Pós-Graduação *Lato Sensu* Modelagem em Ciência e Tecnologia - Turma 3
Trabalho de Conclusão de Curso aprovado em 13 de Junho 2025

Agrupamento em texto: análise semântica latente e *k-means* aplicados em dados de avaliação institucional

Clustering text data. Latent Semantic Analisys and *k-means* applied institutional evaluation

Carlos Bino de Souza
Universidade Federal de Alfenas - UNIFAL-MG
carlos.bino@sou.unifal-mg.edu.br

Deive Ciro de Oliveira
Universidade Federal de Alfenas - UNIFAL-MG
deive.oliveira@unifal-mg.edu.br

Resumo

Este artigo aborda o uso de técnicas de Processamento de Linguagem Natural (NLP) na análise automatizada de comentários qualitativos submetidos à Comissão Própria de Avaliação (CPA) da Universidade Federal de Alfenas (Unifal-MG). O objetivo é realizar a análise semântica automatizada dos comentários qualitativos enviados à CPA da Unifal-MG, aplicando técnicas de NLP e LSA, com SVD, seguidas da aplicação do algoritmo de clusterização *k-means*. Justifica-se a pesquisa pela dificuldade enfrentada pela CPA em sistematizar e interpretar, de forma objetiva e escalável, grandes volumes de dados textuais, geralmente tratados de maneira manual e subjetiva. Para tanto, foram aplicadas técnicas de NLP, Análise Semântica Latente (LSA), vetorização TF-IDF, redução de dimensionalidade com Decomposição por Valores Singulares (SVD), agrupamento por meio do algoritmo *k-means* e a visualização dos cluster por meio de nuvens de palavras. A definição do número de *clusters* baseou-se na análise do dendrograma. Os resultados revelaram dez agrupamentos que permitem identificar os assuntos recorrentes de cada *cluster*.

Palavras-chave

CPA, Processamento de Linguagem Natural, Análise Semântica Latente, Agrupamento, Avaliação Institucional

Abstract

This article addresses the use of Natural Language Processing (NLP) techniques in the automated analysis of qualitative comments submitted to the Self-Assessment Commission (CPA) of the Federal University of Alfenas (Unifal-MG). The objective is to perform automated semantic analysis of the qualitative comments sent to the Unifal-MG CPA, applying NLP and LSA techniques, with SVD, followed by the application of the k-means clustering algorithm. The research is justified by the difficulty faced by the CPA in objectively and scalably systematizing and interpreting large volumes of textual data, which are generally treated manually and subjectively. To this end, NLP techniques, Latent Semantic Analysis (LSA), TF-IDF vectorization, dimensionality reduction with Singular Value Decomposition (SVD), clustering through the k-means algorithm, and the visualization of the clusters through word clouds were applied. The definition of the number of clusters was based on the analysis of the dendrogram. The results revealed ten groupings that allow for the identification of the recurring topics within each cluster.

Keywords

CPA, Natural Language Processing, Latent Semantic Analysis, Clustering, Institutional Assessment

1 Introdução

A avaliação institucional é um dos principais instrumentos para garantir a qualidade da educação superior no Brasil. Regulada pelo Sistema Nacional de Avaliação da Educação Superior (SINAES), instituído pela Lei nº 10.861/2004, a avaliação busca fornecer um diagnóstico preciso do desempenho das instituições, contribuindo para o aprimoramento contínuo das políticas educacionais Brasil (2004). Nesse contexto, a Comissão Própria de Avaliação (CPA) é responsável por conduzir processos de autoavaliação institucional e sistematizar informações que subsidiam a tomada de decisões acadêmicas e administrativas (Universidade Federal de Alfenas, 2018).

Um dos mecanismos utilizados pela CPA para coletar dados sobre a percepção da comunidade acadêmica são os formulários de avaliação, nos quais estudantes, docentes e técnicos adminis-

trativos expressam suas opiniões sobre diferentes aspectos da instituição. No entanto, a análise desses comentários é, muitas vezes, realizada de forma manual e subjetiva, tornando difícil a extração de padrões significativos. Essa dificuldade pode comprometer a efetividade das ações corretivas e de melhoria propostas pelas universidades, uma vez que a interpretação dos dados depende da capacidade individual dos avaliadores e da estrutura disponível para o processamento dessas informações.

Nesse cenário, a análise manual dos comentários enviados à CPA da Universidade Federal de Alfenas-MG (Unifal-MG) revela-se limitada diante do volume, diversidade e subjetividade das manifestações textuais coletadas nos formulários de autoavaliação. A ausência de padronização, aliada à expressividade espontânea dos respondentes, dificulta a sistematização das informações e a extração de padrões significativos, comprometendo a efetividade das ações institucionais baseadas nesses dados. Nesse contexto, o Processamento de Linguagem Natural (NLP - sigla em inglês *Natural Language Processing*) surge como uma abordagem promissora para o tratamento computacional de textos em linguagem natural, viabilizando a análise objetiva e escalável de conteúdos discursivos.

Entre as técnicas de NLP aplicáveis, destaca-se a Análise Semântica Latente (LSA - sigla em inglês de *Latent Semantic Analysis*), que projeta documentos e termos em espaços vetoriais de baixa dimensionalidade por meio da Decomposição de Valores Singulares (SVD - sigla em inglês *Singular Value Decomposition*), permitindo identificar estruturas semânticas ocultas e relações latentes de significado. Essa abordagem é especialmente relevante para organizar comentários institucionais em temas recorrentes e semanticamente similares, superando as limitações das análises baseadas apenas em frequência de palavras. Ao incorporar tais técnicas, a CPA amplia sua capacidade de interpretar percepções da comunidade acadêmica de forma mais sistemática e orientada por dados.

Complementando as etapas de pré-processamento e análise semântica, o agrupamento (ou clusterização) desponta como técnica essencial para a organização automatizada dos comentários em grupos tematicamente coesos. Por meio de algoritmos de aprendizado não supervisionado, como o *k-means*, é possível segmentar os documentos textuais com base em suas representações vetoriais reduzidas, agrupando aqueles que compartilham padrões latentes de significado. Essa abordagem permite não apenas identificar temas recorrentes nas percepções da comunidade acadêmica, mas também estruturar o corpo textual em conjuntos semanticamente interpretáveis, ampliando a capacidade de extração de conhecimento dos dados avaliativos. Quando combinada com técnicas como Frequência do Termo – Frequência Inversa nos Documentos (TF-IDF, sigla em inglês de *Term Frequency–Inverse Document Frequency*) e SVD, a clusterização proporciona uma análise mais profunda e orientada, revelando nuances discursivas que podem passar despercebidas em processos avaliativos convencionais baseados apenas na leitura humana.

