ISSN: 2317-0840

Métodos de Agrupamento para Análise de Dados Textuais dos Indicadores de Planejamento Estratégico do Estado de Mato Grosso

Lia Hanna Martins Morita[1†], Wellington Martins[1], Anderson Oliveira[1]

[1] *Departamento de Estatística, Universidade Federal de Mato Grosso (UFMT), Cuiabá, Mato Grosso*

Resumo: *Indicadores desempenham um papel essencial na administração eficaz de recursos e no monitoramento de políticas públicas. No estado de Mato Grosso, o Programa de Gerenciamento do Planejamento Estratégico visa aprimorar a governança pública por meio da adoção de indicadores padronizados. Este estudo explora o uso de métodos de mineração de texto e agrupamentos para analisar 160 indicadores distribuídos em 10 dimensões estratégicas: Aprendizagem e Conhecimento; Desenvolvimento Econômico; Educação; Esportes, Cultura e Lazer; Estrutura Organizacional; Fiscal; Infraestrutura, Saneamento Básico e Meio Ambiente; Satisfação da Sociedade; Saúde e Vulnerabilidade Social. Utilizando o método de Ward para agrupamento e a métrica de distância de cossenos, os indicadores foram agrupados em dois grupos formados por características textuais. Os indicadores da dimensão Educação situaram-se majoritariamente no primeiro grupo, enquanto os demais indicadores foram agregados no segundo grupo. Os resultados deste trabalho são úteis para os gestores públicos na tomada de decisões estratégicas, promovendo melhorias nas políticas públicas municipais.*

Palavras-chave: *Agrupamento de Textos; Análise Exploratória de Dados Textuais; Políticas Públicas.*

Clustering Methods for Textual Data Analysis of Strategic Planning Indicators of the State of Mato Grosso

Abstract: *Indicators play an essential role in effective resource management and monitoring public policies. In Mato Grosso, the Strategic Planning Management Program aims to improve public governance by adopting standardized indicators. This study explores the use of text mining and clustering methods to analyze 160 indicators distributed across ten strategic dimensions: Learning and Knowledge; Economic development; Education; Sports, Culture, and Leisure; Organizational structure; Supervisor; Infrastructure, Basic Sanitation, and Environment; Society Satisfaction; Health and Social Vulnerability. Using Ward's method for grouping and the cosine distance metric, the indicators were grouped into two groups formed by textual characteristics. The indicators of the Education dimension were mainly located in the first group, while the other indicators were aggregated in the second group. The results of this work are useful for public managers in making strategic decisions and promoting improvements in municipal public policies.*

Keywords: *Text Clustering; Exploratory Analysis of Textual Data; Public Policy.*

[†]Autor correspondente: profaliaufmt@gmail.com

Manuscrito recebido em: 29/07/2024
Manuscrito revisado em: 01/10/2024
Manuscrito aceito em: 03/10/2024

Introdução

Indicadores são essenciais para a gestão de recursos e o monitoramento eficiente de políticas públicas nos municípios (MARSHALL; MEIER, 2004). Desenvolvido por Kaplan e Norton em 1992, o Balanced Scorecard (BSC), ou Indicadores Balanceados de Desempenho (em português), é uma abordagem estratégica que exprime as necessidades de uma organização em um conjunto de indicadores de desempenho, destacando-se por alinhar as atividades e metas de curto prazo com a visão de longo prazo da organização (KAPLAN; NORTON, 1992).

Desde 2009, o TCE-MT adota o BSC como metodologia de construção de seus planos estratégicos, visando avaliar o desempenho da administração pública de maneira mais ampla e integrada. Em 2022, o TCE-MT iniciou o Programa de Gerenciamento do Planejamento Estratégico (GPE), que consiste em uma iniciativa colaborativa com os municípios, com a adoção de indicadores padronizados que visam aprimorar a qualidade dos serviços públicos e os resultados das políticas públicas (MATO GROSSO. TRIBUNAL DE CONTAS DO ESTADO, 2022). Esta abordagem mede e quantifica as condições socioeconômicas das regiões, ao mesmo tempo que fornece informações valiosas sobre o desempenho e a eficácia dos programas governamentais.

As metodologias estatísticas contribuem para a análise e a interpretação efetivas dos dados, permitindo diagnósticos precisos e a identificação de tendências emergentes. Jannuzzi (2018) salienta a relevância de integrar abordagens qualitativas, quantitativas e participativas para obter uma visão mais abrangente da realidade.

