnicas existentes, o *UMAP* (McInnes et al., 2018) destaca-se por combinar fundamentos matemáticos da geometria riemanniana e da topologia algébrica, permitindo reduzir a dimensionalidade de *embeddings* de centenas de dimensões para espaços bidimensionais ou tridimensionais, sem comprometer a coerência semântica. Essa redução possibilita que algoritmos de clusterização, como o *HDBSCAN*, operem de maneira mais eficiente.

O *HDBSCAN* (Campello, Moulavi & Sander, 2013) é um algoritmo de aprendizado não supervisionado baseado em densidade que identifica regiões densas em conjuntos de dados sem a necessidade de definir previamente o número de clusters. Diferentemente do *K-means*, o *HDBSCAN* é capaz de lidar com ruídos, detectar *outliers* e extrair *clusters* hierárquicos estáveis, tornando-o ideal para a análise exploratória de dados textuais complexos.

A combinação das técnicas *UMAP* e *HDBSCAN* representa uma solução robusta e eficaz para lidar com dados textuais de alta dimensionalidade. O *UMAP* reduz o espaço vetorial preservando a estrutura semântica, enquanto o *HDBSCAN* identifica agrupamentos densos e significativos de forma autônoma. Essa integração metodológica resulta em modelos mais interpretáveis, coerentes e aplicáveis em múltiplos domínios, mineração de tópicos, detecção de padrões sociais e identificação de tendências comportamentais.

Assim, a adoção combinada dessas técnicas demonstra o potencial do aprendizado não supervisionado em descobrir relações latentes e revelar estruturas semânticas emergentes em grandes volumes de dados textuais, contribuindo para o avanço das aplicações de Processamento de Linguagem Natural em contextos de análise social e da segurança pública.

2.2.1. Clusterização

A clusterização é uma técnica de aprendizado não supervisionado que busca identificar estruturas ocultas em conjuntos de dados sem rótulos pré-definidos. Seu objetivo é agrupar elementos com alta similaridade dentro de um mesmo grupo e alta dissimilaridade entre grupos distintos (SINAGA; YANG, 2020). Além disso, representa um processo que visa organizar dados de modo a facilitar a interpretação e a compreensão humana (MCINNES; HEALY, 2017).

Os algoritmos de clusterização podem ser classificados em duas categorias: ordem conhecida e ordem desconhecida. No primeiro caso, o número de clusters é definido como parâmetro de entrada, enquanto no segundo, esse número é determinado automaticamente com base em critérios como densidade ou proximidade dos dados (MOUTON; FERREIRA; HELBERG, 2020).

2.2.2. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) constitui uma área interdisciplinar que integra conhecimentos da computação, da inteligência artificial e da linguística, voltada à compreensão e ao tratamento automático da linguagem humana (RESHAMWALA; MISHRA; PAWAR, 2013; KHURANA et al., 2022). Desde os anos 2000, técnicas como

representações vetoriais e modelos baseados em aprendizado profundo ampliaram significativamente as aplicações em PLN (KHURANA et al., 2022).

Com o crescimento dos grandes volumes de dados digitais, surgiu a necessidade de representações numéricas capazes de transformar textos em vetores compreensíveis para máquinas. O modelo de espaço vetorial permanece como uma das abordagens mais utilizadas para a representação de documentos, possibilitando a análise automatizada e a classificação de textos (BAFNA; PRAMOD; VAIDYA, 2016; SINOARA et al., 2019).

2.2.3. Topic Modeling

O *topic modeling* é uma técnica de aprendizado de máquina não supervisionado utilizada para extrair significado de textos, identificando temas latentes a partir das palavras-chave mais representativas em coleções de documentos não estruturados (RAO; TABOADA, 2021). Essa abordagem permite organizar e resumir grandes volumes de informações, tornando possível uma análise que seria inviável manualmente (BLEI, 2012).

Neste estudo, foi utilizado o algoritmo *BERTopic*, desenvolvido por Maarten Grootendorst (2022), que se baseia em arquiteturas *Transformer* e aproveita o modelo de linguagem *BERT* para gerar *embeddings,* ou seja, representações numéricas que capturam o significado semântico das palavras e frases em um espaço vetorial contínuo. Esses *embeddings* são posteriormente utilizados na clusterização de tópicos, permitindo identificar padrões semânticos em grandes conjuntos de textos.

Além de modelos baseados em embeddings (como BERT) e abordagens de modelagem de tópicos, observa-se o avanço do uso de modelos de linguagem de grande porte (LLMs) em tarefas de apoio à produção e estruturação textual. Uma revisão recente sobre o impacto do ChatGPT em processos criativos reporta benefícios como ganho de produtividade e auxílio na geração de diálogos e tramas, mas ressalta limitações relacionadas à profundidade emocional e à originalidade, além de desafios éticos e laborais associados ao uso acrítico dessas ferramentas. Assim, em contextos sensíveis, como narrativas de violência, recomenda-se que a IA seja empregada de forma complementar e com supervisão humana, de modo a preservar nuances semânticas e reduzir riscos interpretativos (ZALDÍVAR-COLADO; TRIPP-BARBA; AGUILAR-CALDERÓN, 2025).

Conforme Figura 2 – Fluxo do algoritmo *BERTopic*, o processo é dividido em três etapas principais. Na primeira, cada documento é convertido em sua representação vetorial (*embedding*) por meio de um modelo pré-treinado. Na segunda, ocorre a redução de dimensionalidade para otimizar a clusterização dos dados. Por fim, na terceira etapa, as representações de tópicos são extraídas com base em uma variação do *TF-IDF*, denominada *c-TF-IDF*, que identifica os termos mais relevantes em cada grupo (GROOTENDORST, 2022).

A flexibilidade do *BERTopic* permite substituir as técnicas utilizadas em cada etapa sem comprometer o desempenho do modelo. Essa característica o torna uma ferramenta eficiente e inovadora para a identificação e organização de tópicos emergentes em relatos de mulheres vítimas de violência doméstica e familiar, contribuindo para uma análise automatizada e semanticamente consistente dos dados coletados.