Dessa forma, este estudo tem como objetivo central realizar a análise semântica automatizada dos comentários qualitativos enviados à CPA da Unifal-MG, aplicando técnicas de NLP e LSA, com SVD, seguidas da aplicação do algoritmo de clusterização *k-means*. Todos os procedimentos metodológicos serão implementados na linguagem *Python*, utilizando bibliotecas específicas para manipulação textual, vetorização e agrupamento de dados. Busca-se, com isso, identificar padrões discursivos e temas latentes expressos pela comunidade acadêmica, organizando-os em agrupamentos semanticamente coesos que possam subsidiar ações institucionais baseadas em evidências. Especificamente, o estudo propõe: (i) aplicar métodos de pré-processamento textual, incluindo correção ortográfica, lematização, remoção de *stopwords*, tokenização e filtragem lexical; (ii) representar os comentários por meio de uma matriz TF-IDF; (iii) reduzir a dimensionalidade da matriz utilizando SVD; (iv) agrupar os comentários com base nas estruturas semânticas extraídas, utilizando o algoritmo *k-means*; e (v) avaliar a qualidade dos agrupamentos formados com métricas como o coeficiente de silhoeta.

2 Revisão Bibliográfica

2.1 A Comissão Própria de Avaliação e a Regulação da Educação Superior no Brasil

A avaliação institucional no ensino superior brasileiro é regida por um arcabouço legal que estabelece diretrizes para a regulação, supervisão e autoavaliação das instituições de ensino. A CPA desempenha um papel essencial nesse contexto, garantindo que os processos avaliativos sejam conduzidos de maneira autônoma e com a participação da comunidade acadêmica. As referências selecionadas abordam diferentes aspectos desse processo, desde a normatização da au-

toavaliação até a regulação do ensino superior.

A Resolução nº 24/2018 da Unifal-MG estabelece o Regimento Interno da CPA[1], atribuindo-lhe a responsabilidade de coordenar os processos de avaliação institucional, promover a cultura avaliativa, sistematizar os dados coletados e utilizá-los na melhoria contínua da universidade (Universidade Federal de Alfenas, 2018). A resolução também reforça a atuação autônoma da comissão, a participação dos diversos segmentos da comunidade acadêmica e a divulgação transparente dos resultados para embasar a gestão acadêmica e administrativa. Em nível nacional, o Decreto nº 9.235/2017 define as diretrizes para regulação, supervisão e avaliação das instituições de ensino superior, estabelecendo o papel do Ministério da Educação na fiscalização e na avaliação periódica de instituições e cursos (Brasil, 2017). O decreto também valoriza a autoavaliação institucional e a atuação da CPA como mecanismos fundamentais para a qualidade do ensino, complementando as diretrizes do SINAES ao determinar que os resultados da autoavaliação sejam utilizados no planejamento estratégico e na conformidade com as políticas educacionais do país.

Além disso, a Lei nº 10.861, de 14 de abril de 2004, institui o SINAES e formaliza a obrigatoriedade da autoavaliação institucional como um dos pilares do sistema de avaliação da educação superior (Brasil, 2004). A lei estabelece que a avaliação das instituições deve abranger diferentes dimensões, incluindo a missão e o desenvolvimento institucional, a responsabilidade social, a política de ensino e pesquisa, a infraestrutura, entre outros aspectos. A CPA deve atuar de forma contínua para identificar pontos fortes e desafios enfrentados pela instituição, propondo medidas que contribuam para a melhoria da qualidade educacional. A Lei nº 10.861/2004 também define os instrumentos de avaliação que devem ser utilizados, incluindo os formulários aplicados à comunidade acadêmica, cujos resultados subsidiam as políticas institucionais.

Assim, a Resolução nº 24/2018 da Unifal-MG, o Decreto nº 9.235/2017 e a Lei nº 10.861/2004 formam o alicerce normativo para a atuação da CPA e reforçam a relevância da autoavaliação como ferramenta de gestão e desenvolvimento institucional.

Diante disso, é cada vez mais importante aperfeiçoar a análise dos dados qualitativos coletados pela CPA, como comentários e sugestões da comunidade acadêmica. O uso de técnicas de NLP com clusterização permite identificar padrões e temas recorrentes de forma objetiva, ampliando a capacidade da CPA de interpretar as percepções institucionais e fortalecendo uma gestão mais atuante, transparente e alinhada às reais demandas da universidade.

2.2 Trabalhos relacionados

Em nossas investigações, identificamos que pesquisas voltadas à Mineração de Textos com uso da LSA, associada a técnicas de clusterização, apresentam significativa proximidade com os objetivos e estratégias adotados no presente trabalho. Esses estudos correlatos evidenciam abordagens computacionais que buscam identificar padrões semânticos ocultos em grandes volumes textuais (*corpora*, plural de *corpus*), de modo semelhante ao que propomos ao aplicar LSA na análise de padrões de *software*. A utilização da LSA como ferramenta para redução de dimensionalidade e identificação de similaridades conceituais, combinada à clusterização para agrupar documentos semanticamente similares, tem se mostrado eficaz em diversos contextos de mineração de informação textual. Considerando essa convergência metodológica, passamos a apresentar, a seguir, os procedimentos metodológicos e os principais resultados obtidos em alguns dos trabalhos mais representativos e correlatos à nossa proposta de pesquisa.

No estudo de Marcolin et al. (2019), foram aplicados métodos computacionais de mineração de texto para analisar os discursos dos deputados durante a votação do impeachment da presidenta Dilma Rousseff, com o objetivo de identificar temas latentes e padrões de argumentação. Inicialmente, os dados textuais passaram por um rigoroso pré-processamento com *scripts* em R, incluindo a remoção de acentos, caracteres especiais, *stopwords* e a padronização para letras minúsculas. Em seguida, construiu-se uma matriz TF-IDF. Com base nessa matriz, foi aplicada a LSA, utilizando SVD para identificar estruturas semânticas ocultas e organizar os discursos em tópicos. Além disso, foi realizada uma análise de correlação entre termos, destacando palavras associadas ao conceito de "corrupção". Os resultados revelaram que não houve distinção clara entre os discursos dos deputados favoráveis e contrários ao impeachment, sugerindo que o conteúdo discursivo não foi suficiente para prever o posicionamento político, evidenciando a presença de dissonância cognitiva e a influência de fatores como ideologia e interesses pessoais na tomada de de-

[1]Todos os documentos referentes à CPA da Unifal-MG, incluindo Regimento, instrumentos de avaliação e relatórios, estão disponível no endereço https://www.unifal-mg.edu.br/cpa/

cisão.