A mineração de texto envolve a extração de informações úteis de grandes volumes de dados textuais, em que utilizam-se técnicas de processamento de linguagem natural (PLN). Morita, Cruz e Oliveira (2023) utilizaram métodos de mineração de texto para unificar as nomenclaturas dos indicadores nos municípios participantes do GPE, facilitando a gestão e o acompanhamento destes. A análise crítica e objetiva destes indicadores é uma área de interesse para os gestores públicos.

O objetivo geral deste trabalho é utilizar técnicas de mineração de texto para agrupar os indicadores do GPE dos municípios do estado de Mato Grosso em grandes grupos (*clusters*), por meio de suas características textuais.

Referencial Teórico

Indicadores de Planejamento Estratégico

Os indicadores de planejamento estratégico consistem em um material metodológico que oferece apoio aos gestores para elaborar, executar e monitorar os indicadores dos municípios participantes do GPE. Estes indicadores são baseados nos fundamentos teóricos do BSC, conectando os objetivos da gestão de longo prazo com os resultados e projetos da organização no momento atual.

Métodos de Mineração de Texto

A mineração de texto é um campo emergente na ciência de dados que envolve a extração de informações úteis de grandes volumes de dados textuais. Utilizando uma variedade de técnicas de PLN e análise estatística, a mineração de texto transforma dados não estruturados em conhecimento. Com o crescimento exponencial de dados textuais na internet e no meio corporativo, essa abordagem se tornou essencial em várias áreas, incluindo saúde, finanças e administração pública (MINER et al., 2012; FELDMAN; SANGER, 2007). Um componente fundamental no processo de mineração de texto é o pré-processamento dos dados. Isso inclui a limpeza de dados, remoção de ruídos, normalização de texto e a extração de características.

Um *corpus* é um conjunto estruturado de textos utilizados para conduzir análises e pesquisas linguísticas. Esses textos são coletados e organizados de forma sistemática para representar uma língua ou um domínio específico.

O desenvolvimento e a utilização de *corpora* permitem a análise quantitativa e qualitativa de fenômenos linguísticos, proporcionando descobertas sobre o uso da língua, padrões sintáticos, léxicos e semânticos. A criação de um *corpus* envolve a seleção criteriosa de textos que representem adequadamente a diversidade e a variação da língua.

Embeddings são representações vetoriais densas de dados, que são utilizadas para representar palavras, frases, ou documentos em espaços de alta dimensão, capturando relações semânticas e sintáticas nos textos. Os modelos de *embeddings* são construídos com a metodologia de redes neurais profundas, utilizando grandes *corpora* textuais. Estes modelos são capazes de capturar o significado semântico das palavras e frases, produzindo *embeddings* de alta qualidade que podem ser utilizados para várias tarefas de PLN, como classificação de texto, recuperação de informação e análise de sentimentos (OPENAI, 2024).

Metodologia

Catálogo Unificado de Indicadores

O conjunto de dados textual foi cedido pelo Tribunal de Contas do Estado de Mato Grosso, contendo 160 indicadores organizados em 10 dimensões estratégicas: Aprendizagem e Conhecimento; Desenvolvimento Econômico; Educação; Esportes, Cultura e Lazer; Estrutura Organizacional; Fiscal; Infraestrutura, Saneamento Básico e Meio Ambiente; Satisfação da Sociedade; Saúde e Vulnerabilidade Social. A Tabela 1 ilustra o catálogo de indicadores do GPE.

Table 1: Description of the unified catalog of GPE indicators.

Dimensões	Número de Indicadores
Aprendizagem e Conhecimento	8
Desenvolvimento Econômico	26
Educação	35
Esportes, Cultura e Lazer	20
Estrutura Organizacional	6
Fiscal	8
Infraestrutura, Saneamento Básico e Meio Ambiente	9
Satisfação da Sociedade	6
Saúde	39
Vulnerabilidade Social	3
Total	**160**

Source: from the authors (2024).

Todos os indicadores foram apresentados mediante uma ficha de qualificação contendo as seguintes informações:

- Indicador: nome do indicador;
- Conceituação: informações que definem o indicador como ele se expressa, se necessário agregando elementos para a compreensão do seu conteúdo;
- Polaridade: descrição da polaridade do resultado de acompanhamento do indicador, indicando se um valor "maior" é considerado melhor ou se um valor "menor" é considerado melhor;
- Interpretação: explicação sucinta do tipo de informação obtida e seu significado;

- Como medir (Método de cálculo): fórmula utilizada para calcular o indicador, definindo precisamente os elementos que a compõem;
- Fonte: Fonte e/ou bibliografias utilizadas para obter o indicador.