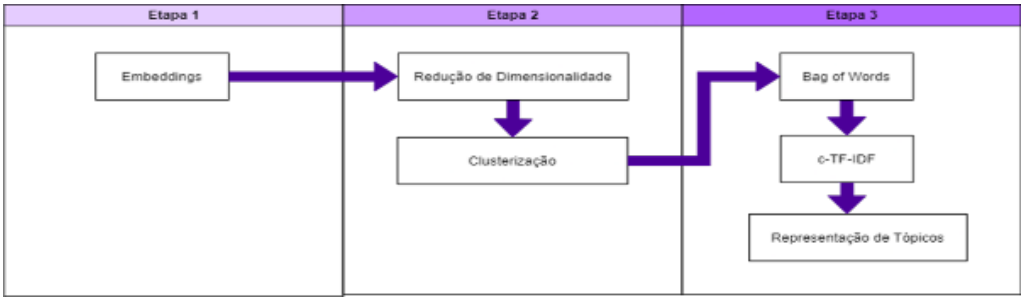

Figura 2 – Fluxo do Algoritmo *BERTopic*

2.3. Trabalhos Correlatos

Embora ainda seja um campo emergente, os estudos que utilizam PLN e IA no fenômeno da violência doméstica e familiar contra a mulher têm se expandido, sobretudo a partir de dados de mídias sociais, registros públicos e bases clínicas.

Na literatura internacional, destacam-se investigações que exploram o potencial do PLN na identificação de padrões de discurso associados à violência. Xue, Chen e Gelles (2019) analisaram 322.863 mensagens do *Twitter* entre outubro de 2015 e janeiro de 2016, aplicando o método *Latent Dirichlet Allocation (LDA)* para revelar tópicos latentes e estruturas temáticas sobre violência doméstica. O estudo demonstrou a eficácia do aprendizado de máquina não supervisionado na detecção de padrões semânticos em grandes volumes de texto.

Em uma linha similar, Soldevilla e Flores (2021) desenvolveram um classificador baseado no modelo *BERT* com *fine-tuning*, aplicado a mensagens do *Reddit* e *Twitter*, para distinguir conteúdos relacionados à violência de gênero. Utilizando o modelo *bert-base-uncased*, os autores obtiveram acurácia de 0,8909 e AUC de 0,9603, validando a aplicabilidade dos transformadores em contextos sensíveis e socialmente relevantes.

No contexto brasileiro, Corrêa e Faria (2021) utilizaram técnicas de mineração de texto, como nuvem de palavras, modelagem de tópicos (*LDA*) e classificação automática (*Naive Bayes*), sobre 125 relatos de violência extraídos de notícias em português, alcançando 71,2% de acurácia na classificação de casos como “constantes” ou “esporádicos”.

Ampliando essa vertente nacional, Campos e Andrêo (2024) propuseram a análise automatizada de notícias online por meio de *web crawling*, análise de sentimentos e redes neurais artificiais, com o objetivo de quantificar ocorrências de violência contra a mulher noticiadas na internet. O estudo apresentou o potencial das *RNA* como ferramenta para geração de indicadores sociais e apoio a políticas públicas de enfrentamento.

Em um recorte mais conceitual, Santos (2025) discute as possíveis aplicações da IA na segurança pública, destacando seu papel estratégico na redução dos índices de violência contra a mulher. Por meio de uma revisão qualitativa, o autor argumenta que a IA, quando estruturada de modo integrado a dados institucionais e sociais, pode fortalecer políticas de prevenção e ampliar a capacidade preditiva do Estado no monitoramento de casos de VDF.

No cenário global contemporâneo, observa-se o avanço do uso de modelos de linguagem de grande porte (*LLMs*) para a compreensão e apoio a vítimas de violência. Guan et al. (2025) desenvolveram um modelo *GPT*-3.5 ajustado para identificar necessidades de informação de sobreviventes de violência doméstica em comunidades de saúde online (*Reddit*). O modelo classificou oito tipos de necessidades, alcançando *F1-score* de 70,49% em textos reais e 84,58% em textos sintéticos, superando modelos como *Llama* 2-7B, *Llama* 3-8B e *LSTM*, o que evidencia a eficácia dos *LLMs* em classificação multicategórica contextualizada.

Na mesma direção, Kouzani e Nouman (2025) propuseram um modelo híbrido de aprendizado profundo combinando *BERT, BiLSTM* e *BiGRU* para detecção de indicadores de violência em textos digitais, utilizando 39.650 postagens do X (antigo *Twitter*). O modelo obteve performance preditiva, demonstrando capacidade de capturar dependências de longo prazo e contextos linguísticos complexos, o que reforça a utilidade de arquiteturas híbridas na análise semântica de violência online.

Em ambientes institucionais, Botelle et al. (2022) e Zeidan et al. (2022) aplicaram modelos *BioBERT* e análises de notas clínicas para identificar casos de violência interpessoal em registros eletrônicos de saúde, revelando o aumento da violência por parceiro íntimo (VPI) durante a pandemia de COVID-19. Esses estudos ampliam a compreensão da VDF como problema multidimensional, apontando a relevância do PLN aplicado a dados clínicos e sociais.

Em síntese, observa-se a consolidação de um ecossistema interdisciplinar que une dados textuais de mídias sociais, registros públicos e fontes médicas, evidenciando a eficácia da Inteligência Artificial e do Processamento de Linguagem Natural na detecção automatizada, análise semântica e previsão de padrões de violência. As abordagens revisadas demonstram não apenas o potencial técnico dos modelos de linguagem, mas também sua contribuição ética e social na formulação de políticas públicas e estratégias de intervenção preventiva.

3. Materiais e Métodos

Nesta seção, estão apresentados os materiais e métodos empregados na condução da pesquisa, com detalhamento das etapas dos experimentos computacionais.

3.1. Etapas Metodológicas

A Figura 3 apresenta o fluxo metodológico adotado, composto por três etapas principais: (1) Coleta de Dados, (2) Pré-processamento dos Dados e (3) Modelagem de Tópicos (*Topic Modeling*).

