ISSN: 2317-0840

Métodos de mineração de texto para unificação de indicadores de planejamento estratégico nos municípios de Mato Grosso

Lia H. M. Morita[1†], Rita C. Cruz[2], Anderson C. S. Oliveira[1]

[1]*Departamento de Estatística – Universidade Federal do Mato Grosso, Cuiabá - MT*
[2]*Curso de Graduação Bacharelado em Estatística – Universidade Federal do Mato Grosso, Cuiabá - MT*

Resumo: *Indicadores são ferramentas importantes na gestão de recursos e no monitoramento de políticas públicas nos municípios. No Estado de Mato Grosso, o programa Gerenciamento do Planejamento Estratégico (GPE), desenvolvido em parceria com o Tribunal de Contas, é uma iniciativa que promove o aperfeiçoamento de políticas públicas por meio da adoção de indicadores padronizados. Neste contexto, a mineração de texto surge como uma técnica valiosa para analisar e processar grande volume de informações em documentos e relatórios. Expressões regulares são sequências de caracteres em textos que obedecem a uma regra, como, por exemplo, palavras acentuadas com acento agudo, circunflexo, til ou crase. Com o uso de algoritmos, estes padrões textuais são detectados e substituídos ou eliminados. Por exemplo, se deseja-se retirar a acentuação nas palavras de uma frase, utilizam-se algoritmos para detectar estes acentos e retirá-los. Outro problema de interesse é a padronização das palavras para minúsculo ou inicial maiúscula com uso de expressão regular. Estas aplicações auxiliam nas tarefas cotidianas, pois textos padronizados são mais fáceis de serem analisados para tirar conclusões em decisões de negócio. Neste trabalho, foram utilizados métodos de mineração de texto e expressões regulares para a unificação das nomenclaturas dos indicadores nos municípios participantes do GPE, auxiliando na gestão e no acompanhamento destes. A mineração de texto permitiu a análise sistemática das informações, identificando melhorias, corrigindo as inconsistências e aprimorando a efetividade das políticas públicas.*

Palavras-chave: *Mineração de Texto; Expressões Regulares; Planejamento Estratégico; Indicadores de Monitoramento; Municípios do Estado de Mato Grosso.*

Text mining methods for unifying strategic planning indicators in the municipalities of Mato Grosso

Abstract: *Indicators are essential tools for resource management and monitoring public policies in municipalities. In the state of Mato Grosso, the Strategic Planning Management Program (GPE), developed in collaboration with the Tribunal de Contas, is an initiative that seeks to enhance public policies by adopting standardized indicators. Within this framework, text mining emerges as a valuable technique for analyzing and processing vast amounts of data in documents and reports. Regular expressions are sequences of characters in a text that follow a specific pattern, such as words accented with acute, circumflex, tilde, or grave accents. These patterns can be detected through algorithms, then replaced or removed. For example, if the objective is to remove the accent from all words in a sentence, algorithms can be employed likewise. Another assignment of interest is standardizing text to lowercase or title case using regular expressions. Such modifications streamline daily tasks and aid in the framing of managing reports, as standardized texts are more straightforward to analyze, derive insights from, and base business decisions on. In this study, text mining techniques and regular expressions were employed to standardize the nomenclature of indicators from municipalities participating in the GPE, thereby enhancing their management and oversight. Text mining allowed for a systematic analysis of the data, pinpointing improvement, correcting inconsistencies, and thereby bolstering the efficacy of public policies.*

Keywords: *Text Mining; Regular Expressions; Strategic Planning; Monitoring indicators; Municipalities of Mato Grosso State.*

[†]Autora correspondente: profaliaufmt@gmail.com.

Introdução

Lidar com os desafios complexos que a sociedade enfrenta no século XXI exige uma abordagem estratégica e eficaz por parte das autoridades públicas. A gestão eficiente dos recursos municipais e a formulação de políticas públicas bem-sucedidas desempenham um papel fundamental no progresso sustentável das regiões. Nesse contexto, a importância dos indicadores sociais como ferramentas de avaliação e monitoramento se torna evidente (MOREIRA; SANTINI, 2022; SILVA, 2022).

Os indicadores sociais são instrumentos que permitem mensurar e quantificar diversos aspectos da realidade socioeconômica de uma região ou comunidade. Eles fornecem informações valiosas sobre áreas como saúde, educação, segurança, meio ambiente, entre outras. Através da coleta e análise de dados, os indicadores oferecem percepções sobre o desempenho das políticas públicas, a eficácia dos programas e a alocação de recursos (JANUZZI, 2018; MOREIRA; SANTINI, 2022; SILVA, 2022).