Castro (2006) aplicou métodos computacionais de Mineração de Texto e LSA com o objetivo de automatizar a descoberta de relacionamentos entre padrões de *software* armazenados em repositórios digitais. O processo iniciou-se com uma etapa detalhada de pré-processamento textual. Nessa etapa, os textos extraídos dos campos dos padrões – como *Nome*, *Contexto*, *Problema*, *Solução* e *Padrões Relacionados* – foram normalizados por meio da remoção de *stopwords*, padronização ortográfica, conversão para minúsculas e eliminação de caracteres especiais. Além disso, a estrutura dos padrões, baseada em seus *templates*, foi preservada para garantir que os significados contextuais fossem mantidos durante a vetorização. Após essa preparação, os dados foram organizados em uma matriz TF-IDF, que passou pela SVD. A seguir, foi aplicada a análise de similaridade semântica, utilizando o cosseno do ângulo entre vetores para medir o grau de proximidade entre os padrões. Por fim, Castro incorporou regras de associação estruturais, baseadas nos vínculos semânticos entre campos específicos (por exemplo, entre o "Contexto" de um padrão e o "Resultado" de outro), para validar e reforçar os relacionamentos detectados. Como resultado, a proposta foi capaz de identificar automaticamente conexões relevantes entre padrões, superando limitações de abordagens tradicionais e contribuindo para um reuso mais eficiente de soluções em engenharia de *software*.

A dissertação de Amaral (2021) propõe um modelo computacional para seleção de tópicos relevantes em documentos aplicados ao *compliance*, utilizando técnicas de NLP e aprendizagem não supervisionada. O objetivo foi automatizar a análise de documentos normativos por meio da mineração de texto, estruturando o trabalho conforme o modelo CRISP-DM (sigla em inglês para Processo Padrão Inter-Indústrias para Mineração de Dados). O pré-processamento textual envolveu remoção de *stopwords*, lematização e vetorização com TF-IDF. Na modelagem, foram utilizados os métodos *K-means*, LSA e Alocação de Dirichlet Latente (LDA - *Latent Dirichlet Allocation*). O LSA foi aplicado para redução de dimensionalidade via decomposição em valores singulares, enquanto o LDA identificou os tópicos latentes em cada base textual. Os testes, realizados com relatórios da SCGE-PE[2], acórdãos do TCU[3] e leis europeias, mostraram que a combinação LSA-LDA produziu os melhores agrupamentos, avaliados por métricas como coeficiente de Silhoueta e coerência semântica. Os melhores resultados foram observados na base de leis europeias, indicando que o modelo proposto é eficaz na extração automatizada de conhecimento em textos normativos e pode contribuir significativamente para os processos de verificação de conformidade.

Na dissertação de Rezende (2019), foi desenvolvido um sistema de mineração de patentes com foco na identificação de tendências tecnológicas e análise de similaridade textual. A metodologia envolveu coleta de dados do USPTO (sigla em inglês para Escritório de Patentes e Marcas dos Estados Unidos), pré-processamento dos textos (tokenização, remoção de *stopwords* e *stemming*) e vetorização por meio da matriz TF-IFD. Essa matriz foi utilizada como base para a aplicação do algoritmo LSA. Além da LSA, foram empregados os algoritmos *Word2Vec* e *Word Mover's Distance* (WMD) para comparação entre resumos de patentes. Os resultados demonstraram que o sistema foi eficaz tanto na prospecção tecnológica, por meio da Curva S e nuvens de palavras, quanto na ranqueação por similaridade, oferecendo suporte ao mapeamento e à análise estratégica de documentos de propriedade intelectual.

Os estudos analisados compartilham uma estrutura metodológica convergente, baseada em técnicas de Mineração de Texto e LSA, articulada em quatro etapas fundamentais: pré-processamento textual, vetorização com TF-IDF, aplicação da LSA via SVD e análise de similaridade semântica. O pré-processamento é rigoroso e essencial para reduzir ruídos e padronizar a linguagem, incluindo a remoção de *stopwords*, eliminação de acentos e caracteres especiais, normalização ortográfica e conversão para minúsculas (Marcolin et al., 2019; Castro, 2006; Amaral, 2021; Rezende, 2019). Algumas abordagens ainda incorporam técnicas como *stemming* ou lematização, que reduzem palavras à forma canônica (Amaral, 2021; Rezende, 2019). Após essa etapa, os textos são vetorizados por meio da matriz TF-IDF, que atribui maior peso a termos distintivos com base em sua frequência e raridade, possibilitando uma representação matemática mais informativa dos dados textuais (Marcolin et al., 2019; Amaral, 2021).

Com os dados vetorizados, aplica-se a LSA, que utiliza a SVD para reduzir a dimensionalidade da matriz TF-IDF e revelar estruturas semânticas latentes, identificando relações ocultas entre termos e documentos (Castro, 2006; Rezende, 2019). Essa projeção em um espaço

[2]Secretaria da Controladoria Geral do Estado de Pernambuco

[3]Tribunal de Contas de União

de menor dimensão ajuda a captar a essência dos textos e eliminar redundâncias. Na etapa final, os estudos realizam análise de similaridade semântica com o objetivo de comparar documentos, identificar padrões e agrupar conteúdos semanticamente próximos. Essa análise se dá por correlação entre vetores, similaridade cosseno, algoritmos de agrupamento como *K-means* ou métodos de ranqueamento (Marcolin et al., 2019; Castro, 2006; Amaral, 2021; Rezende, 2019). Em conjunto, essas etapas formam uma cadeia metodológica robusta e adaptável, capaz de extrair conhecimento profundo de *corpora* não estruturados, com aplicações em diferentes domínios como política, tecnologia, direito e ciência da informação.

Vamos apresentar estudos relacionados que exploram técnicas de agrupamento aplicadas a conjuntos de dados que podem ser representados geometricamente em espaços vetoriais. O objetivo é destacar abordagens que se beneficiam da estrutura espacial dos dados para identificar padrões, similaridades e segmentações por meio de métodos de aprendizado não supervisionado, com ênfase em representações vetoriais que permitem a aplicação eficiente de algoritmos de clusterização.