O processo inicial consistiu na definição de uma estratégia de busca capaz de extrair, com o mínimo de ruído, conteúdos autênticos sobre relatos de mulheres vítimas de violência doméstica e familiar (VDF). Foram combinados os termos “relato”, “violencia” e “mulher” sem acentuação para capturar narrativas públicas relevantes.

A escolha da plataforma *YouTube* foi considerada com base em três critérios:

a. quantidade de material disponível;

b. profundidade dos relatos, geralmente mais longos e descritivos;
c. disponibilidade de *API* pública para extração de dados.

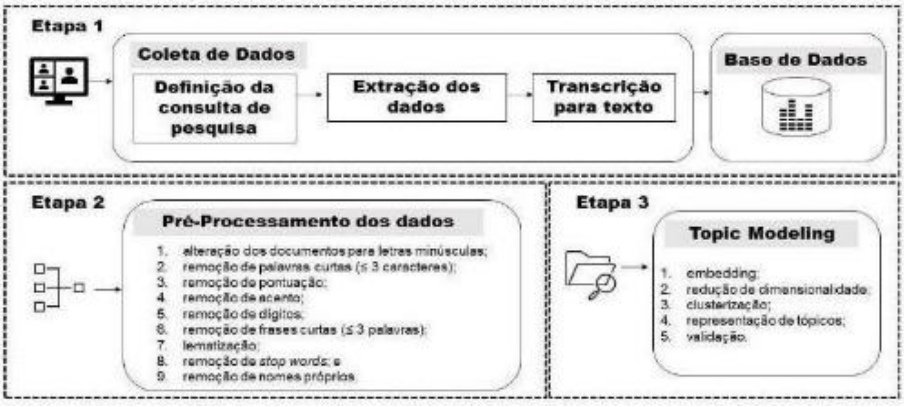

Figura 3 – Processo Metodológico e Experimental

A coleta foi realizada com o módulo *urllib.request* (*Python* 3.8), utilizando as palavras-chave "relato+violencia+mulher". Devido às limitações da *API*, que indexa apenas títulos e descrições, foi necessário complementar a busca com uma seleção manual de vídeos.

Em seguida, foi aplicado o modelo *Whisper* (Radford et al., 2022), desenvolvido pela *OpenAI*, para transcrever automaticamente o áudio dos vídeos em texto. O *Whisper* é um sistema de reconhecimento automático de fala (*Automatic Speech Recognition – ASR*) treinado em 680 mil horas de dados multilíngues, sendo gratuito e de código aberto.

As transcrições resultantes foram validadas manualmente para garantir precisão semântica e exclusão de partes irrelevantes como perguntas de repórteres. Apenas os relatos das vítimas, em português do Brasil, foram mantidos.

Com as transcrições validadas, foi construída uma base de dados estruturada em *Python*, contendo as colunas:

- ID do Vídeo (identificação única);
- Título do Vídeo;
- Transcrição completa; e
- Transcrição validada (apenas o relato da vítima).

Essa estrutura assegura rastreabilidade e controle de qualidade sobre o material analisado. Os trechos que não pertenciam aos relatos, como introduções jornalísticas ou comentários de terceiros, foram eliminados. Assim, cada registro da base representa um documento textual limpo, pronto para o processamento linguístico.

O pré-processamento é uma etapa essencial em análises de PLN, especialmente quando se lida com dados textuais não estruturados. O objetivo é normalizar e limpar os textos para reduzir ruído e aumentar a precisão das etapas posteriores de modelagem.

As operações aplicadas incluíram:

1. conversão para letras minúsculas;
2. remoção de palavras curtas (≤ 3 caracteres);
3. remoção de pontuação, acentos e dígitos;
4. exclusão de frases curtas (≤ 3 palavras);
5. lematização, conforme Lucca e Nunes (2002), para representar palavras em suas formas canônicas;
6. remoção de *stopwords* (Fox, 1989), incluindo termos genéricos e palavras específicas do fenômeno, "violência", "contra", "mulher" e "doméstica", que poderiam enviesar a análise;
7. exclusão de nomes próprios, visando à proteção da identidade das mulheres.

Essas transformações resultaram em um conjunto textual limpo, anonimizado e normalizado, pronto para as análises semânticas e estatísticas subsequentes.

Por fim, foi aplicada a etapa de modelagem de tópicos, voltada à identificação de temas latentes nos relatos coletados. O processo seguiu as fases metodológicas conforme demonstrado na Figura 4.

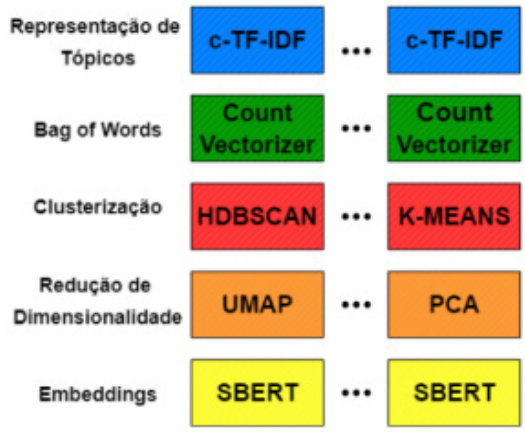

Figura 4 – Etapas para criar as representações de tópicos

Esses procedimentos permitiram agrupar automaticamente relatos com similaridades semânticas, possibilitando a extração de tópicos emergentes relacionados às experiências de VDF. O método mostrou-se adequado para identificar padrões discursivos, emoções recorrentes e aspectos sociais e psicológicos subjacentes aos relatos.

O processo metodológico adotado, representado na Figura 3, reflete uma estrutura de experimentação replicável e transparente, composta por três macro etapas integradas:

- Coleta e transcrição automatizada de dados autênticos (*YouTube*);
- Tratamento e limpeza textual com foco em qualidade linguística e ética; e
- Modelagem não supervisionada de tópicos, como ferramenta exploratória de padrões narrativos.

Essa sequência de procedimentos assegura a validade empírica e a reprodutibilidade científica da pesquisa, consolidando sua natureza experimental e aplicada no campo de Processamento de Linguagem Natural aplicado às Ciências Sociais.