No entanto, a mera coleta de indicadores não é suficiente. É fundamental a utilização de uma abordagem estatística robusta para interpretar e analisar esses dados. A abordagem estatística desempenha um papel crucial na gestão estratégica, permitindo a realização de diagnósticos detalhados e a identificação de padrões e tendências. Ela também viabiliza a construção de sistemas de indicadores que monitoram ações, avaliam processos e resultados de programas, e investigam possíveis impactos e externalidades negativas (JANUZZI, 2018).

A importância da abordagem estatística é ressaltada por Januzzi (2018), que destaca a necessidade de uma combinação de metodologias qualitativas, quantitativas e participativas. Essa diversificação permite uma compreensão mais abrangente da realidade, envolvendo os diversos atores interessados, como beneficiários, usuários, profissionais e gestores.

Kirch *et al*. (2019) investigaram os indicadores de desempenho das universidades federais brasileiras, demonstrando que tais indicadores não só servem como instrumentos de avaliação de desempenho, mas também permitem identificar tendências e agrupar vastos conjuntos de dados através de técnicas estatísticas multivariadas, potencializando a gestão destas instituições.

Além disso, a utilização de indicadores como ferramentas operacionais no monitoramento e avaliação de políticas públicas ao longo do ciclo das políticas é uma prática essencial. Isso contribui para subsidiar decisões informadas, identificar áreas de melhoria e direcionar os esforços para atingir metas e objetivos preestabelecidos (JANNUZZI, 2016).

No estado de Mato Grosso, o Tribunal de Contas do Estado (TCE-MT) desempenha um papel significativo no aprimoramento da gestão pública. Desde 2009, o TCE-MT adota o Balanced Scorecard (BSC) na construção de planos estratégicos, avaliando o desempenho público em efetividade, eficácia e eficiência. Em 2012, lançou o Programa de Desenvolvimento Institucional Integrado (PDI) para compartilhar conhecimento em Planejamento Estratégico, governança e tecnologias eficientes, especialmente com os municípios (MATO GROSSO, 2013).

Em 2022, o TCE-MT lançou o Programa de Apoio ao Gerenciamento do Planejamento Estratégico (GPE), em colaboração com os municípios. Este visa melhorar a qualidade dos serviços públicos e os resultados das políticas públicas. O GPE promove a cultura do planejamento e oferece suporte na implementação das estratégias, buscando otimizar a eficácia das políticas e impulsionar o desenvolvimento sustentável (MATO GROSSO, 2022).

Inicialmente, 24 dos 141 municípios já participantes do PDI aderiram ao GPE, apoiando seus Planejamentos Estratégicos e monitorando-os por meio de indicadores. No entanto, a falta de padronização nesses indicadores resulta em abordagens, nomenclaturas e métodos distintos, dificultando a comparação socioeconômica e de resultados entre os municípios (DE OLIVEIRA *et. al*., 2019).

A disparidade nas fontes de dados e a falta de uniformidade nos indicadores adotados pelos órgãos governamentais são questões que desempenham um papel crucial no cenário da gestão pública e na tomada

de decisões informadas. Essas divergências podem ter um impacto significativo na eficácia das políticas públicas, na avaliação do desempenho e no monitoramento do progresso (JANUZZI, 2018; SILVA, 2022).

A falta de uniformidade nos indicadores adotados também pode gerar ambiguidade e confusão na interpretação dos resultados. Indicadores similares com terminologias distintas ou métodos de medição diferentes podem levar a conclusões equivocadas ou incompletas. Além disso, a falta de padrões torna difícil a comparação entre diferentes regiões ou municípios, dificultando a identificação de tendências ou áreas que requerem atenção especial (FITZPATRICK, SANDERS; WORTHEN, 2011; JANNUZZI, 2016).

Para abordar essa problemática, é essencial promover a padronização e dos indicadores utilizados pelos órgãos governamentais. Isso não apenas melhoraria a qualidade das informações, mas também facilitaria a comparação e a análise, permitindo uma compreensão mais precisa do panorama geral e um planejamento mais eficiente. Tecnologias como a mineração de texto podem ser empregadas para extrair informações valiosas de documentos e relatórios, contribuindo para a criação de uma base de dados mais coesa e integrada (LEE; YOON, 2021; SHIBUI, 2018; MODRUŠAN, MRŠIĆ; RABUZIN, 2021).