Os Quadros 1 e 2 exemplificam as fichas de indicadores específicos, como a "Índice de infestação predial por *Aedes Aegypti*" e a "Taxa de abandono escolar na educação infantil (pré-escola)", demonstrando como cada indicador é definido, medido e interpretado.

Chart 1: Aedes Aegypti Building Infestation Index indicator sheet.

Nome	Índice de Infestação Predial por *Aedes Aegypti*
Conceituação	Percentual de imóveis que apresentam larvas do *Aedes Aegypti* em relação ao total de imóveis pesquisados.
Polaridade	Quanto menor melhor.
Interpretação	Este indicador varia de 0 a 100%, sendo que quanto mais próximo de 100%, pior é a situação dos imóveis pesquisados, dado que os índices mais baixos significam menor presença de larvas do mosquito nos imóveis.
Como medir (método de cálculo)	$$IIP = \frac{I}{P} \times 100$$ em que: • IIP é o índice de infestação predial por *Aedes Aegypti*; • I é o número de imóveis com presença de larvas por *Aedes Aegypti*; • P é o número total de imóveis pesquisados.

Source: Brasil (2005).

Chart 2: Indicator sheet School Dropout Rate in Early Childhood Education (pre-school).

Nome	Taxa de Abandono Escolar na Educação Infantil (pré-escola)
Conceituação	A taxa de abandono escolar na educação infantil refere-se à proporção de crianças, com idade entre 4 e 5 anos, que deixam de frequentar a pré-escola antes de completar o período letivo. Este indicador mede especificamente os casos em que não há formalização da transferência para outra instituição de ensino. Ele é crucial para entender os fatores que influenciam a descontinuidade precoce da educação em idades fundamentais para o desenvolvimento infantil.
Polaridade	Quanto menor melhor.
Interpretação	Este indicador varia de 0 a 100%, sendo que quanto mais próximo de 0%, menor é a taxa de abandono na educação infantil. Valores elevados indicam mais crianças que interrompem seus estudos precocemente.
Como medir (método de cálculo)	$$TAEI = \frac{A}{M} \times 100$$ em que: • TAEI é a taxa de abandono escolar na educação infantil; • A é o número de crianças que abandonaram a escola de educação infantil; • M é o número total de matrículas nas escolas de educação infantil.

Source: Brasil (2014).

Obtenção do Corpus textual

O *corpus* textual foi obtido por meio da descrição textual dos indicadores na ficha de qualificação.

API da OpenAI

A API da OpenAI oferece uma maneira poderosa e flexível de acessar modelos avançados de aprendizado de máquina, como os modelos de *embeddings* textuais. O uso da API da OpenAI envolveu o processo de registro, com a verificação do e-mail e do número de telefone. Após a criação da conta, foi possível acessar o painel de controle da OpenAI e gerar uma nova chave de API na seção "API Keys". Esta chave foi necessária para autenticar as solicitações à API. Com a chave de API, foi possível enviar os textos dos 160 indicadores utilizando linguagem de programação em Python conforme a Figura 1.

A resposta da API contém uma lista de vetores ou *embeddings* (OPENAI, 2024). O modelo *text-embedding-3-small* consiste em uma rede neural profunda, treinado em grandes *corpora* textuais, capturando o significado semântico das palavras e frases.

Na terceira etapa, foi efetuado o cálculo da matriz de distâncias entre as unidades dos 160 indicadores, utilizando a medida do cosseno entre vetores em um espaço de alta dimensão (MANNING; RAGHAVAN; SCHÜTZE, 2008).

Foi realizado o cálculo do Índice de Calinski-Harabasz (CH) por número de *clusters*, em que os valores elevados indicam melhor separação destes grupos (CALIŃSKI; HARABASZ, 1974).

Figure 1: Example of using the OpenAI API to obtain textual embeddings with the text-embedding-3-small model.

