Analysis of Criminal Patterns in Police Report Narratives using Spectral Clustering with K-means

Ricardo Rodrigues Barcelar; Flavia Rosane de Mendonça Luis; Claudia Aparecida Martins; Raphael de Souza Rosa Gomes; Anderson Castro Soares de Oliveira; Thiago Meirelles Ventura

Instituto de Computação – Universidade Federal de Mato Grosso (UFMT) Caixa Postal 78060-900 – Cuiabá – MT – Brasil

ricardo.barcelar@sou.ufmt.br, flavia.luis@sou.ufmt.br, claudia@ic.ufmt.br, raphael@ic.ufmt.br, anderson.oliveira@ufmt.br, thiago@ic.ufmt.br

Abstract. The high volume and heterogeneity of police report narratives in Brazil pose challenges for manual analysis and investigative prioritization. This work proposes an approach for identifying criminal patterns using clustering techniques applied to unstructured textual data. The methodology integrates Spectral Clustering with K-means, leveraging MPNet embeddings for vector representation and UMAP for dimensionality reduction. The resulting six clusters revealed thematic coherence, highlighting patterns such as bank fraud, judicial scams, social media crimes, and account hacking. Comparative experiments with Agglomerative Clustering were conducted using different linkage strategies, with Spectral Clustering achieving the highest silhouette score (0.38), indicating better-defined groups. A manual qualitative analysis of samples from each cluster supported the thematic distinctions. The study demonstrates that automatic clustering can contribute to investigative triage, offering relevant insights for public security applications.

CCS Concepts: • **Unsupervised learning** → **Clustering**.

Keywords: Spectral clustering, Police report, Public security, Clustering, UMAP

1. INTRODUÇÃO

A análise de dados criminais desempenha papel fundamental na gestão estratégica da segurança pública ao viabilizar a identificação de padrões e a otimização dos recursos policiais [Chainey e Ratcliffe, 2013]. No Brasil, o volume expressivo de boletins de ocorrência (BOs) registrados diariamente, associado à diversidade de crimes quanto a tipo, localização e modus operandi, impõe desafios significativos à alocação eficiente de recursos investigativos. Para ilustrar essa magnitude, a Secretaria de Segurança Pública do Estado de Mato Grosso reportou o registro de 394.896 boletins de ocorrência apenas em 2024 [SSPMT, 2025].

Entretanto, as narrativas textuais desses boletins, apresentam desafios substanciais para análise, sobretudo em razão de sua natureza não estruturada. A extração manual de padrões é dificultada pela ausência de padronização, pela heterogeneidade e subjetividade presentes nos relatos, bem como por imprecisões e inserção de informações redundantes ou pouco claras. Além disso, métodos tradicionais de análise, como buscas por palavras-chave, revelam-se insuficientes para capturar similaridades contextuais e relações mais complexas existentes nesses textos [Aggarwal e Zhai, 2012].

Nesse contexto, técnicas de processamento de linguagem natural aliadas a métodos de agrupamento, surgem como alternativas promissoras para organizar automaticamente ocorrências com base em características textuais. Essas abordagens revelam padrões e relações semânticas pouco acessíveis por

O presente estudo foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT), processo FAPEMAT-PRO.0000138/2025, o apoio da Secretaria Segurança Pública do Estado de Mato Grosso (SESP-MT) por meio do contrato n. 057/2023/SESP com a Universidade Federal de Mato Grosso (UFMT).

métodos tradicionais. No entanto, sua aplicação em narrativas de boletins de ocorrência ainda é incipiente, especialmente no Brasil, evidenciando uma lacuna na literatura.

Este trabalho tem como objetivo aplicar técnicas de agrupamento em narrativas de boletins de ocorrência, com foco na identificação de padrões criminais recorrentes. Utiliza-se a combinação de Spectral Clustering com K-means, além de experimentos comparativos com Agglomerative Clustering, para avaliar o desempenho dos métodos e sua contribuição à triagem automática e à priorização de investigações.

2. FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

O Spectral Clustering é uma técnica baseada em grafos que captura relações não lineares em dados complexos, como narrativas textuais. O método constrói uma matriz de adjacência para representar similaridades, calcula o Laplaciano do grafo (L = D - A), e projeta os dados em um espaço reduzido por meio dos autovetores associados aos menores autovalores [Von Luxburg, 2007]. Essa nova representação torna viável o uso de algoritmos tradicionais de agrupamento, com melhor separação de estruturas não lineares.

Nesse espaço, o algoritmo k-means é aplicado como etapa final para particionar os dados em clusters, aproveitando sua eficiência computacional para minimizar a variância intra-cluster. Essa integração combina a capacidade do spectral clustering de lidar com estruturas complexas com a praticidade do k-means [Jain, 2010].

O K-means é um algoritmo de particionamento que agrupa dados com base na minimização da variância intra-cluster, utilizando centróides como referência. Cada ponto é atribuído ao centróide mais próximo, geralmente com base na distância euclidiana. Embora eficiente, o K-means é sensível à escolha de k e à estrutura linear dos dados, o que limita sua aplicação isolada em textos complexos ou de alta dimensionalidade, justificando seu uso complementar no Spectral Clustering [Jain, 2010].

As técnicas de agrupamento têm sido exploradas em análises criminais, embora com foco predominante em dados estruturados. Oliveira e Zanusso [2005] combinaram k-means com um Mapa Auto-Organizável de Kohonen para segmentar vítimas de lesão corporal, mas limitaram-se a atributos numéricos, ignorando a riqueza semântica de narrativas. Andrade e Faria [2023] aplicaram DBSCAN a dados geográficos de ocorrências, analisando impactos da pandemia, enquanto Joshi, Sabitha e Choudhury [2017] usaram k-means para mapear taxas de criminalidade, sem considerar dinâmicas textuais.

Em outra análise, o algoritmo de agrupamento aglomerativo tem se mostrado eficaz na segmentação temática de textos curtos e não estruturados, como demonstrado por Jáñez-Martino et al. [2023], que utilizaram essa abordagem para rotular conjuntos de dados de spam em categorias temáticas por meio do Agglomerative Hierarchical Clustering.

Embora o agrupamento a partir de textos seja um método bastante viável, estudos de revisão sobre mineração de dados aplicados à análise criminal ressaltam que grande parte das abordagens permanece centrada em dados estruturados, com pouca ênfase na exploração de narrativas textuais, apesar do seu potencial informativo. Aggarwal e Zhai [2012] reforçam que métodos tradicionais, como buscas por palavras-chave, falham em capturar similaridades contextuais em textos não estruturados, validando a necessidade de técnicas avançadas como o spectral clustering com agrupamento final por k-means.

Entre os trabalhos analisados, o estudo de Lal Beejal et al. [2021] se destaca por aplicar o K-means a documentos textuais, enfatizando o papel do pré-processamento. No entanto, lacunas permanecem, pois não são consideradas as particularidades das narrativas criminais, como a variabilidade lexical e a subjetividade presentes nos boletins de ocorrência.

3. MATERIAIS E MÉTODOS

Esta seção descreve os procedimentos metodológicos adotados para o agrupamento de narrativas textuais de boletins de ocorrência.

3.1 Conjunto de Dados e Análise Exploratória

Foram analisados 15.580 boletins de ocorrência registrados entre 01/01/2024 e 31/12/2024, distribuídos para investigação na Delegacia Especializada de Estelionato de Cuiabá, Mato Grosso, Brasil. Cada boletim de ocorrência continha dados estruturados (natureza da ocorrência, município, bairro, latitude, longitude, data/hora, meios empregados, sexo, idade estimada, cor, orientação sexual, grau de instrução) e narrativas textuais descrevendo a dinâmica do crime.

A análise exploratória identificou importantes limitações nos dados estruturados: mais de 80% de valores ausentes em variáveis como cor, orientação sexual e grau de instrução; desbalanceamento em tipos criminais menos frequentes; presença de ruído (erros de digitação, duplicatas); heterogeneidade em escalas (variáveis numéricas e categóricas) e características redundantes (ex.: sobreposição entre informações georreferenciadas e dados de bairro/município), além de outliers em atributos como idade (valores <0 ou >100 anos).