O presente trabalho tem como objetivos a aplicação da mineração de texto em dados dos indicadores do programa GPE e a obtenção de uma base unificada desses indicadores.

Referencial Teórico

Mineração de texto

Na era da informação, onde dados são gerados em grande escala, extrair conhecimento de volumes extensos de texto é fundamental. A mineração de texto (do inglês *Text Mining*) é uma técnica crucial nesse contexto, permitindo a análise e extração de informações de textos, frases e palavras. Ao utilizar algoritmos computacionais, identificam-se padrões e informações que frequentemente não são acessíveis por métodos convencionais de busca, especialmente em formatos não estruturados, comuns em documentos (FELDMAN; SANGER, 2006; JURAFSKY; MARTIN, 2020; ZIZKA; DARENA; SVOBODA, 2019).

Neste processo, examinam-se termos relevantes em um conjunto de textos, permitindo identificar padrões e categorizar tópicos com base na frequência de palavras. Explora textos de forma sistemática, descobrindo temas subjacentes em domínios específicos (FELDMAN; SANGER, 2006; JURAFSKY; MARTIN, 2020; ZIZKA; DARENA; SVOBODA, 2019). O método permite transformar documentos em formatos numéricos, como matrizes, para analisar a distribuição de palavras e reconhecer os padrões. Os dados brutos são refinados por técnicas estatísticas e processamento de linguagem natural (do inglês *Natural Language Processing*) para extrair informações valiosas (FELDMAN; SANGER, 2006; JURAFSKY; MARTIN, 2020; ZIZKA; DARENA; SVOBODA, 2019).

As expressões regulares, do inglês *regex*, são ferramentas poderosas para manipular e analisar caracteres em textos. Estas ferramentas apresentam aplicação em áreas diversas, como linguística computacional, bioinformática, mineração de dados, segurança da informação e desenvolvimento de software (FELDMAN; SANGER, 2006; FRIEDL, 2006; JURAFSKY; MARTIN, 2020). O processamento de linguagem natural envolve tarefas essenciais de *tokenização*, remoção de caracteres indesejados e identificação de padrões linguísticos. Na mineração de dados e análise textual, identificam-se, extraem-se e analisam-se padrões em grandes conjuntos de dados textuais (FRIEDL, 2006; JURAFSKY; MARTIN, 2020).

A mineração de texto inclui tarefas variadas como associação, sumarização, classificação, agrupamentos e análise linguística, buscando extrair conhecimento de textos extensos, classificar textos e identificar padrões nestes textos. Essas técnicas são fundamentais para compreender informações complexas em textos, auxiliando na tomada de decisões (FELDMAN; SANGER, 2006; JURAFSKY; MARTIN, 2020; ZIZKA; DARENA; SVOBODA, 2019).

O agrupamento de textos é uma técnica importante na análise de dados textuais, pois permite agrupar documentos similares. Esta técnica utiliza algoritmos para encontrar padrões e relações entre documentos, oferecendo percepções significativas e tomadas de decisões informadas. Envolve a representação numérica dos documentos e algoritmos como *k-means* e *clustering* hierárquico. Essencial em áreas como análise de conteúdo digital, marketing e processamento de linguagem natural, ajuda a compreender e organizar dados textuais (FELDMAN; SANGER, 2006; JURAFSKY; MARTIN, 2020; ZIZKA; DARENA; SVOBODA, 2019).

Balanced Scorecard

O *Balanced Scorecard (BSC)*, conhecido em português como Indicadores Balanceados de Desempenho, é uma abordagem estratégica de gestão que visa traduzir a visão e estratégia de uma organização em um conjunto completo de indicadores de desempenho. Criado por Kaplan e Norton em 1992, o BSC se destaca como uma ferramenta valiosa para alinhar as atividades e metas de curto prazo de uma empresa com sua visão de longo prazo.

A essência do BSC é reconhecer que a performance de uma organização não pode ser avaliada apenas por meio de indicadores financeiros. Em vez disso, propõe uma abordagem equilibrada considerando quatro perspectivas: Financeira, Processos Internos, Aprendizado e Conhecimento, e Cidadãos (Sociedade). A partir dessas perspectivas, são definidos objetivos específicos, indicadores, metas e iniciativas para alcançar a visão estabelecida (KAPLAN; NORTON, 1996).