4. Resultados e Discussão

O método adotado é flexível e modular, permitindo a escolha de diferentes algoritmos em cada etapa. Foram coletados 119 vídeos no *YouTube* com relatos de mulheres vítimas de VDF, dos quais 70 foram selecionados para análise e 49 descartados por não tratarem do fenômeno.

Dois experimentos foram conduzidos: (i) *UMAP + HDBSCAN* com *stopwords*; e (ii) *UMAP + HDBSCAN* sem *stopwords*. O segundo experimento apresentou melhor coerência semântica, identificando tópicos recorrentes relacionados ao ciclo da violência, ameaças, impactos psicológicos, apoio familiar e busca por ajuda institucional.

A análise demonstrou que a remoção de *stopwords* aumentou a clareza dos tópicos gerados.

4.1. *UMAP + HDBSCAN* + com *stopwords*

Nessa configuração, foi utilizado o *UMAP* para a redução de dimensionalidade e o *HDBSCAN* para a clusterização, destacando a vantagem de não ser necessário informar previamente o número de clusters, aspecto essencial para este estudo, uma vez que o total de agrupamentos a serem identificados não era conhecido a princípio (GROOTENDORST, 2022).

A aplicação do modelo, com o conjunto de dados sem remoção de *stopwords*, resultou na identificação de 40 tópicos, os quais estão apresentados os 20 primeiros no Quadro 1 devido a sua extensão e assim estamos apresentando os mais significativos.

Tópicos	Palavras por tópicos	Quantidade
-1	ter, meu, de, muito, porque, esse, falar, para, estar, ele	1750
0	estar, isso, ter, para, porque, gente, falar, muito, de, meu	174
1	voce, falar, assim, poder, ninguem, este, pessoa, querer, morrer, nada	142
2	entao, situacao, conhecer, esse, tudo, decisao, em, ter, sensacao, gente	114
3	para, meu, poder, isso, porque, percebi, falar, pessoa, demonio, ter	92
4	nunca, querer, mais, voltar, momento, viver, medo, terminar, tudo, aquele	89
5	gente, acontecer, isso, achar, luxo, coisa, tudo, aqui, comum, sociedade	67
6	tapa, soco, olho, cabelo, braco, puxar, murro, chao, sequela, facao	62

Tópicos	Palavras por tópicos	Quantidade
7	ser, sempre, estar, relacao, perdao, agressoes, ter, coisa, perdoar, algum	61
8	mulher, como, basear, perguntar, acreditar, apoio, quem, ajudar, ele, saber	60
9	delegacia, policia, medida, protetivo, denunciar, chamar, queixa, procurar, para, dizer	60
10	fiquei, medo, lembro, muito, calado, ficar, tempo, aquilo, aquele, cheguei	58
11	comecar, namorar, comeco, gente, comeca, anal, complicar, virgem, partir, começar	55
12	chegar, ir, cima, ligar, carona, jogar, telefone, pegar, afiliar, presa	52
13	gente, meia, noite, cafe, tomar, dormer, hora, tar, tarde, conversar	45
14	violencia, mulher, contra, violenciar, feminismo, genero, sobre, negro, luta, sofrer	42
15	peguei, cheguei, carro, casa, lembrar, meu, voltar, quando, ela, pai	42
16	casa, droga, sair, onde, chegar, semana, voltar, quando, embora, saer	40
17	muito, gente, pessoa, gentil, carinhoso, educado, sempre, atencioso, simpatico, prestativo	39
18	casamento, ano, familia, casado, casar, marido, emprego, trabalhei, virgem, meu	38
19	agredir, controle, totalmente, indelicado, alcoolisar, agressivo, bebir, beber, bebar, verbalmente	38
20	violencia, domestico, saude, violenciar, familiar, mulher, tipo, enfrentamento, cadastrar, contra	37

Quadro 1 – Resultado dos tópicos identificados (*UMAP* + *HDBSCAN* com *stopwords*)

Observa-se uma elevada concentração no tópico -1, classificado pelo *HDBSCAN* como *outlier*. Esse comportamento favorece a qualidade da representação, pois evita a alocação forçada de documentos em *clusters* inadequados.

Apesar disso, apenas 25 tópicos destacados no Quadro 2 apresentaram coerência semântica interpretável, correspondendo a 27,30% do total de documentos analisados. Os demais não revelaram clareza suficiente quanto aos temas representados, o que limita a análise interpretativa.

Diante dessa baixa representatividade, uma nova abordagem metodológica será apresentada na seção seguinte, buscando maior consistência nos resultados.

4.2. *UMAP* + *HDBSCAN* + sem *stopwords*

Nesta configuração, foi aplicado o *UMAP* para redução de dimensionalidade em conjunto com o *HDBSCAN* para a clusterização, desta vez com a remoção de *stopwords*, em comparação ao experimento anterior, seção 5.2.

A Figura 5 apresenta a distribuição visual dos tópicos gerados, apresentando uma concentração significativa de documentos nos tópicos 0 e 1, além da presença da cor

cinza-claro, correspondente ao tópico -1, que reúne outliers e ruídos identificados nos documentos.

Esse resultado indica que, mesmo após a remoção de *stopwords*, há uma parcela de registros não alocados em clusters semanticamente claros.