Em contraste, as narrativas textuais dos boletins preservam a riqueza semântica e contextual dos eventos, permitindo a captura de informações que frequentemente escapam dos campos estruturados. Técnicas de processamento de linguagem natural possibilitam transformar esses textos em representações vetoriais capazes de expressar relações complexas e padrões latentes. Embora a integração com variáveis estruturadas pudesse ser considerada, desafios técnicos como a incompatibilidade de escalas, a predominância da dimensionalidade dos embeddings textuais sobre os demais atributos e a natureza eminentemente qualitativa das narrativas dificultam uma análise conjunta significativa [Aggarwal e Zhai, 2012].

Dessa forma, o presente trabalho opta por concentrar a análise nas narrativas, visando explorar todo o seu potencial descritivo e investigativo para a identificação de padrões criminais relevantes.

3.2 Pré-processamento das Narrativas

Com o intuito de agrupar apenas crimes de interesse da investigação policial, inicialmente foram removidos registros duplicados, vazios ou de fatos que não constituíam crimes identificados a partir do atributo "natureza da ocorrência", refletindo a exclusão de 4.549 registros, restando para as demais etapas 11.031 narrativas.

As narrativas passaram por um processo de limpeza textual voltado à remoção de ruídos, com o objetivo de aprimorar a qualidade do agrupamento. Por meio da biblioteca NLTK, foi realizada uma análise de frequência que permitiu identificar a presença de frases iniciais repetitivas, como por exemplo:

—Frase: 'compareceu nesta delegacia de policia o(a) comunicante' | Frequência: 1910

—Frase: 'sobre os fatos, o comunicante narra que no dia' | Frequência: 319

—Frase: 'compareceu nesta central de ocorrências policiais o comunicante noticiando que' | Frequência: 153

Além disso, a análise de n-grams de cinco palavras revelou agrupamentos recorrentes de termos sem relevância semântica, conforme exemplos a seguir:

—N-gram: 'as informações prestadas nesta ocorrência' | Frequência: 3810

—N-gram: 'inteira responsabilidade civil e criminal' | Frequência: 3754

—N-gram: 'todas as informações prestadas nesta' | Frequência: 3729

Posteriormente, o pré-processamento das narrativas envolveu etapas destinadas a aprimorar a qualidade dos dados textuais para o agrupamento. Além da eliminação dos agrupamentos repetitivos de palavras ou frases sem significado semântico, duplicidades e narrativas vazias foram removidas, assegurando a integridade do conjunto analisado.

A filtragem de stopwords utilizou a lista do NLTK adaptada ao português, visando eliminar termos de alta frequência e baixo valor informacional, o que contribui para a redução do ruído textual. Por fim, foi realizada uma anonimização complementar por meio de expressões regulares (regex) e do pacote SpaCy, para remover números de contas bancárias, documentos pessoais e endereços, fortalecendo a privacidade dos dados e complementando as rotinas de pré-anonimização previamente adotadas.

3.3 Geração de Embeddings e Redução de Dimensionalidade

Para a vetorização das narrativas, foi utilizado o modelo paraphrase-multilingual-mpnet-base-v2, da biblioteca sentence-transformers [Reimers e Gurevych, 2019]. A escolha se deu com base em comparações anteriores reportadas na literatura, nas quais o MPNet demonstrou desempenho superior em tarefas de similaridade semântica, superando modelos como BERT-base e RoBERTa-base em benchmarks como o STS Benchmark e o MTEB [Muennighoff et al., 2022]. Além disso, apresenta vetores de 768 dimensões, compatíveis com técnicas de redução e agrupamento aplicadas neste estudo.