O BSC vem ganhando visibilidade na gestão pública, pois possibilita alinhar objetivos de longo prazo com ações operacionais. Ele cria um mapa estratégico que liga metas a indicadores específicos, melhorando serviços, otimizando processos, desenvolvendo equipes e garantindo transparência. Sua aplicação visa aumentar a eficiência e eficácia dos serviços públicos, fortalecendo a relação entre a administração e os cidadãos.

Além disso, De Oliveira et al. (2019) propôs um quadro teórico para trabalhar com os indicadores municipais no estado de Mato Grosso. Esse quadro envolve não apenas as perspectivas, mas também 10 dimensões e 63 subdimensões, visando uma análise mais abrangente dos indicadores.

Metodologia

A metodologia adotada para este estudo compreendeu diversas etapas, visando à organização e análise dos indicadores presentes no banco de dados do planejamento estratégico de 21 municípios participantes do Programa GPE.

Inicialmente, o banco de dados coletou informações sobre os indicadores em formato de texto, abrangendo sete características essenciais: nome do município, objetivo, metas de longo e curto prazo, nome do indicador e descrição de como medir.

A primeira etapa consistiu na coleta das informações textuais, englobando o nome do indicador e a respectiva descrição de como era medido. Com um total de 1305 registros, os dados coletados possuíam uma rica diversidade de conteúdo, mas também apresentavam variações em termos de formato, tornando necessário o emprego de técnicas de mineração de texto para extrair conhecimentos de maneira significativa. Para promover a uniformidade nos dados, a segunda etapa concentrou-se na utilização de expressões regulares. Essas ferramentas permitiram a padronização dos textos de acordo com critérios previamente estabelecidos, garantindo uma base consistente para a análise subsequente (FRIEDL, 2006; JURAFSKY; MARTIN, 2020).

Na terceira fase, foram identificados e tratados os indicadores cujos nomes faziam parte de outros indicadores. Essa etapa foi crucial para evitar redundâncias e ambiguidades nos dados, assegurando uma representação precisa dos indicadores.

Na quarta fase, procedeu-se à verificação da distância entre os caracteres (denominados *strings*), com o intuito de identificar a similaridade entre os indicadores. Essa análise permitiu a detecção de padrões e relações semânticas entre os termos, contribuindo para uma segmentação mais precisa dos dados (FRIEDL, 2006; JURAFSKY; MARTIN, 2020).

A quinta fase envolveu a geração de uma matriz de documentos e termos, na qual as linhas representavam os documentos (indicadores) e as colunas representavam os termos relevantes. Utilizando a distância euclidiana, aplicou-se o escalonamento multidimensional para visualizar a disposição dos indicadores no espaço multidimensional (FRIEDL, 2006; JURAFSKY; MARTIN, 2020).

Em seguida, na sexta fase, procedeu-se à unificação manual das nomenclaturas dos indicadores. Adotou-se um nome padrão relacionado a indicadores já reconhecidos por órgãos como a Organização Mundial da Saúde (OMS) e o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Nos casos em que não havia correspondência em tais órgãos, o nome original da base inicial foi mantido.

Finalmente, na última etapa, cada indicador foi atribuído a uma perspectiva e dimensão de acordo com o Balanced Scorecard (BSC), conferindo uma estrutura organizada e abrangente à análise dos indicadores de desempenho. O resultado foi uma abordagem sólida e completa que proporcionou descobertas valiosas para o planejamento estratégico dos municípios participantes do Programa GPE.

Todos os procedimentos de mineração de texto foram conduzidos por meio do software R, utilizando os pacotes *tm* (FEINERER; HORNIK, 2023), *textclean* (RINKER, 2018) e *LexisNexisTools* (GRUBER, 2023). O pacote *tm* permitiu a manipulação e transformação dos textos, enquanto o *textclean* desempenhou um papel crucial na aplicação de expressões regulares e na padronização dos dados. Além disso, o pacote *LexisNexisTools* facilitou a detecção de sobreposições e similaridades entre os indicadores, contribuindo para a identificação de padrões semânticos e a formação de grupos de indicadores.