No trabalho desenvolvido por Pilatti (2023), foram aplicados métodos computacionais baseados em técnicas de aprendizado de máquina não supervisionado, com ênfase na análise de agrupamento de dados. O objetivo foi realizar a segmentação comportamental de usuários de cartão de crédito a partir de um conjunto de dados contendo 8950 registros, considerando 18 atributos comportamentais. Inicialmente, foi conduzida uma etapa de pré-processamento fundamental, que envolveu a exclusão de atributos irrelevantes (como o identificador do cliente), imputação de valores ausentes pela média e a normalização dos dados por padronização (*Z-score*), garantindo que todos os atributos tivessem média zero e desvio padrão igual a um, evitando que atributos com diferentes escalas influenciassem os resultados do agrupamento. O autor adotou o algoritmo *K-means* para agrupar os clientes com base em suas similaridades, conforme as distâncias entre pontos e centróides em um espaço vetorial de múltiplas dimensões. A escolha do número ideal de clusters (K) foi realizada por meio do *método do cotovelo*, que identifica o ponto de inflexão na curva da soma dos quadrados intra-*cluster* (*WCSS*), conforme descrito por Bholowalia & Kumar (2014, *apud* Pilatti, 2023). Além disso, foram utilizados dois índices de validação interna: o coeficiente de silhoeta, que mede o quão bem cada ponto está ajustado ao seu *cluster*, e o índice de Davies-Bouldin, que avalia a separação e coesão dos grupos formados (Rousseeuw, 1987; Gustriansyah et al., 2020). Esses métodos permitiram identificar que o agrupamento mais adequado foi aquele composto por três *clusters*, possibilitando *insights* sobre padrões de gastos, uso de crédito e comportamento de pagamento dos clientes. A aplicação do algoritmo *K-means*, em conjunto com os métodos de validação, demonstrou ser eficaz na tarefa de segmentação e descoberta de perfis distintos dentro da base analisada, evidenciando o potencial das técnicas de mineração de dados no apoio à tomada de decisões em contextos comerciais.

Com o objetivo de desenvolver um Índice de Qualidade de Água Subterrânea (IQA_{SUB}) aplicável a áreas com potencial minerário, Ribeiro (2024) estruturou uma metodologia que integra técnicas de pré-processamento, agrupamento e análise multivariada para avaliar a qualidade de águas subterrâneas em poços monitorados pelo Instituto Mineiro de Gestão das Águas (IGAM). A autora trabalhou com uma base de 795 registros oriundos das redes Guarani, Norte de Minas, Velhas e do Projeto Águas do Norte de Minas (PANM), realizando uma pré-seleção dos parâmetros com base em sua frequência e relevância normativa. O pré-processamento envolveu a padronização dos dados pelo método de *Z-score*, garantindo comparabilidade entre variáveis de diferentes escalas. Para o agrupamento dos dados, foi aplicada a técnica de clusterização *k-means* no ambiente *RStudio*, com a definição do número ótimo de *clusters* pelo método do cotovelo (*elbow method*), o que resultou na formação de três grupos: dois com qualidade "péssima" e um com qualidade "regular", segundo os escores calculados pelo IQA_{SUB}. A avaliação da estrutura dos dados foi complementada pela Análise de Componentes Principais (PCA - *Principal Component Analysis*), que explicou 79,9% da variância total com cinco componentes principais. Os resultados evidenciaram que o IQA_{SUB} apresentou classificações mais conservadoras do que o índice canadense IQACCME[4], reflexo da ponderação atribuída aos parâmetros, e que os agrupamentos obtidos foram coerentes com as classificações fornecidas pelo índice proposto, validando sua eficácia como ferramenta de apoio à gestão da qualidade da água subterrânea em regiões minerárias.

O trabalho de Silva (2021) teve como objetivo identificar similaridades nos preços da ga-

[4]Sigla em inglês para Conselho Canadense de Ministros do Meio Ambiente

solina comum e do etanol praticados em postos de combustíveis na cidade de Campina Grande – PB, durante o ano de 2019. Para isso, o autor aplicou técnicas de análise multivariada, com destaque para a análise de *cluster*, utilizando métodos computacionais na Linguagem R. O pré-processamento dos dados consistiu na verificação de registros faltantes e inconsistências, seguido por análise descritiva e geração de gráficos exploratórios. Em seguida, foram aplicados o teste de normalidade de *Shapiro-Wilk* e o cálculo do coeficiente de correlação de *Spearman*, apropriado devido à não normalidade da variável gasolina comum. A análise de agrupamento foi realizada por dois métodos complementares: o método não hierárquico, por meio do algoritmo *k-means*, e o método hierárquico, utilizando distância euclidiana e ligação completa (*complete linkage*). Para determinar o número ideal de grupos, foi utilizado o método do cotovelo (*Elbow Method*), que indicou a formação de três clusters. O estudo revelou que os clusters formados por ambos os métodos foram consistentes entre si, o que reforça a estabilidade da classificação. Como resultado, observou-se que os postos pertencentes ao cluster 1 apresentaram os menores preços médios para gasolina e etanol, enquanto os do cluster 3 exibiram os maiores valores. A convergência entre os dois métodos na composição dos grupos indica a adequação das técnicas aplicadas à estrutura dos dados analisados.

Os trabalhos de Pilatti (2023), Ribeiro (2024) e Silva (2021) evidenciam a versatilidade do algoritmo *k-means*, aplicável em distintos contextos como análise comportamental de consumidores, avaliação da qualidade da água subterrânea e estudo de preços de combustíveis. Em comum, os três autores adotaram o pré-processamento com padronização por *Z-score*, o método do cotovelo para definição do número ótimo de clusters e a aplicação do *k-means* como técnica de agrupamento baseada em distância euclidiana, demonstrando a eficácia da abordagem em tarefas exploratórias com dados multivariados. As divergências entre os estudos residem no domínio de aplicação e nas estratégias de validação: enquanto Pilatti (2023) utilizou coeficientes de silhoeta e *Davies-Bouldin* para avaliar a coesão e separação dos grupos em um contexto comercial, Ribeiro (2024) integrou a clusterização à PCA para interpretar padrões ambientais em poços monitorados, e Silva (2021) comparou os resultados do *k-means* com os de um método hierárquico, reforçando a robustez da segmentação em dados de natureza econômica. Essas abordagens demonstram que, embora a base metodológica seja semelhante, o *k-means* pode ser adaptado e validado de forma diversa conforme os objetivos, as características da base de dados e o contexto de aplicação.