Para mitigar a chamada "maldição da dimensionalidade" [Bellman, 1957], que dificulta a análise em espaços de alta dimensão, foi empregada a técnica UMAP (*Uniform Manifold Approximation and Projection*) [Mcinnes et al., 2018]. Parâmetros de `n_neighbors = 15` e `n_components = 5` foram definidos com base em testes exploratórios e, ao mesmo tempo, reduzir a dimensionalidade para uma representação vetorial adequada ao agrupamento. Valores inferiores ou superiores de `n_neighbors` resultaram em menor coerência temática nos agrupamentos identificados.

3.4 Agrupamento

Utilizando a biblioteca `scikit-learn`, o agrupamento foi conduzido utilizando o algoritmo *Spectral Clustering* com separação final dos grupos usando *K-means* dada sua simplicidade e por se mostrar eficaz após a transformação espectral. O algoritmo *Agglomerative Clustering* foi igualmente avaliado, visando à comparação metodológica com o método principal.

Inicialmente, foram realizados testes com diferentes configurações de afinidade e número de vizinhos para avaliar o impacto dos parâmetros no desempenho do modelo. Após experimentação empírica, optou-se por `affinity = nearest_neighbors` e `n_neighbors = 15`, por apresentarem melhor separação visual e valor de *Silhouette Score* [Rousseeuw, 1987]. O valor de `n_clusters = 6` foi definido com base nos resultados combinados do método do cotovelo e da métrica de Silhouette.

Para fins de visualização, os *embeddings* foram reduzidos para duas dimensões utilizando o UMAP (`n_components = 2`), facilitando a análise gráfica da distribuição dos *clusters*.

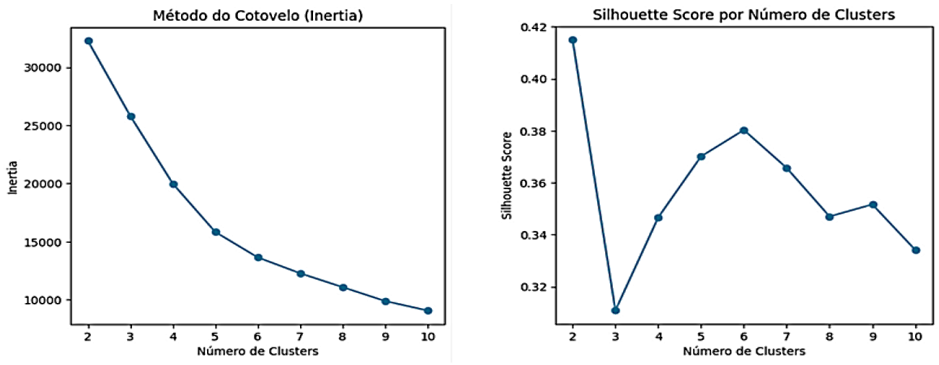
A partir dessas métricas, confirmou-se que a escolha de 6 *clusters* foi adequada, sendo esse valor respaldado tanto pelo comportamento do índice de inércia quanto por um *Silhouette Score* de 0,38. Foram realizadas simulações variando o número de *clusters* de 2 a 10, empregando o método do cotovelo (inércia) e a métrica de Silhouette para avaliar a qualidade dos agrupamentos.

Adicionalmente, com o objetivo de comparar o desempenho do *Spectral Clustering* com uma abordagem alternativa, foi aplicado o algoritmo *Agglomerative Clustering* sobre o mesmo conjunto de *embeddings* reduzidos, com parâmetros `n_clusters= [5,6,7]` e `linkage=[ward, complete, average]`.

Ao final do agrupamento, cada registro recebeu um rótulo correspondente ao seu *cluster*, possibilitando análises posteriores, como a investigação de padrões criminais, geração de estatísticas segmentadas e integração com outras bases ou métodos analíticos.

4. RESULTADOS E DISCUSSÕES

Para avaliar o desempenho, foram utilizados o Silhouette Score, a inércia e a análise qualitativa das amostras. A definição do número de clusters (k) considerou tanto o Silhouette Score quanto o método do cotovelo, cujo gráfico revelou um ponto de inflexão em k = 6, sugerindo uma boa relação entre redução de inércia e número de grupos. Essa escolha permite capturar a estrutura dos dados sem criar divisões excessivamente específicas, o que contribui para interpretações mais estáveis e generalizáveis, conforme ilustrado na Figura 1a.