Resultados e Discussão

Na Tabela 1, é fornecida uma visão geral do número de indicadores por município. Destaca-se que o município com o maior número de indicadores é o município de Campo Verde, com um total de 200 indicadores. Em contraste, o município de Alta Floresta possui 28 indicadores. Agregando-se todas as informações dos municípios listados, chega-se a um total de 1305 indicadores que compõem a base de dados. Essa variedade de indicadores evidencia a riqueza e a diversidade das informações coletadas, oferecendo uma visão abrangente dos esforços de planejamento estratégico em diferentes localidades.

Na fase inicial da análise dos 1305 indicadores, foram utilizadas expressões regulares para aplicar diversos padrões de formatação. Isso incluiu a normalização das letras maiúsculas e minúsculas, a remoção de acentos e a exclusão de números nos nomes dos indicadores. Por meio dessa abordagem, foi possível identificar e agrupar 933 registros únicos, como demonstrado no Quadro 1.

Posteriormente, esse mesmo procedimento foi estendido para as descrições que explicavam como os indicadores eram medidos. Isso resultou em uma redução para 830 registros. Além disso, 153 indicadores foram eliminados por apresentarem composições idênticas, porém com denominações diferentes.

No Quadro 2, estão exemplificados indicadores cujos nomes possuem fragmentos presentes em outros indicadores. Após a conclusão dessa verificação, observou-se uma redução no número total de indicadores para 763 registros únicos. Isso ocorreu devido ao processo de identificação e eliminação de indicadores redundantes ou que compartilhavam parte de seus nomes com outros indicadores. Essa etapa foi crucial para garantir a integridade e a precisão dos dados, bem como para evitar duplicações e ambiguidades nas análises subsequentes.

Tabela 1: Número de indicadores por municípios.

Município	Número de Indicadores
Água Boa	47
Alta Floresta	28
Cáceres	59
Campo Verde	200
Cuiabá	77
Diamantino	45
Itiquira	36
Juína	35
Juscimeira	43
Lucas do Rio Verde	53
Nortelândia	57
Primavera do Leste	97
Querência	53
Rondonópolis	66
São Félix do Araguaia	67
São José dos Quatro Marcos	52
Sapezal	60
Sinop	45
Tangará da Serra	62
Tapurah	56
Várzea Grande	67
Total	1305

Fonte: Dos autores.

Quadro 1: Exemplos de aplicação das expressões regulares na formatação do nome do indicador.

Texto	**Descrição**	**Texto após formatação**
Cobertura Potencial na Educação Infantil	padronização de caixa (minúsculo)	cobertura potencial na educação infantil
1.1:taxa de mortalidade infantil	retirar índices numéricos antes do texto	taxa de mortalidade infantil
- taxa: de incidência ? de dengue.	retirar dois pontos, ponto final e ponto de interrogação	taxa de incidência de dengue
número de óbitos por acidentes de trânsito	retirar acentos	numero de obitos por acidentes de transito
indice da educacao básica	mudar para sigla	ideb

Fonte: Dos autores.

Em seguida, foi realizada a verificação da similaridade entre os textos. Nesse contexto, a atenção foi direcionada para a inspeção dos indicadores com nomes que demonstravam semelhanças entre si. Essa abordagem teve o objetivo de identificar possíveis sobreposições ou coincidências nos nomes dos indicadores, indicando redundâncias ou variações mínimas entre os registros. Isso resultou na redução do número de indicadores para 656 registros únicos. O quadro 3 ilustra exemplos desses indicadores, proporcionando uma visão detalhada das semelhanças encontradas e permitindo uma análise aprofundada dos padrões observados.

Quadro 2: Exemplos de indicadores cujos nomes possuem fragmentos de outros indicadores.

Nome do indicador	Texto após formatação
percentual de arrecadacao do iptu	1. percentual de arrecadacao do iptu em relacao a receita prevista 2. percentual de arrecadacao do iptu em relacao a receita
area verde por habitante	1. indice de area verde por habitante 2. numero de metros quadrados de area verde por habitante (oms) 3. metros quadrados de area verde por habitante
vias urbanas pavimentadas	1. taxa de vias urbanas pavimentadas 2. km de vias urbanas pavimentadas 3. percentual de vias urbanas pavimentadas 4. numero de vias urbanas pavimentadas

Fonte: Dos autores.

Quadro 3: Exemplo da verificação de similaridade dos nomes dos indicadores.