Diante da análise dos trabalhos correlatos e da constatação da eficácia das técnicas de análise semântica e clusterização em diferentes domínios, evidencia-se a importância de aprofundar as etapas que fundamentam essas abordagens. Em especial, o NLP se apresenta como componente metodológico essencial para a transformação de dados textuais não estruturados em representações vetoriais compatíveis com métodos computacionais de análise.

Os estudos analisados evidenciam a eficácia de técnicas de análise semântica combinadas com métodos de clusterização para a identificação de padrões discursivos em *corpora*. Apesar da diversidade de contextos — que inclui política, *compliance*, qualidade ambiental e consumo — observa-se uma estrutura metodológica comum baseada na vetorização TF-IDF, redução de dimensionalidade via SVD e agrupamento com o algoritmo *k-means*. No entanto, poucos trabalhos aplicaram essas estratégias no âmbito da avaliação institucional do ensino superior, especialmente no tratamento de comentários espontâneos submetidos à CPA. Assim, este estudo propõe preencher essa lacuna ao adaptar esse arcabouço metodológico ao contexto da Unifal-MG, com foco na extração automatizada de temas latentes a partir da linguagem natural da comunidade acadêmica, contribuindo tanto para a gestão universitária quanto para a formação docente em uma perspectiva interdisciplinar.

2.3 Processamento de Linguagem Natual

O NLP é um campo interdisciplinar dedicado à investigação e à proposição de métodos e sistemas para o tratamento computacional da linguagem humana. Segundo Caseli et al. (2024), o termo "natural" refere-se às línguas humanas, em oposição a outras linguagens formais, e o NLP está intimamente relacionado à Inteligência Artificial (IA) e à Linguística Computacional. A área é tradicionalmente dividida em duas subáreas principais: a Interpretação ou Compreensão de Linguagem Natural (NLU – *Natural Language Understanding*), que se ocupa da análise e interpretação de textos, e a Geração de Linguagem Natural (NLG – *Natural Language Generation*), voltada à produção automática de linguagem a partir de dados estruturados.

O NLP pode ser desenvolvido com base em diferentes paradigmas, que evoluíram ao longo do tempo. Inicialmente, predominou o paradigma

simbólico, no qual o conhecimento linguístico é representado por meio de formalismos explícitos, como léxicos, regras e linguagens lógicas, compreensíveis ao ser humano (Caseli et al., 2024, p. 14). Nessa abordagem, cada aspecto da língua é descrito manualmente, exigindo a atuação de especialistas para formular regras voltadas à concordância verbal, à formação do plural ou à construção de sentenças interrogativas. A partir da década de 1990, com o aumento da capacidade computacional, surgiu o paradigma estatístico, que se baseia na extração de padrões a partir de grandes *corpora*, modelando os fenômenos linguísticos com base na frequência de ocorrência de exemplos reais. Nesse paradigma, as regras explícitas são substituídas por inferências extraídas dos dados, permitindo que o sistema capte regularidades linguísticas de forma automática, sem a necessidade de codificação manual por especialistas. Mais recentemente, com o avanço das redes neurais profundas, consolidou-se o paradigma neural, caracterizado pela aprendizagem de representações numéricas da linguagem por meio de arquiteturas com múltiplas camadas, capazes de generalização, mesmo que seja difícil compreender o funcionamento interno.

Reconhecendo as limitações de cada abordagem, foi introduzido o paradigma híbrido, que combina elementos do paradigma simbólico com abordagens estatísticas ou neurais, buscando equilibrar flexibilidade de modelagem e explicitação de conhecimento. Essa combinação favorece a interpretabilidade dos modelos e amplia a robustez dos sistemas de NLP (Caseli et al., 2024, p. 15).

O desenvolvimento de projetos de NLP segue etapas bem definidas que, segundo Madureira (2024a, p. 286), são descritas da seguinte forma:

- **Escolha da Aplicação**: Definição clara do objetivo e do escopo do sistema a ser desenvolvido, etapa que orienta todo o processo subsequente.
- **Caracterização dos Fenômenos Linguísticos**: Identificação dos fenômenos linguísticos relevantes, como morfologia, sintaxe, semântica ou pragmática, a serem tratados no sistema.
- **Seleção das Teorias Linguísticas**: Escolha das bases teóricas que embasarão a modelagem linguística, etapa especialmente importante em abordagens simbólicas.
- **Desenvolvimento e Teste de Algoritmos**: Implementação dos métodos de processamento — simbólicos, estatísticos, neurais ou híbridos — e realização de testes para avaliar o desempenho inicial.
- **Implementação de Versões Progressivas**: Construção de versões sucessivas do sistema, com melhorias contínuas em eficiência, robustez e cobertura linguística, até alcançar condições ideais para uso prático.

Embora os modelos neurais predominem em muitas aplicações atuais, salientamos que, na maioria das técnicas contemporâneas, os sistemas de NLP não compreendem o significado semântico pleno das expressões linguísticas. Em geral, eles reconhecem apenas padrões de coocorrência entre palavras e frases. Modelos de aprendizado profundo, por exemplo, não captam que o termo "casa"se refere a um local de moradia, limitando-se a reproduzir padrões extraídos dos dados de treinamento (Caseli et al., 2024, p. 15). Nesse contexto, o NLP moderno busca integrar abordagens simbólicas, estatísticas e neurais para enfrentar a complexidade da linguagem natural, promovendo o desenvolvimento de sistemas mais explicáveis, adaptáveis e robustos.

Dentre as técnicas aplicáveis à construção desses sistemas, aquelas voltadas à análise semântica exercem um papel central na interpretação de textos. Após a coleta de dados, que pode incluir o uso de estratégias como *web crawling* e *web scraping*[5], realiza-se o **pré-processamento textual**, etapa que compreende a **tokenização**, a **normalização ortográfica**, a **remoção de pontuação** e a **eliminação de palavras irrelevantes**, conhecidas como *stopwords* (Madureira, 2024b, p. 419).