(a) Gráfico de inércia **(b) Gráfico de Silhouette Score**

Fig. 1. Métricas de Avaliação

O Silhouette Score, por sua vez, apresenta o maior valor absoluto para k=2 (0,415), o que indica uma separação mais clara entre dois grandes grupos. Contudo, valores baixos em k=3 (0,15) e em k=4 (0,35) sugerem dificuldades de separação nesses pontos. Observa-se um novo patamar elevado a partir de k=5 (0,37), atingindo um valor próximo ao máximo em k=6 (0,38), e mantendo-se relativamente estável até k=7 (0,365). Portanto, a escolha por k=6 representa um compromisso entre a redução de inércia (cotovelo) e a manutenção de uma qualidade de agrupamento avaliada pelo Silhouette Score, permitindo a identificação de subgrupos mais informativos e úteis para análise criminal, sem perda significativa de coesão interna dos clusters, conforme Figura 1b.

Um fator que pode justificar o Silhouette Score moderado está relacionado à qualidade limitada das narrativas, que podem introduzir vieses e inconsistências. Esse cenário indica que a separação entre os clusters pode ser aprimorada com a adoção de técnicas adicionais de pré-processamento textual.

A metodologia proposta resultou em seis agrupamentos temáticos, cuja visualização bidimensional é apresentada na Figura 2. Essa representação, obtida por meio da técnica UMAP, ilustra os pontos coloridos conforme os rótulos de cluster, evidenciando uma separação visual entre os grupos. Observa-se, contudo, a presença de sobreposições em regiões de fronteira, o que sugere que os clusters apresentam contornos não perfeitamente definidos.

Para fins comparativos, também foi aplicado o algoritmo *Agglomerative Clustering* sobre os mesmos *embeddings* vetoriais reduzidos. Apesar de formar grupos visualmente definidos em algumas configurações, seus resultados foram inferiores aos do *Spectral Clustering*. A melhor configuração hierárquica, com `n_clusters = 6` e `linkage = complete`, obteve *Silhouette Score* de 0,35, abaixo dos 0,38 alcançados pelo *Spectral Clustering* com o mesmo número de grupos.

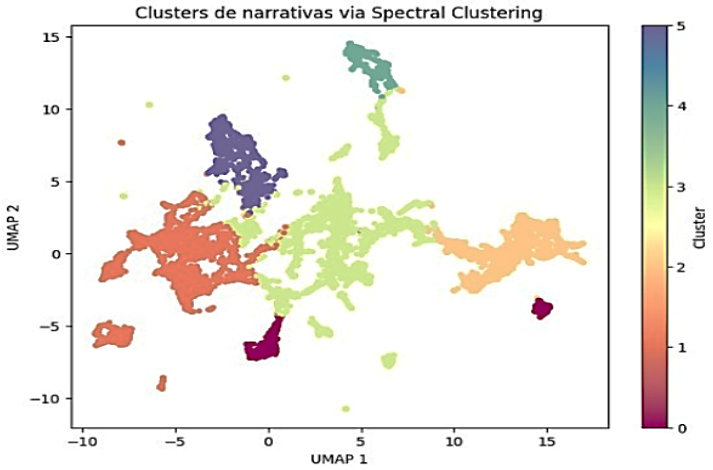


Fig. 2. Visualização dos Clusters

Além disso, observou-se maior sobreposição entre os agrupamentos, conforme ilustrado na Figura 3, indicando menor capacidade do método hierárquico em capturar estruturas semânticas não lineares presentes nas narrativas.