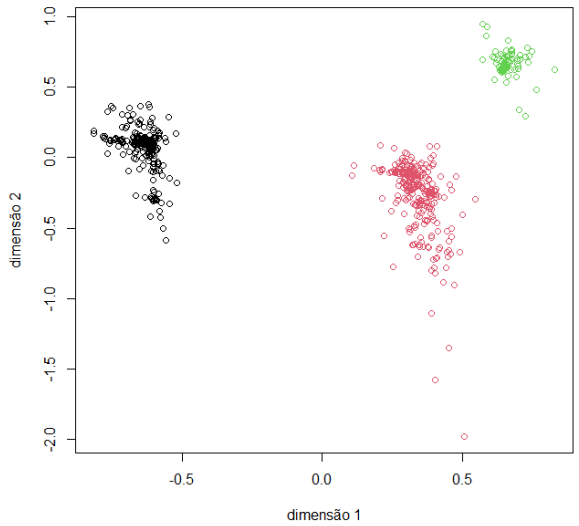
Texto Original	Texto Duplicado	Similaridade
percentual de cobertura do atendimento na educacao infantil (4 a 5 anos)	percentual de cobertura do acesso a educacao infantil para criancas na faixa etaria de 4 a 5 anos	0,80
taxa de acidentes de transito	numero de acidentes de transito	0,86
taxa de acidentes de transito	indice de acidentes de transito	0,86
taxa de acidentes de transito	percentual de acidentes de transito	0,86
percentual de cobertura da educacao infantil (4 a 5 anos)	taxa de cobertura na educacao infantil (4 a 5 anos)	0,90
taxa de participante em reunioes de tomada de decisoes em politicas publicas	taxa de participante em reunioes de tomada de decisao em politicas publicas	0,95
percentual de cobertura da educacao infantil (0 a 3 anos)	percentual de cobertura da educacao infantil publica (0 a 3 anos)	0,95
percentual de cobertura da educacao infantil (4 a 5 anos)	percentual de cobertura da educacao infantil publica (4 a 5 anos)	0,95

Fonte: Dos autores.

A Figura 1 apresenta os grupos de indicadores que foram identificados por meio da aplicação do método *K-Means*, levando em consideração o espaço de escalonamento multidimensional. A análise realizada resultou na criação de três grupos distintos nos quais os indicadores foram agrupados com base na similaridade dos textos que compõem seus nomes. Essa abordagem permitiu uma visualização clara das relações entre os indicadores, revelando padrões de similaridade semântica em suas denominações.

Na Figura 2, são apresentadas as nuvens de palavras dos termos que compõem cada um dos grupos (clusters) identificados. O Cluster 1 é composto por indicadores cujas denominações estão relacionadas a valores percentuais, enquanto o Cluster 2 abrange indicadores cujas denominações estão associadas a números. O Cluster 3 não possui um termo predominante. A visualização desses termos em grupos distintos proporciona uma compreensão clara das características semânticas compartilhadas entre os indicadores de cada cluster.

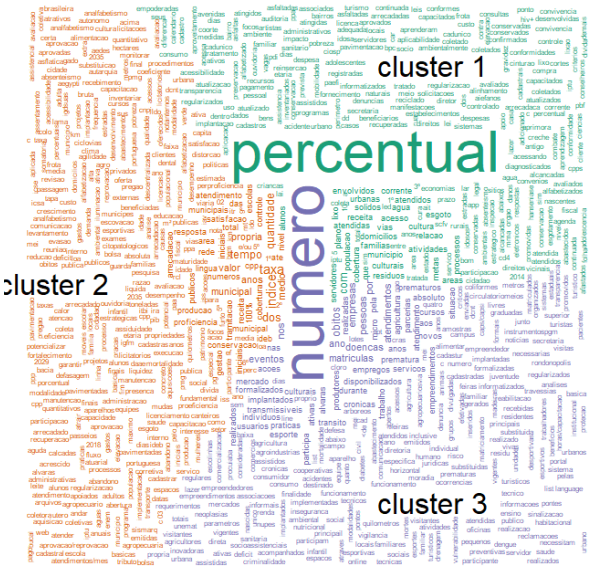
Figura 1: Agrupamento de Indicadores pelo Método *K-Means* no Espaço de Escalonamento Multidimensional.



Fonte: Dos autores.