Superado esse nível preliminar, a análise semântica permite explorar camadas mais profundas de significado. Técnicas de extração de informação, como o de entidades nomeadas e a identificação de eventos, contribuem para a representação estruturada do texto, favorecendo a detecção de relações e interpretações contextuais (Madureira, 2024b, p. 422). Além disso, a resolução de correferência aprimora a coesão textual ao identificar diferentes expressões que se referem à mesma entidade.

A sumarização automática, seja por métodos extrativos ou abstrativos, é igualmente relevante, pois permite a condensação de grandes volumes de conteúdo textual preservando os núcleos semânticos essenciais (Paes & Freitas, 2024,

[5]O *web crawling* refere-se à navegação automatizada por sites para localizar e indexar conteúdos disponíveis, enquanto o *web scraping* diz respeito à extração seletiva de informações específicas desses conteúdos, organizando-as em estruturas utilizáveis por sistemas de NLP.

p. 425). A classificação de textos complementa esse processo ao viabilizar a organização de documentos por categorias temáticas, padrões opinativos ou domínios de aplicação, com base em modelos capazes de reconhecer estruturas de significado.

Técnicas de predição e modelagem de tendências textuais também podem ser empregadas para identificar padrões de significação recorrentes em grandes corpora. Em casos que envolvem informações sensíveis, a utilização de métodos de anonimização textual torna-se essencial para preservar as relações semânticas enquanto se protege a identidade de indivíduos ou instituições mencionadas (Moreira, 2024, p. 463).

Essas técnicas, empregadas de maneira articulada, compõem um arcabouço metodológico robusto para a construção de sistemas capazes de interpretar a linguagem natural em níveis que superam sua estrutura superficial, promovendo a extração de significados latentes e contribuindo significativamente para a compreensão computacional da linguagem.

A adoção de estratégias inspiradas nas técnicas de NLP aqui apresentadas mostra-se especialmente pertinente para a análise dos comentários enviados à CPA. A partir de métodos como o pré-processamento textual, a extração de entidades e eventos, a classificação temática e a anonimização de conteúdos sensíveis, torna-se possível estruturar e interpretar os dados qualitativos de forma mais sistemática e eficiente. Essas abordagens favorecem a identificação de padrões discursivos, tendências de percepção e áreas críticas apontadas pela comunidade acadêmica, ajudando a CPA a analisar melhor as informações e a propor ações para a instituição que sejam mais bem fundamentadas e que respondam melhor às necessidades da comunidade.

2.3.1 Tokenização

A tokenização, uma das etapas do pré-processamento no NLP, é responsável por segmentar uma sequência de caracteres em unidades linguísticas mínimas denominadas *tokens*. Esses tokens podem corresponder a palavras, sinais de pontuação ou outros elementos com relevância interpretativa para tarefas computacionais de análise textual (Finatto et al., 2024, p. 78).

No português, a separação baseada em espaços e pontuações é uma estratégia comum, mas insuficiente para lidar com todas as estruturas linguísticas da língua. Expressões como "da"e "pela"geralmente são descontraídas corretamente para "de"+ "a"e "por"+ "a", mas há casos ambíguos em que essa segmentação gera interpretações incorretas. Por exemplo, a palavra "pelo"pode representar o substantivo ou a preposição "por"combinada com o artigo "o". Situações semelhantes ocorrem com "consigo", que pode ser pronome reflexivo ou forma verbal, e com palavras hifenizadas como "sinto-me", que podem ser decompostas em três tokens distintos: "sinto", -"e "me"(Finatto et al., 2024, p. 79).

Para lidar com palavras raras ou desconhecidas (*out-of-vocabulary*), os sistemas modernos de NLP adotam a tokenização em subpalavras, que fragmenta os vocábulos em unidades menores. Essa técnica permite que qualquer palavra seja representada a partir de um vocabulário reduzido, melhorando a generalização do modelo e reduzindo o número de palavras ausentes. Por exemplo, "felizmente"pode ser decomposta em "feliz"+ "mente", e "desfazer"em "de"+ "s"+ "fazer"(Finatto et al., 2024, p. 80).

No contexto institucional, a tokenização assume um papel estratégico na análise automatizada dos comentários enviados à CPA, pois constitui a etapa inicial que viabiliza a extração de informações linguísticas a partir de dados textuais brutos. Ao segmentar adequadamente os textos em unidades interpretáveis, a tokenização permite que sistemas de NLP identifiquem padrões de opinião, agrupem termos semanticamente relevantes e reconheçam estruturas discursivas recorrentes. A precisão dessa etapa impacta diretamente a eficácia de análises posteriores, como a identificação de temas latentes, a categorização de percepções da comunidade acadêmica e a geração de relatórios com base em evidências linguísticas. Assim, investir em estratégias de tokenização sensíveis às particularidades do português brasileiro é essencial para transformar comentários livres em insumos confiáveis para a gestão institucional e para o aprimoramento contínuo da qualidade educacional.

2.3.2 Normalização

A normalização textual é outra etapa no pré-processamento de dados linguísticos em sistemas de NLP, tendo como objetivo a padronização formal dos textos, a fim de reduzir sua variabilidade superficial e promover consistência na análise computacional. Segundo Finatto et al. (2024, p. 81), "a normalização é a tarefa que converte as palavras para alguma forma padrão". Essa padronização permite que diferentes formas de uma mesma expressão ou conceito sejam tra-

tadas de maneira uniforme pelo sistema.

As etapas da normalização envolvem tanto transformações gráficas quanto linguísticas. Entre os procedimentos formais estão: a conversão de todos os caracteres para caixa baixa, a remoção de sinais de pontuação e acentuação, e a substituição de abreviações por suas formas completas. Além disso, a normalização também compreende processos como a lematização, que associa palavras flexionadas à sua forma de dicionário (por exemplo, "somos"→ "ser"), e a radicalização, que busca identificar o radical comum entre palavras derivadas (como "retrabalho"→ "trabalho") (Finatto et al., 2024, p. 81).

Apesar de seu papel estratégico, a normalização apresenta desafios importantes, especialmente quando aplicada a textos informais e contextos variados. Um dos principais riscos é a aplicação inadequada de regras de substituição, que pode gerar interpretações equivocadas. As autoras exemplificam com o caso da abreviação "rs", que pode representar riso em mensagens informais, mas cuja substituição irrestrita pode levar a distorções como "youtuberisos" no lugar de "youtubers" (Finatto et al., 2024, p. 81). Para evitar esses erros, é fundamental empregar mecanismos sensíveis ao contexto e realizar substituições apenas sobre *tokens* identificados como relevantes.