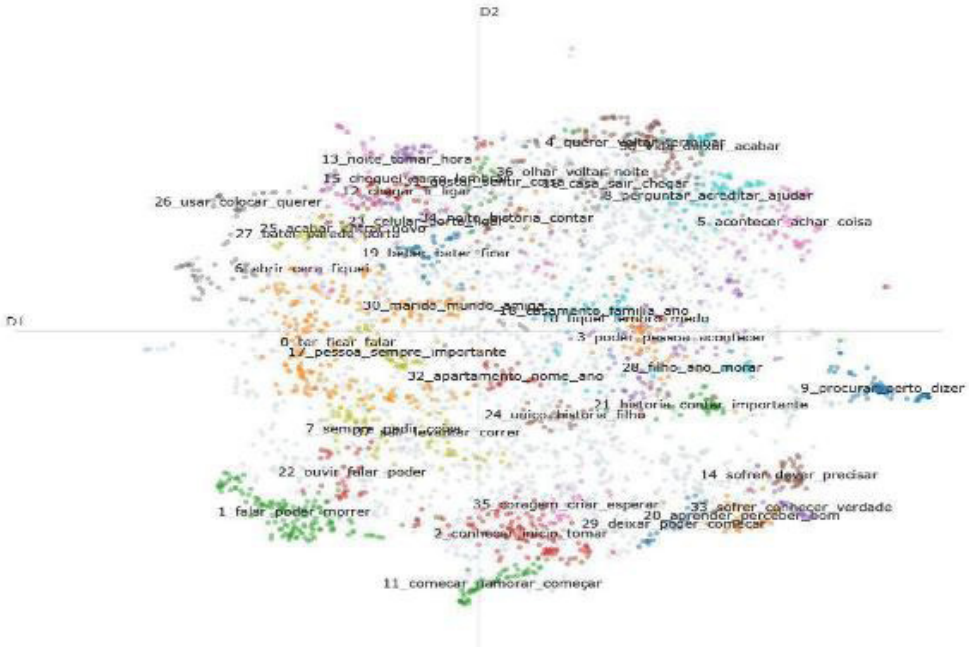

Figura 5 – Distribuição dos Tópicos

A etapa seguinte complementa a análise com a Figura 5, que apresenta um mapa de calor das similaridades entre tópicos. Observa-se que a maioria apresenta similaridade (≥ 0,90), o que sugere redundância entre os clusters e limita a diversidade semântica dos tópicos extraídos.

Com a remoção de *stopwords*, o modelo *UMAP + HDBSCAN* apresentou uma melhora na qualidade semântica dos tópicos gerados. A Figura 6 demonstra que os tópicos 10, 11 e 21 possuem menor similaridade em relação aos demais, ainda que mantenham níveis consideráveis de proximidade. Essa configuração resulta em interpretações mais consistentes do que as obtidas nas seções anteriores.

O Quadro 3 apresenta os tópicos identificados, nos quais foi possível interpretar 27 tópicos semanticamente relevantes, representando 31,54% do total de documentos analisados. Esse valor indica um avanço em relação ao resultado obtido com stopwords, Quadro 2. Vale ressaltar que neste quadro também são apresentados os primeiros 20 tópicos por sua maior representatividade.

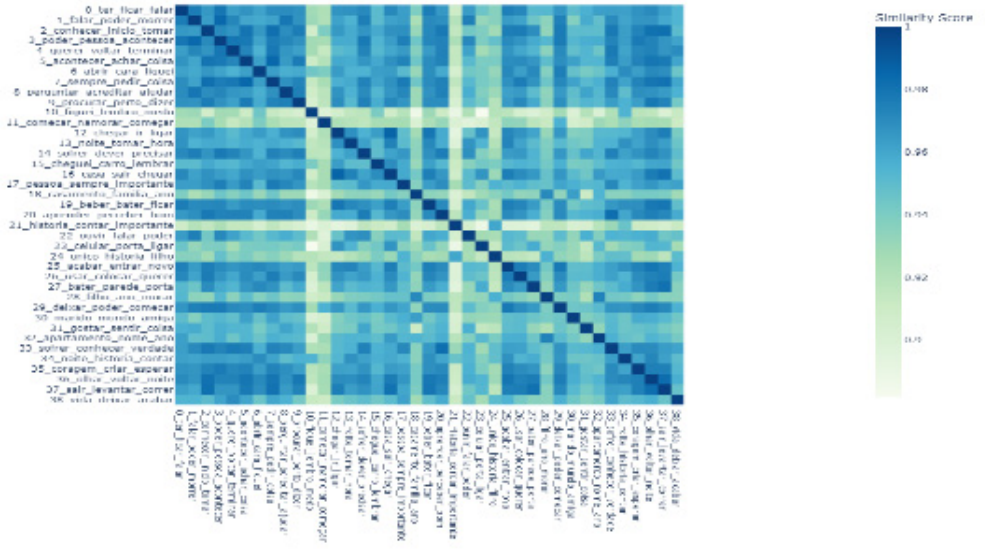

Figura 6 – Matriz de Similaridade

Tópicos	Palavras por tópicos	Quantidade
-1	ter, falar, ficar, dizer, casa, comecar, fazer, querer, passar, ano	1750
0	ter, ficar, falar, morar, dizer, filha, pessoa, fazer, sentir, trabalhar	174
1	falar, poder, morrer, entaar, pessoa, perguntar, voltar, passar, querer, saber	142
2	conhecer, inicio, tomar, separar, bater, achar, pensar, mudar, ter, entaar	114
3	poder, pessoa, acontecer, pensar, falar, mudar, comigo, problema, filho, acabar	92
4	querer, voltar, terminar, momento, viver, medo, amor, feliz, poder, acontecer	89
5	acontecer, achar, coisa, menina, vez, mundo, normal, forte, vida, saber	67
6	abrir, cara, fiquei, pegar, continuar, dar, parede, briga, correr, hora	62
7	sempre, pedir, coisa, ter, trabalho, momento, acontecer, aceitar, conseguir, grande	61
8	perguntar, acreditar, ajudar, saber, aceitar, pessoa, bom, chegar, olhar, trabalhar	60
9	procurar, perto, dizer, caso, perguntar, precisar, ter, olhar, chegar, dar	60
10	fiquei, lembro, medo, tempo, ficar, cheguei, comigo, sofrer, beber, momento	58
11	comecar, namorar, começar, novo, afastar, usar, sair, amigo, amiga, momento	55
12	chegar, ir, ligar, jogar, pegar, deixar, sair, casa, perguntar, porta	52
13	noite, tomar, hora, conversar, dormir, trabalhar, chegar, casa, conhecer, sair	45
14	sofrer, dever, precisar, existir, vir, direito, dia, falar, conversar, ano	42

Tópicos	Palavras por tópicos	Quantidade
15	cheguei, carro, lembrar, pai, voltar, casa, comprar, pegar, embora, porta	42
16	casa, sair, chegar, voltar, embora, ponto, dizer, trabalho, ir, entrar	40
17	pessoa, sempre, importante, conversar, mundo, inicio, pensar, perto, amigo, ajudar	39
18	casamento, familia, ano, marido, filho, morar, comprar, feliz, criar, acabar	38
19	beber, bater, ficar, chegar, jogar, sempre, voltar, noite, passar, pessoa	38
20	aprender, perceber, bom, lugar, direito, sofrer, contar, conhecer, mundo, começar	37

Quadro 2 – Resultado dos tópicos identificados (*UMAP* + *HDBSCAN* sem *stopwords*)

A análise mostra que a remoção das *stopwords* contribuiu para uma representação mais precisa das palavras nos tópicos, favorecendo sua compreensão semântica.