Dessa forma, em todos os indicadores que continham o termo "número", procedeu-se à substituição por "percentual". Em seguida, foi realizada nova análise de similaridade, resultando na redução para 423 indicadores únicos. Posteriormente, repetiu-se a análise de cluster, o que resultou na obtenção de dois clusters distintos. Esse processo de ajuste e refinamento permitiu uma segmentação mais precisa dos indicadores, refletindo nuances semânticas mais relevantes e culminando na formação desses dois clusters bem definidos. O primeiro é formado por indicadores cujas denominações estão relacionadas a valores percentuais, enquanto o segundo não possui um termo predominante (Figura 3).

Figura 2: Nuvem de palavras para cada grupo identificado.



Fonte: Dos autores.

Figura 3: Nuvem de palavras para cada grupo identificado.

Fonte: Dos autores.

A etapa seguinte do processo englobou a unificação manual das nomenclaturas de 423 indicadores. Essa unificação foi realizada com base em indicadores já existentes em órgãos oficiais. Nos casos em que não foi possível encontrar um indicador correspondente nesses órgãos, optou-se por atribuir o nome descrito no banco de dados original. Além disso, durante esse processo de unificação, foram identificados e removidos 78 indicadores duplicados. Essa ação de detecção e eliminação de duplicatas assegurou que apenas informações únicas e relevantes fossem preservadas, contribuindo para a integridade e qualidade dos dados.

Tabela 2: Número de indicadores unificados por perspectiva e dimensão.

Perspectiva	Dimensão	Número de indicadores
Sociedade	Educação	36
Sociedade	Esportes cultura e lazer	23
Sociedade	Indústria e comercio	28
Sociedade	Infraestrutura	45
Sociedade	Saneamento básico e meio ambiente	33
Sociedade	Saúde	44
Sociedade	Segurança	3
Sociedade	Vulnerabilidade social	22
Processos	Atendimentos	3
Processos	Estrutura operacional	1
Processos	Gestão	57
Processos	Infraestrutura	1
Processos	Processos	2
Processos	Satisfação da sociedade	4
Financeira	Fiscal	32
Aprendizagem e conhecimento	Desenvolvimento humano	11
Total		345

Fonte: Dos autores.

Foram unificados um total de 345 indicadores após a etapa de unificação manual das nomenclaturas. Essa ação permitiu a consolidação de informações e a eliminação de redundâncias, resultando em um conjunto mais coeso e de maior qualidade. A configuração desses indicadores unificados por perspectiva e dimensão é apresentada de maneira clara e organizada na Tabela 2. Nessa tabela, é possível verificar a quantidade de indicadores agrupados em cada perspectiva e dimensão do Balanced Scorecard (BSC), proporcionando uma visão abrangente e estruturada da contribuição de cada área para o planejamento estratégico. Esse processo de unificação e categorização reflete o compromisso em simplificar e otimizar a interpretação dos dados, facilitando a análise e a tomada de decisões informadas com base nas informações presentes nos indicadores.

Conclusão

A metodologia empregada neste estudo desempenhou um papel fundamental na organização e análise dos indicadores presentes no banco de dados do planejamento estratégico de 21 municípios participantes do Programa GPE. A análise minuciosa e sequencial resultou em uma representação mais precisa e estruturada dos indicadores, atuando na compreensão do cenário municipal.

Os procedimentos iniciais envolveram a coleta de informações textuais, seguida pela aplicação de expressões regulares para a padronização e normalização dos dados. Esse processo garantiu que as análises subsequentes fossem conduzidas de forma consistente e precisa. A identificação de indicadores com nomes semelhantes, bem como a eliminação de duplicações e redundâncias, contribuiu para uma base de dados mais limpa e confiável.

A análise de similaridade e o uso do método *K-Means* permitiram a formação de grupos de indicadores com base em suas características semânticas. Essa segmentação revelou padrões de relacionamento entre os indicadores e proporcionou uma visão mais clara das tendências e semelhanças nas denominações. A aplicação de técnicas de mineração de texto, aliada ao uso dos pacotes do software R, contribuiu para a eficiência e a precisão do processo.

Os resultados obtidos demonstram a importância da metodologia adotada para a organização, limpeza e análise dos indicadores, oferecendo conhecimentos significativos para a compreensão das abordagens de planejamento estratégico em diferentes municípios. Esse estudo exemplifica a aplicação bem-sucedida de técnicas de mineração de texto na área de gestão pública, oferecendo contribuições valiosas para a otimização do processo de análise de indicadores de desempenho.