Os benefícios da normalização incluem a simplificação da estrutura textual, a redução do vocabulário efetivo e a mitigação de ruídos linguísticos. Esses benefícios favorecem diretamente tarefas posteriores do NLP, como análise sintática, lematização, classificação de textos e extração de informações. Conforme destacam (Finatto et al., 2024), o uso adequado da normalização contribui para tornar os sistemas computacionais mais robustos, eficazes e sensíveis às variações da linguagem natural.

Dessa forma, a normalização textual não apenas otimiza o processamento linguístico automático, como também amplia a capacidade dos sistemas de compreender, categorizar e extrair informações relevantes de grandes volumes de dados textuais.

2.3.3 Análise Semântica

A análise semântica, no contexto da representação vetorial de texto, é uma abordagem que busca retratar o significado dos documentos ou palavras por meio de vetores em um espaço multidimensional, baseando-se nos padrões de coocorrência entre termos em um *corpus* de treinamento (Seno et al., 2024, p. 192). Essa abordagem fundamenta-se na hipótese distribucional, segundo a qual palavras que ocorrem em contextos linguísticos semelhantes tendem a apresentar significados próximos (Seno et al., 2024, p. 190).

A representação inicial é realizada mediante a construção de uma matriz termo-documento, na qual cada linha representa um termo e cada coluna corresponde a um documento, preenchida com valores relacionados à frequência de ocorrência de termos (Seno et al., 2024, p. 193). Para aprimorar essa representação, aplica-se o modelo TF-IDF, que pondera a importância de cada termo considerando tanto sua frequência em um documento específico quanto sua raridade em toda a coleção textual. Dessa forma, termos mais relevantes em documentos específicos e menos frequentes no *corpus* recebem pesos mais elevados (Seno et al., 2024, p. 197).

Entretanto, as matrizes TF-IDF tendem a ser extremamente esparsas, contendo muitos elementos nulos, o que aumenta a complexidade computacional. Para atenuar esse problema e capturar relações semânticas subjacentes, emprega-se a técnica SVD (Seno et al., 2024, p. 199). A SVD decompõe a matriz original em três matrizes fatoradas, permitindo a identificação dos eixos principais de variação e possibilitando a reconstrução de uma versão reduzida da matriz inicial, preservando os k maiores valores singulares (Seno et al., 2024, p. 200).

A Análise Semântica Latente (LSA – Latent Semantic Analysis) é, então, utilizada para projetar termos e documentos em um espaço vetorial de baixa dimensionalidade, eliminando ruídos e realçando padrões semânticos relevantes. Essa projeção permite agrupar palavras e documentos semanticamente similares em regiões próximas no espaço vetorial (Seno et al., 2024, p. 199), possibilitando a inferência de relações semânticas ocultas, mesmo entre termos que raramente coocorrem no corpus original.

2.3.4 Modelo TF-IDF

O modelo TF-IDF é uma técnica amplamente utilizada no NLP para medir a importância de um termo em um documento dentro de um *corpus*. Segundo Seno et al. (2024, p. 197), em vez de usarmos a frequência de ocorrência dos termos no documento, utilizamos a frequência do termo no documento normalizada pela frequência total de palavras do documento.

Conforme Scikit-learn (2023), o Term Frequency (TF) é definido como:

$$\mathrm{tf}(t, d) \tag{1}$$

em que $\text{tf}(t, d)$ representa o número de vezes que o termo t aparece no documento d.

A componente *Inverse Document Frequency* (IDF) é apresentada como:

$$\text{idf}(t) = \ln\left(\frac{1+N}{1+n_t}\right) + 1 \tag{2}$$

onde N é o número total de documentos na coleção e n_t o número de documentos que contêm o termo t. De acordo com Seno et al. (2024, p. 197), a ideia é que termos que ocorrem em muitos documentos não sejam bons discriminadores, enquanto termos que ocorrem em poucos documentos devem ter um peso maior.

Scikit-learn (2023) define o componente IDF assim, pois é comumente aplicado em base de dados de teste onde é possível que n_t seja igual a zero, desta forma, evita-se que o denominador do logaritmando seja zero.

A combinação das duas medidas resulta no valor do componente TF-IDF:

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t) \tag{3}$$

Esse valor, conforme destacado por Seno et al. (2024, p. 197), é maior para termos que aparecem muitas vezes no documento e aparecem em poucos documentos do *corpus*. Essa característica é especialmente útil na análise dos comentários enviados à CPA, pois permite destacar termos que são recorrentes em determinados relatos individuais, mas que não são comuns no conjunto total de comentários, sinalizando aspectos específicos ou preocupações singulares da comunidade acadêmica.

A fórmula pode ser modificada dependendo dos parâmetros utilizados na função Scikit-learn (2023), todavia, a equação (3) é aplicada por padrão, convertendo cada documento em um vetor tf-idf, v, que ainda é normalizado pela norma Euclidiana.

$$v_{norm} = \frac{v}{\|v\|} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}} \tag{4}$$

onde v_{norm} é o vetor tf-idf v normalizado, $\|v\|$ é a norma de v e $(v_1, v_2, \cdots, v_n)$ são os componentes de v. Desta forma, cada termo é analisado com todos os vetores tendo o mesmo comprimento.

2.3.5 Decomposição por valores singulares - SVD

O principal objetivo da SVD é tornar as representações vetoriais dos textos mais compactas, eliminando ruídos e evidenciando relações semânticas que não são diretamente observáveis.

Conforme Seno et al. (2024, p. 199), a SVD decompõe uma matriz A em três componentes:

$$A = U\Sigma V^T \tag{5}$$

A matriz U é uma matriz ortogonal cujas colunas representam os vetores próprios da matriz AA^T, sendo responsável por capturar os conceitos latentes associados aos termos. A matriz Σ é uma matriz diagonal composta pelos valores singulares, organizados em ordem decrescente, os quais indicam a relevância de cada conceito extraído. Já V^T representa a transposta de uma matriz ortogonal cujas linhas correspondem aos vetores próprios da matriz A^TA, associados aos documentos. Essa decomposição permite reduzir a dimensionalidade dos dados, preservando os k maiores valores singulares, o que concentra a maior parte da informação semântica contida no conjunto original (Seno et al., 2024, p. 200).