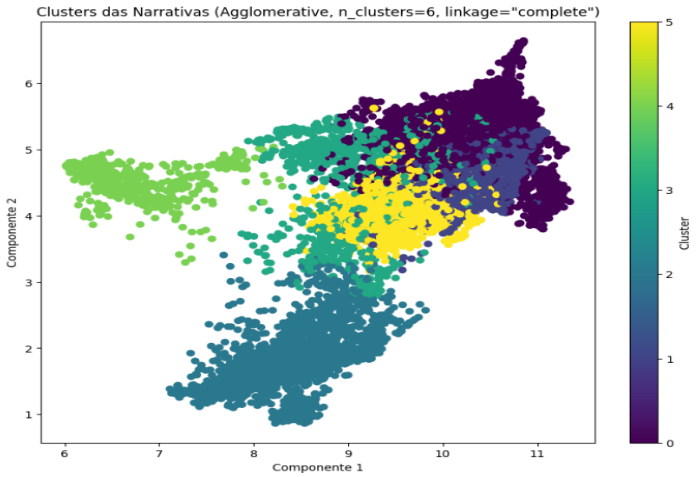


Fig. 3. Visualização dos Clusters por Agglomerative Clustering

Com base na melhor configuração do Spectral Clustering, cada narrativa recebeu um rótulo de grupo, permitindo a análise da distribuição das ocorrências entre os clusters. A Tabela 1 mostra a quantidade de boletins por agrupamento, evidenciando variações na concentração de registros entre os grupos.

Para compreender o perfil temático de cada grupo, foi realizada uma análise manual de uma amostra aleatória de 20 narrativas por cluster. A avaliação foi conduzida por dois pesquisadores com experiência em segurança pública, de forma independente, observando elementos característicos da narrativa em relação ao fato. Após a leitura, os temas predominantes de cada grupo foram consensualmente definidos com base na recorrência dos padrões observados:

—Cluster 0 (Questões Judiciais/Advogado/Processos): Narrativas associadas a crimes envolvendo supostos processos judiciais ou se passando por advogados (ex.: falsa taxa judicial, recebimento de

valores de causas judiciais, etc).

—Cluster 1 (Movimentação Bancária/Instituição Financeira): Casos relacionados a transações bancárias e instituições financeiras (ex.: transferências, empréstimos e pagamentos fraudulentos).

—Cluster 2 (Golpe envolvendo redes sociais): Estelionatos via mensagens ou publicações em redes sociais como instagram, facebook e whatsapp (ex.: clonagem de conta, falso perfil).

—Cluster 3 (Fraude em compra e venda de produtos e serviços): Casos relacionados à venda de produtos, especialmente na internet (ex.: falsa venda de celular, oferta de descontos em produtos).

—Cluster 4 (Invasão e hackeamento de contas): Casos de uso de identidades de terceiros, contas de redes sociais e acesso indevido a contas na internet (ex.: Criação de whatsapp em perfil de terceiros).

—Cluster 5: (Golpes em geral): Caso em que pessoas se passam por outras para obter vantagem indevida (ex.: criminoso se passando por algum parente para pedir dinheiro)

Cluster	Quantidade de Boletins de Ocorrência
0	626
1	3.134
2	1.904
3	3.632
4	500
5	1.236

Table I. Distribuição de boletins de ocorrência por cluster

Os resultados revelaram agrupamentos tematicamente coesos, cada um associado a padrões específicos de atividade criminosa. A análise manual de amostras permitiu identificar temas predominantes, como golpes envolvendo processos judiciais, atuação de falsos advogados, fraudes bancárias e crimes em redes sociais. Essa segmentação automática evidencia o potencial da abordagem para qualificar a triagem inicial, facilitar a detecção de *modus operandi* e apoiar a priorização investigativa.

Sob a perspectiva da gestão da segurança pública, a distribuição temática dos clusters favorece o encaminhamento de casos a equipes especializadas, aprimora a alocação de recursos e contribui para identificar redes criminosas por modalidade de delito.

5. CONSIDERAÇÕES FINAIS

Este estudo propôs uma abordagem para o agrupamento de boletins de ocorrência com base em narrativas textuais, utilizando a combinação de UMAP para redução de dimensionalidade e Spectral Clustering com particionamento final por K-means. A metodologia possibilitou identificar agrupamentos temáticos, como “Golpe envolvendo redes sociais” e “Movimentação bancária/instituição financeira”, demonstrando a aplicabilidade da técnica usada para segmentar ocorrências e superar as limitações de métodos tradicionais.