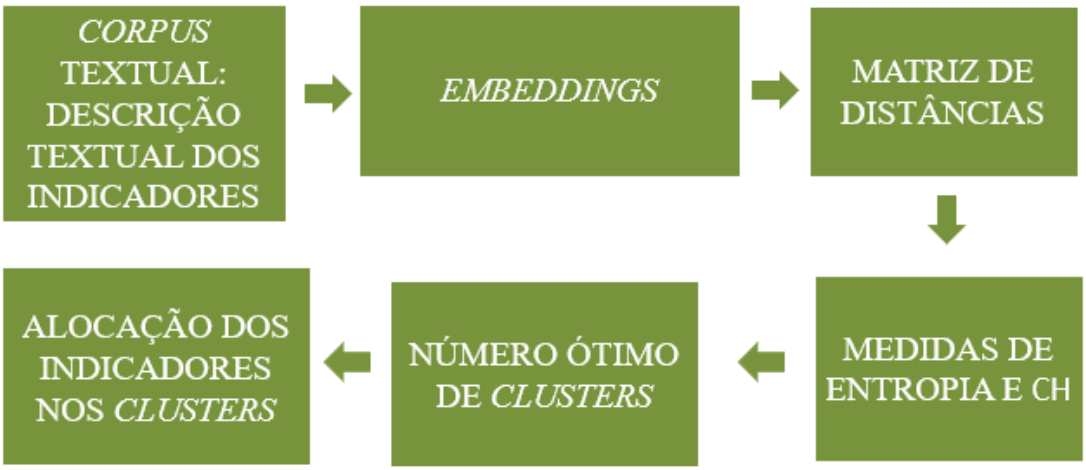
```python
import openai

# Substitua 'your-api-key' pela sua chave de API
openai.api_key = 'your-api-key'

response = openai.Embedding.create(
    input="Seu texto aqui",
    model="text-embedding-3-small"
)

embeddings = response['data'][0]['embedding']
print(embeddings)
```

Source: from the authors (2024).

Também foi realizado o cálculo da medida de entropia por número de *clusters*, em que os valores baixos indicam que os *clusters* são mais homogêneos.

Por fim, alocam-se os indicadores dentros dos *clusters*, baseado no número ótimo de *clusters*, pelo método de Ward (MURTAGH; LEGENDRE, 2014). A descrição destes passos está exibida no organograma da Figura 2.

Figure 2: Steps for building clusters.

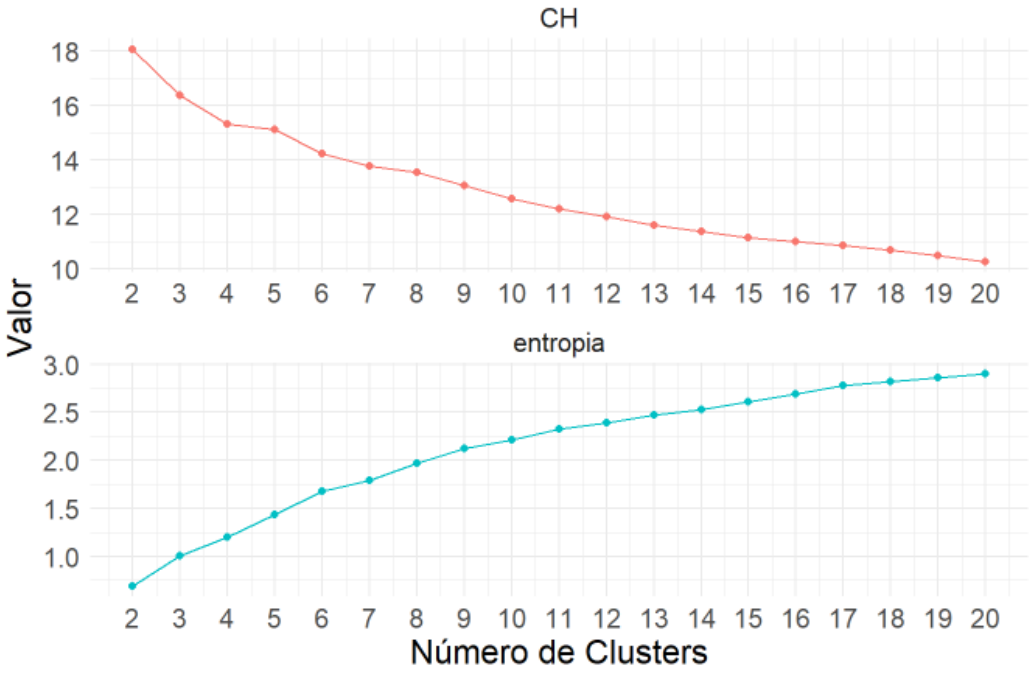


Source: from the authors (2024).

Resultados e Discussão

A Figura 3 exibe os resultados de CH e medida de entropia, em que o número ótimo de *clusters* é igual a 2.

Figure 3: CH metrics and entropy versus number of clusters.



Source: from the authors (2024).

Sendo assim, os 160 indicadores do planejamento estratégico foram alocados em dois *clusters*, baseado nas *embeddings* resultantes de suas características textuais: nome, descrição e interpretação, conforme exibido na Tabela 2.