Assim como observado no Quadro 1, o tópico -1 concentrou o maior número de documentos, classificados como ruído ou *outliers*, com menor relevância para a representação dos tópicos. Diante disso, o aprofundamento da análise está sobre os tópicos destacados (2, 4, 5, 6 e 8) por apresentarem maior clareza semântica, conforme apresentado na Figura 7.

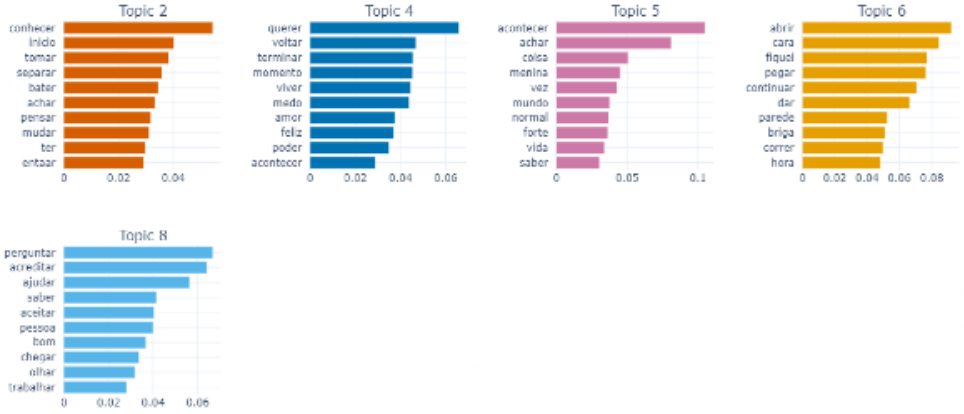

Figura 7 – Representação dos Tópicos 2, 4, 5, 6 e 8

A análise revelou tópicos com interpretação mais clara e consistente a partir da aplicação do *UMAP* + *HDBSCAN* com remoção de stopwords. O tópico 2, Quadro 3 traz narrativas de mulheres que buscaram forças para romper relacionamentos abusivos, refletindo decisões de separação e superação.

Tópico	Palavras por tópico	Documento
2	conhecer - inicio - tomar - separar - bater - achar - pensar - mudar - ter - entaar	“Então, eu tive muitas perdas até eu conseguir sair disso.”
2	conhecer - inicio - tomar - separar - bater - achar - pensar - mudar - ter - entaar	“Então eu passei por várias situações que já estavam ali me dando dica de que o relacionamento não era legal.”

Quadro 3 – Documentos associados ao tópico 2

O tópico 4, apresentado no Quadro 4, retrata a confusão emocional em relações violentas, onde sentimentos de medo, amor e arrependimento coexistem na fase do ato de violência do ciclo de Walker (1979).

Tópico	Palavras por tópico	Documento
4	querer - voltar - terminar - momento - viver - medo - amor - feliz - poder - acontecer	“Eu não queria que meu pai e minha mãe soubessem o que eu estava passando...”
4	querer - voltar - terminar - momento - viver - medo - amor - feliz - poder - acontecer	“Ele queria que eu acreditasse que o fato de ele ter ciúme demais era amor.”

Quadro 4 – Documentos associados ao tópico 4

O tópico 5, mostrado no Quadro 5, desta relatos de empoderamento e incentivo, nos quais as vítimas reafirmam a possibilidade de sair de situações abusivas e encorajam outras mulheres a buscarem ajuda.

Tópico	Palavras por tópico	Documento
5	acontecer - achar - coisa - menina - vez - mundo - normal - forte - vida - saber	“Eu saí e eu digo para as mulheres que a gente pode sair dessa.”
5	acontecer - achar - coisa - menina - vez - mundo - normal - forte - vida - saber	“Quem passa por isso vai me entender.”

Quadro 5 – Documentos associados ao tópico 5

O tópico 6, Quadro 6, concentra descrições explícitas de violência física severa, incluindo agressões, ameaças e sequelas permanentes, caracterizando o ápice da explosão violenta.

Tópico	Palavras por tópico	Documento
6	abrir - cara - fiquei - pegar - continuar - dar - parede - briga - correr - hora	“E... hoje eu sou uma mulher... que não consegue trabalhar... porque eu fiquei com sequelas.”
6	abrir - cara - fiquei - pegar - continuar - dar - parede - briga - correr - hora	“Ele socava a minha barriga, ele batia a minha cabeça na parede...”

Quadro 6 – Documentos associados ao tópico 6

Por fim, o tópico 8, Quadro 7 mostra a descrença social nos relatos das vítimas, expondo a dificuldade de reconhecimento da violência e o estigma que contribui para a perpetuação do ciclo abusivo.

Tópico	Palavras por tópico	Documento
8	perguntar - acreditar - ajudar - saber - aceitar - pessoa - bom - chegar - olhar - trabalhar	"Se eu falasse às vezes para algumas pessoas, elas não iam acreditar."
8	perguntar - acreditar - ajudar - saber - aceitar - pessoa - bom - chegar - olhar - trabalhar	"A sociedade aponta o dedo muito e colabora para isso."

Quadro 7 – Documentos associados ao tópico 8

No geral, os resultados confirmam que a estratégia adotada possibilitou a extração de tópicos semanticamente mais precisos, evidenciando os impactos da violência na vida das mulheres, e ao mesmo tempo, os desafios sociais e emocionais enfrentados por elas.