A contribuição central deste estudo reside na combinação de técnicas de processamento de linguagem natural e aprendizado não supervisionado para estruturar e interpretar narrativas policiais. A integração entre embeddings, redução de dimensionalidade e agrupamento permitiu identificar padrões criminais recorrentes, oferecendo uma base analítica sólida para apoiar a triagem e a gestão investigativa em contextos com grande volume de registros.

Entre as limitações, destaca-se a influência da baixa qualidade de algumas narrativas nos resultados, o que reforça a necessidade de aprimorar o pré-processamento textual. Além disso, os experimentos foram realizados com dados restritos, obtidos por meio de cooperação institucional, o que limita a generalização. Pesquisas futuras poderão explorar bases de outras regiões, além de comparar os resultados com abordagens baseadas em LLMs ou extração de tópicos, como BERTopic e TF-IDF.

Indo além, a integração entre dados estruturados (ex.: localização, tipificação penal) e narrativas também se mostra promissora, embora enfrente desafios técnicos relacionados à padronização e heterogeneidade das informações. Ainda assim, os resultados indicam que a abordagem pode apoiar a triagem automatizada e a priorização de investigações em contextos de alta demanda.

REFERÊNCIAS

Andrade, Rafael Lara Mazoni and de Faria, Bruno Lopes COVID-19 E Clusters de Homicídios em Belo Horizonte: Analise dos Impactos da Pandemia Sobre a Distribuição Espacial de Crimes, Caderno de Geografia, PP.489–489, 2023.

Aggarwal, Charu C and Zhai, ChengXiang, A survey of text clustering algorithms, Springer, pp.77-128, 2012.

Bellman, Richard E and Dreyfus, Stuart E, Applied dynamic programming, Princeton university press, 1957.

Chainey, Spencer, Crime mapping, Springer New York, Encyclopedia of Criminology and Criminal Justice, pp.699-709, 2013.

Jain, Anil K, Data clustering: 50 years beyond K-means, Pattern recognition letters, Elsevier, pp.651-666, 2010.

Janez-Martino, Francisco and Alaiz-Rodriguez, Rocio and Gonzalez-Castro, Victor and Fidalgo, Eduardo and Alegre, Enrique, Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach, Elsevier, Applied Soft Computing, 2023.

Joshi, Anant and Sabitha, A Sai and Choudhury, Tanupriya, Crime analysis using K-means clustering, 2017 3rd International conference on computational intelligence and networks (CINE), pp.33-39, 2017.

Lal Beejal, Chaman and Ahmed, Awais and Siyal, Reshma and Kumar, Suresh and Aftab, Shagufta and Jamali, Arshad, Text Clustering using K-MEAN, International Journal of Advanced Trends in Computer Science and Engineering, pp. 2892-2897, 2021.

McInnes, Leland and Healy, John and Melville, James, Umap: Uniform manifold approximation and projection for dimension reduction, Journal of Open Source Software, pp. 861, 2025.

Muennighoff, Niklas and Tazi, Nouamane and Magne, Loic and Reimers, Nils, Mteb: Massive text embedding benchmark, Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2023.

Oliveira, Fabiano R and Zanusso, Maria B, Clusterização de ocorrências policiais utilizando k-means e um mapa auto-organizável, CBRN, 2005.

Reimers, N.; Gurevych, I., Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982-3992, 2019.

Rousseeuw, Peter J., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Elsevier, Journal of computational and applied mathematics, pp. 53-65, 1987.

SSPMT - Sistema de Registro de Ocorências Policiais do Estado de Mato Grosso. Dados extraídos do módulo SROP, referente ao registro de boletins de ocorrência. Cuiabá: SSP-MT, 2025. Dados obtidos via acesso interno , 2025.

Von Luxburg, Ulrike, A tutorial on spectral clustering, Statistics and computing, Springer pp. 395-416, 2007