Table 2: Alocação dos Indicadores nos *Clusters*.

Indicadores	Total	*Cluster* 1	*Cluster* 2
Aprendizagem e Conhecimento	8	0	8
Desenvolvimento Econômico	26	0	26
Educação	35	34	1
Esportes, Cultura e Lazer	20	0	20
Estrutura Organizacional	6	0	6
Fiscal	8	0	8
Infraestrutura, Saneamento Básico e Meio Ambiente	9	0	9
Satisfação da Sociedade	6	0	6
Saúde	39	37	2
Vulnerabilidade Social	3	3	0
Total	**160**	**74**	**86**

Source: from the authors (2024).

Conclusão

Este artigo destaca a grande importância da padronização dos indicadores para a eficácia das políticas públicas nos municípios do Mato Grosso. Os indicadores das dimensões Educação, Saúde e Vulnerabilidade Social foram alocados majoritariamente no primeiro grupo; os demais indicadores foram agregados no segundo grupo. Estes resultados auxiliam na tomada de decisões e contribuem para a melhoria nas políticas públicas. A pesquisa ressalta o valor de metodologias inovadoras e estratégicas no manejo de dados, demonstrando que a padronização e implementação cuidadosa de indicadores de desempenho são fundamentais para otimizar a governança e os serviços públicos nos municípios.

Agradecimentos

Agradecimentos a Universidade Federal de Mato Grosso, TCE-MT e Fundação Uniselva por todo o suporte dado para o desenvolvimento e publicação desta pesquisa.

Referências

BRASIL. *Plano Nacional de Educação (PNE). Lei nº 13.005, de 25 de junho de 2014*. Brasília: Ministério da Educação (MEC), 2014.

BRASIL. MINISTÉRIO DA SAÚDE. SECRETARIA DE VIGILÂNCIA EM SAÚDE. *Diagnóstico Rápido nos municípios para vigilância entomológica do Aedes Aegypti no Brasil – LIRAa: metodologia para avaliação dos índices de Breteau e Predial*. Diretoria Técnica de Gestão. Brasília: Ministério da Saúde, 2005. P. 60. (Série A. Normas e Manuais Técnicos). p. 19.

CALIŃSKI, T.; HARABASZ, J. *A dendrite method for cluster analysis*. Communications in Statistics, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. DOI: 10.1080/03610927408827101. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/03610927408827101>.

FELDMAN, R.; SANGER, J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. [S.l.]: Cambridge University Press, 2007.

JANNUZZI, P. M. *A importância da informação estatística para as políticas sociais no Brasil: breve reflexão sobre a experiência do passado para considerar no presente*. Revista Brasileira de Estudos de População, São Paulo, v. 35, n. 1, e0055, jan. 2018.

KAPLAN, R.; NORTON, D. *The Balanced Scorecard—Measures That Drive Performance*. Harvard Business Review, n. 1, p. 71–79, jan. 1992. jan-fev.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. [S.l.]: Cambridge University Press, 2008.

MARSHALL, G.; MEIER, K. J. *The Political Economy of Local Government*. In: FERLIES, E.; LYNN, L. E.; POLLITT, C. (Ed.). The Oxford Handbook of Public Management. Oxford: Oxford University Press, 2004. P. 205–228.

MATO GROSSO. TRIBUNAL DE CONTAS DO ESTADO. *Resolução Normativa nº 14/2022 de 28 de junho de 2022: Dispõe sobre a instituição do Programa de Apoio à Gestão do Planejamento Estratégico dos Municípios, denominado GPE, no âmbito do Tribunal de Contas do Estado de Mato Grosso*. Cuiabá: [s.n.], jun. 2022.

MINER, G. et al. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. [S.l.]: Elsevier, 2012.

MORITA, L. H. M.; CRUZ, R. C.; OLIVEIRA, A. C. S. de. *Text Mining Methods for Unifying Strategic Planning Indicators in the Municipalities of Mato Grosso*. Sigmae, v. 12, n. 3, p. 39–50, dez. 2023. Disponível em: <`https://publicacoes.unifal-mg.edu.br/revistas/index.php/sigmae/article/view/2219`>.

MURTAGH, F.; LEGENDRE, P. *Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?* Journal of Classification, v. 31, p. 274–295, 2014. DOI: `10.1007/s00357-014-9161-z`.

OPENAI. *Embeddings API*. [S.l.: s.n.], 2024. Retrieved from https://www.openai.com/embeddings-api.