6. Conclusões

Este estudo apresentou técnicas de aprendizado de máquina e processamento de linguagem natural para identificar padrões semânticos em relatos de mulheres vítimas de violência doméstica e familiar coletados no *YouTube*, suprindo a ausência de repositórios disponíveis para esse tipo de dado, uma vez que, embora o PLN já tenha sido utilizado em estudos sobre violência doméstica em mídias sociais e registros institucionais (Xue, Chen & Gelles, 2019; Soldevilla & Flores, 2021; Corrêa & Faria, 2021; Campos & Andrêo, 2024; Kouzani & Nouman, 2025), não foram identificados trabalhos anteriores que aplicassem técnicas de extração automatizada diretamente a relatos espontâneos das próprias vítimas em língua portuguesa.

Entre os experimentos realizados, a combinação das técnicas *UMAP* + *HDBSCAN*, associada à remoção de *stopwords*, apresentou melhor desempenho na extração de tópicos semanticamente coerentes, permitindo identificar padrões discursivos e temas recorrentes nos relatos. Esses resultados demonstram a capacidade do método em revelar estruturas latentes de significado, oferecendo subsídios para compreender dimensões psicológicas, sociais e culturais da violência. Além disso, evidenciam a relevância da IA como ferramenta analítica e exploratória na investigação de fenômenos humanos complexos, conforme também observado por Guan et al. (2025) e Kouzani & Nouman (2025) em contextos internacionais de detecção de violência e apoio a sobreviventes.

O estudo reafirma o potencial de generalização do PLN e do aprendizado de máquina para outros contextos sociais e institucionais, como denúncias corporativas, discursos de ódio, injúria racial e LGBTfobia, contribuindo para o avanço de soluções baseadas em dados para o monitoramento de vulnerabilidades e comportamentos de risco. De forma alinhada aos achados de Campos & Andrêo (2024) e Santos (2025), reforça-se que o uso ético e supervisionado da IA pode favorecer estratégias de enfrentamento à violência e apoiar políticas públicas baseadas em evidências.

Entretanto, este artigo apresenta limitações. Nem todas as vítimas compartilham relatos em plataformas digitais, o que restringe a representatividade da amostra. Além disso,

a ausência de dados sociodemográficos e de bases comparativas limita a inferência estatística e a extrapolação dos resultados para populações mais amplas. Há, portanto, um desafio metodológico em equilibrar a riqueza qualitativa dos relatos espontâneos com a necessidade de padronização dos dados para análises quantitativas robustas.

Como trabalho futuro, recomenda-se a ampliação da coleta de dados em colaboração com ONGs e instituições de acolhimento, de modo a diversificar as fontes e melhorar a validade externa dos achados, também como sugestão o rotulamento dos tópicos com apoio de especialistas e a integração de modelos supervisionados de aprendizado profundo para análises de sentimentos e predição de risco mais precisas. Tais avanços poderão consolidar o papel da IA e do PLN como instrumentos científicos e sociais na compreensão e prevenção da violência doméstica e familiar contra a mulher.

Referências

Acosta, D. F., Gomes, V. L. de O., De Oliveira, D. C., Gomes, G. C., & Da Fonseca, A. D. (2017). Aspectos éticos e legais no cuidado de enfermagem às vítimas de violência doméstica. *Texto & Contexto - Enfermagem, 26*.

Al-Garadi, M. A., Kim, S., Guo, Y., Warren, E., Yang, Y. C., Lakamana, S., & Sarker, A. (2022). Natural language model for automatic identification of intimate partner violence reports from Twitter. *Array, 15*, 100217.

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. In *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 61–66). Institute of Electrical and Electronics Engineers Inc.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*, 77–84.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*, 662–679.

Brasil. (1994). *Convenção Interamericana para Prevenir, Punir e Erradicar a Violência contra a Mulher*. Convenção de Belém do Pará.

Brasil. (2006). *Lei nº 11.340*. Disponível em http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11340.htm

Brasil. (2022). *Política Nacional de Enfrentamento à Violência contra as Mulheres*.

Campos, E. A. V., & Andrêo, A. C. O. P. (2024). *Violência doméstica: utilização de IA para quantificação das ocorrências noticiadas na internet*. *Revista Camalotes – RECAM*. Faculdade Insted.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). *Density-based clustering based on hierarchical density estimates*. In *Advances in Knowledge Discovery and Data Mining* (pp. 160–172).

Chen, N. C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using machine learning to support qualitative coding in social science. *ACM Transactions on Interactive Intelligent Systems (TiiS), 8*(2), 1–20.

De Lucca, J. L., & Nunes, M. G. V. (2002). *Lematização versus stemming*. Universidade de São Paulo (USP), Universidade Federal de São Carlos (UFSCar) e Universidade Estadual Paulista (UNESP).

Dong, W., Charikar, M., & Li, K. (2011). *Efficient k-nearest neighbor graph construction for generic similarity measures*. Proceedings of the 20th International Conference on World Wide Web.

Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology, 7*, 6.

Elsayed, M., Abdelwahab, A., & Ahdelkader, H. (2019). A proposed framework for improving analysis of big unstructured data in social media. In *2019 14th International Conference on Computer Engineering and Systems (ICCES)* (pp. 61–65). Institute of Electrical and Electronics Engineers Inc.

Estroff, S. E., Penn, D. L., & Topor, A. (1994). The influence of social networks and social support on violence by persons with serious mental illness. *Psychiatric Services, 45*(7), 669–679.

Ferreira, Í. A., & Moraes, S. S. (2020). Subnotificação e Lei Maria da Penha: o registro como instrumento para o enfrentamento dos casos de violência doméstica contra a mulher considerando o anuário brasileiro de segurança pública (2019). *O Público e o Privado, 18*(37), 1–20.

Fox, C. (1989). A stop list for general text. *ACM SIGIR Forum, 24*(1–2), 19–35. https://doi.org/10.1145/378881.378888

Fornari, L. F., Lourenço, R. G., Oliveira, R. N. G. de, Santos, D. L. A. dos, Menegatti, M. S., & Fonseca, R. M. G. S. da. (2021). Domestic violence against women amidst the pandemic: Coping strategies disseminated by digital media. *Revista Brasileira de Enfermagem, 74*(29).