Agradecimentos

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

Referências

DE OLIVEIRA, M.W.P.; NEDER, R.; RAMALHO, P.; MACIEL, C.; FREIRE, N.; PERES, J.; VUOLO, C.; ANJOS, A. MANSILLA, D. Indicators of Municipal Public Management: Study of Multiple Performance Measurement Systems. In: KŐ, A. et al. Electronic Government and the Information Systems Perspective. EGOVIS 2019. *Lecture Notes in Computer Science*, v. 11709. Cham: Springer, 2019. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-27523-5_9

FEINERER, I.; HORNIK, K. *tm: Text Mining Package*. R package v. 0.7-11, 2023. Disponível em: https://cran.r-project.org/package=tm

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

FITZPATRICK, J. L.; SANDERS, J. R.; WORTHEN, B. R. *Program Evaluation: alternative approaches and practical guidelines*. 4. ed. New Jersey: Pearson, 2011.

FRIEDL, J. E. F. *Mastering Regular Expressions*. Sebastopol: O'Reilly Media, 2006.

GRUBER, J. *LexisNexisTools. An R package for working with newspaper data from 'LexisNexis'*. R package v. 0.3.5, 2023. Disponível em: https://github.com/JBGruber/LexisNexisTools.

JANNUZZI, P. M. *Monitoramento e avaliação de programas sociais: uma introdução aos conceitos e técnicas*. Campinas, SP: Editora Alínea, 2016.

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, 2018.

JURAFSKY, D.; MARTIN, J. H. *Speech, and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3. ed. New Jersey: Pearson Education, 2020.

KAPLAN, R.; NORTON, D. *The Balanced Scorecard - Measures That Drive Performance*. Harvard Business Review, n. 01, pp.71-79, jan-fev, 1992.

KAPLAN, R.; NORTON, D. *Using the balanced scorecard as a strategic management system*. Harvard Business Review, n. 01, pp.75-85, jan-fev, 1996.

KIRCH, J. L.; SCHOENHERR, R. P.; VELOSO, T. C. M. A.; HONGYU, K. Aplicação da Análise de Componentes Principais e de Agrupamento para os Indicadores de Desempenho das Universidades Federais do Brasil. *Sigmae*, 8(2), 55-66, 2019. Disponível em: https://publicacoes.unifal-mg.edu.br/revistas/index.php/sigmae/article/view/920

LEE, J.; YOON, Y. Indicators development to support intelligent road infrastructure in urban cities. *Transport Policy*, v. 114, p.252-265, 2021.

MATO GROSSO. PDI – *Programa de Desenvolvimento Institucional Integrado*: TCE-MT, promovendo soluções inovadoras na linha de sua missão orientadora compartilha com os fiscalizados a experiência adquirida a partir da adoção do planejamento estratégico e de novas tecnologias, para a eficiência da administração pública. Cuiabá: TCE, 2013.

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: TCE, 2022.

MODRUŠAN, N.; MRŠIĆ, L.; RABUZIN, K. *Intelligent Public Procurement Monitoring System Powered by Text Mining and Balanced Indicators*. In: Hammoudi, S. et al. Data Management Technologies and Applications. DATA 2020. Communications in Computer and Information Science, v. 1446. Cham: Springer, 2021.

MOREIRA, D.; SANTINI, J. F. *Conectando Pesquisa a Gestão Municipal: Avaliações de Impacto Influenciam a Formação de Política Pública?* In: KOGA, Natália Massaco et al. Políticas Públicas e Usos de Evidências no Brasil: Conceitos, Métodos, Contextos e Práticas. 1. ed. Brasília: Ipea, 2022.

R CORE TEAM. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2022. Disponível em: https://www.R-project.org/.

RINKER, T. W. *textclean: Text Cleaning Tools,* v. 0.9.3. Buffalo, New York, 2018. Disponível em: https://github.com/trinker/textclean.

SHIBUI, S. *Visualization of indicators to enhance the internal quality assurance system in Japanese universities*. International Conference on Research in Education, Teaching and Learning, Paris, p. 16, 2018.

SILVA, I. P. *Indicadores Sociais e Sua Importância para a Gestão Pública Municipal*. 2022. 52 f. Trabalho de Conclusão de Curso (Especialização) - Universidade Federal Rural de Pernambuco, Especialização em Gestão Pública Municipal, Recife, 2022.

ZIZKA, J.; DARENA, F.; SVOBODA, A. *Text Mining with Machine Learning: Principles and Techniques*. 1. ed. Boca Raton: CRC Press, 2019.