Ghani, N. A., Hamid, S., Hashem, I. A. T., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior, 101*, 417–428.

Giglietto, F., Rossi, L., & Bennato, D. (2012). The open laboratory: Limits and possibilities of using Facebook, Twitter, and YouTube as a research data source. *Journal of Technology in Human Services, 30*, 145–159.

Gil, A. C. (2022). *Como elaborar projetos de pesquisa* (7ª ed.). Atlas.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guan, S., Hui, V., Stiglic, G., Constantino, R. E., Lee, Y. J., & Wong, A. K. C. (2025). *Classifying the information needs of survivors of domestic violence in online health communities using large language models: Prediction model development and evaluation study. Journal of Medical Internet Research, 27*(1), e65397. https://doi.org/10.2196/65397

Kepios. (2022). *Global social media statistics — DataReportal – Global Digital Insights*. Disponível em https://datareportal.com/social-media-users

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 1–32.

Kouzani, A. Z., & Nouman, M. (2025). *Detecting indicators of violence in digital text using deep learning*. *Natural Language Processing Journal, 12*, 100175. https://doi.org/10.1016/j.nlp.2025.100175

Levendosky, A. A., Huth-Bocks, A. C., Semel, M. A., & Shapiro, D. L. (2004). The social networks of women experiencing domestic violence. *American Journal of Community Psychology, 34*(1–2), 95–109.

Lisboa, T. K., & Zucco, L. P. (2022). Os 15 anos da Lei Maria da Penha. *Revista Estudos Feministas, 30*(29).

Liu, T., Yang, Y., Huang, G. B., Yeo, Y. K., & Lin, Z. (2015). Driver distraction detection using semi-supervised machine learning. *IEEE Transactions on Intelligent Transportation Systems, 17*, 1108–1120. https://doi.org/10.1109/TITS.2015.2477470

McInnes, L., Healy, J., & Astels, S. (2016). *hdbscan: Hierarchical density based clustering*. *Journal of Open Source Software, 2*(11), 205.

McInnes, L., & Healy, J. (2017). Accelerated hierarchical density clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 33–42). IEEE Computer Society. https://doi.org/10.1109/ICDMW.2017.12

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1983). *Machine learning: An artificial intelligence approach*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-12405-5

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Science/Engineering/Math.

Mouton, J. P., Ferreira, M., & Helberg, A. S. J. (2020). A comparison of clustering algorithms for automatic modulation classification. *Expert Systems with Applications, 151*, 113317. https://doi.org/10.1016/j.eswa.2020.113317

Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: A review. *IEEE Access, 9*, 88364–88376.

Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research, 124*, 389–404.

Ni, Y., Barzman, D., Bachtel, A., Griffey, M., Osborn, A., & Sorter, M. (2020). Finding warning markers: Leveraging natural language processing and machine learning technologies to detect risk of school violence. *International Journal of Medical Informatics, 139*.

Pandove, D., Goel, S., & Rani, R. (2018). *Dimensionality reduction of text data: A comparative study*. *International Journal of Computer Applications, 179*(16), 6–11.

Previatti, J. F., & Milani, M. L. (2016). Violência contra a mulher no território da 25ª Agência de Desenvolvimento Regional (ADR) catarinense: Realidade social, políticas públicas e implicações para o desenvolvimento. *Colóquio - Revista do Desenvolvimento Regional, 13*, 179–197.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision. arXiv preprint* arXiv:2212.04356. https://doi.org/10.48550/arXiv.2212.04356

Rao, P., & Taboada, M. (2021). Gender bias in the news: A scalable topic modelling and visualization framework. *Frontiers in Artificial Intelligence, 4*, 82.

Reshamwala, A., Mishra, D., & Pawar, P. (2013). Review on natural language processing. *IRACST Engineering Science and Technology: An International Journal (ESTIJ), 3*, 113–116.

Russell, S., & Norvig, P. (2010). *Inteligência artificial* (2ª ed.). Elsevier.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*, 210–229.

Santos, W. de P. (2025). *Combate à violência doméstica contra a mulher: possíveis avanços a partir da inteligência artificial. Brazilian Journal of Development, 11*(2), 1–17. https://doi.org/10.34117/bjdv11n2-004

Saravanan, R., & Sujatha, P. (2018). A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 945–949). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICCONS.2018.8663155

Schraiber, L. B., D'Oliveira, A. F. P. L., França-Junior, I., & Diniz, C. S. G. (2007). Violência contra mulheres entre usuárias de serviços públicos de saúde da Grande São Paulo. *Revista de Saúde Pública, 41*, 359–367.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access, 8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., & Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems, 163*, 955–971.

Souza, A. P. de, Marin, M. J. S., & Rodrigues, P. S. (2021). O mundo sombrio das mulheres vítimas de violência: Uma análise qualitativa dos boletins de ocorrência. *New Trends in Qualitative Research, 8*, 97–105.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 11*, 2579–2605.

Vieira, P. R., Garcia, L. P., & Maciel, E. L. N. (2020). Isolamento social e o aumento da violência doméstica: O que isso nos revela? *Revista Brasileira de Epidemiologia, 23*.

Walker, L. (1979). *The battered woman*. Harper and Row.

World Health Organization. (2022). *Violence against women*. Disponível em https://www.who.int/news-room/fact-sheets/detail/violence-against-women

Xie, J., Yu, F. R., Huang, T., Xie, R., Liu, J., Wang, C., & Liu, Y. (2018). A survey of machine learning techniques applied to software defined networking (SDN): Research issues and challenges. *IEEE Communications Surveys & Tutorials, 21*, 393–430. https://doi.org/10.1109/COMST.2018.2866942

Zaldívar-Colado, A., Tripp-Barba, C., & Aguilar-Calderón, J. A. (2025). *Inteligencia artificial y creatividad cinematográfica: Revisión sobre el impacto de ChatGPT en la evolución de la escritura de guiones*. RISTI – Revista Ibérica de Sistemas e Tecnologia de Informação. (58), 30-35. https://doi.org/10.17013/risti.58.20–35.