A utilização da SVD no NLP, em especial na LSA, possibilita a projeção de termos e documentos em um espaço vetorial de baixa dimensionalidade, preservando as relações semânticas mais relevantes. Essa representação vetorial reduzida é particularmente eficaz no tratamento da polissemia[6] e da sinonímia[7], fenômenos linguísticos comuns em textos naturais. Segundo Seno et al. (2024, p. 199-200), a projeção vetorial realizada pela SVD suaviza esses efeitos, aproximando semanticamente documentos e termos mesmo quando não compartilham diretamente termos idênticos ou contextos explícitos.

Dessa forma, a SVD permite não apenas a redução eficiente da complexidade dos dados, mas também a revelação de estruturas semânticas profundas que favorecem o desenvolvimento de sistemas linguísticos mais robustos e inteligentes.

Com o uso de técnicas de NLP, como a análise semântica e a vetorização dos textos, os comentários enviados à CPA da Unifal-MG podem ser transformados em dados que facilitam uma análise mais aprofundada. Isso torna possível agrupar automaticamente os textos por meio do algoritmo *k-means*, revelando padrões de discurso, percepções por áreas específicas da instituição e até mesmo assuntos que passariam despercebidos em uma leitura manual. Ao adotar essa abordagem, a CPA amplia sua capacidade de

[6] A polissemia ocorre quando uma palavra apresenta múltiplos significados, variando conforme o contexto em que é empregada, como no caso do termo "banco", que pode significar tanto uma instituição financeira quanto um assento

[7] A sinonímia refere-se a palavras diferentes que possuem significados semelhantes, como "rápido"e "veloz"

escuta e compreensão das vozes da comunidade acadêmica, de forma mais organizada, sensível e baseada em evidências.

2.4 Clusterização *k-means*

O algoritmo de agrupamento *k-means* é uma técnica de aprendizagem não supervisionada destinada à partição de um conjunto de dados em k grupos distintos, denominados *clusters*. O objetivo é minimizar a variância intra-*cluster* e maximizar a variância inter-cluster, com base em medidas de distância entre os dados e os centroides dos agrupamentos.

De acordo com Seebregts (2022, p. 18), o *k-means* busca organizar observações multidimensionais em k grupos baseando-se na distância entre as observações e o centroide do grupo. A função erro a ser minimizada pode ser expressa por:

$$E = \sum_{i=1}^{k} \sum_{x_n \in C_i} d(x_n, c_i) \tag{6}$$

em que:

- x_n é uma observação pertencente ao agrupamento C_i,
- c_i é o centroide do agrupamento C_i,
- $d(x_n, c_i)$ representa a distância entre a observação x_n e o centroide c_i,
- a métrica de distância comumente usada é a distância euclidiana.

Segundo Soares (2024, p. 9), o algoritmo é iterativo e envolve os seguintes passos:

1. Seleção de k centroides iniciais;
2. Atribuição de cada ponto ao *cluster* mais próximo;
3. Recomputação dos centroides;
4. Repetição até convergência ou critério de parada.

A função de custo que representa a soma das distâncias quadráticas entre os pontos e os centroides é dada por:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2 \tag{7}$$

onde c_i é o centroide do cluster C_i, e $\|\cdot\|$ representa a norma Euclidiana.

Silveira et al. (2017, p. 505) destacam que o algoritmo pode ser aplicado ao agrupamento de vetores em espaços semânticos derivados de textos processados, como os obtidos via *word embeddings*, dada a proximidade semântica ser representada geometricamente. Neste trabalho, estendemos esse entendimento à matriz TF-IDF reduzida por SVD, dado que tal transformação projeta documentos e termos em um espaço vetorial de baixa dimensionalidade, no qual relações semânticas latentes também podem ser interpretadas geometricamente, favorecendo a aplicação de técnicas de clusterização como o *k-means*.

A seguir, apresentaremos técnicas para decidir o número de clusters.

2.4.1 Coeficiente de silhoeta

A escolha do número k de *clusters* é um desafio. Para avaliá-la, pode-se usar o coeficiente de silhoeta, definido por:

$$s = \frac{b - a}{\max(a, b)} \tag{8}$$

onde:

- a é a distância média entre uma amostra e as demais do mesmo grupo,
- b é a menor distância média entre a amostra e os demais grupos.

Esse coeficiente varia de -1 a $+1$, e valores altos indicam boa separação entre os grupos (Seebregts, 2022, p. 41).

2.4.2 Método do Cotovelo

O método do cotovelo é outra técnica amplamente utilizada na determinação do número ótimo de agrupamentos em algoritmos de clusterização, especialmente no *k-Means*. Segundo Syakur et al. (2018), essa abordagem consiste em calcular, para diferentes valores de k, a soma dos erros quadráticos (SSE - *Sum of Squared Errors*) intra-*cluster* representada pela equação:

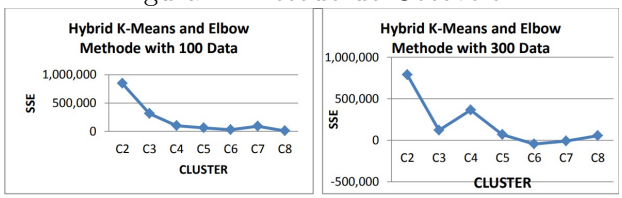
$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

em que k é o número de *clusters*, C_i representa o conjunto de pontos no *cluster* i, x é um ponto de dado pertencente ao *cluster* i, e μ_i é o centróide do *cluster* i. O método propõe a análise gráfica do comportamento do SSE em função de k, observando-se o ponto no qual a redução no erro se torna marginal com o aumento de novos

clusters. Esse ponto é denominado "cotovelo" da curva e indica o valor de k mais adequado para a segmentação dos dados.

A interpretação gráfica é fundamental para o método: o valor ideal de k é aquele no qual a curva SSE vs. k apresenta uma inflexão significativa, após a qual as melhorias tornam-se pouco expressivas. Syakur et al. (2018) ilustram esse processo com testes empíricos em perfis de clientes, demonstrando que o valor $k = 3$ representou o ponto de cotovelo, sendo o mais apropriado para agrupar os dados analisados, como mostra a Figura 1. A aplicação do método do cotovelo, nesse contexto, possibilitou uma segmentação mais eficiente dos clientes, contribuindo para a definição de estratégias comerciais mais adequadas.

Figura 1: Método do Cotovelo



Fonte: Syakur et al. (2018, p. 5)

Entre as vantagens apontadas pelos autores está a facilidade de implementação do método e sua eficiência na redução da dimensionalidade da análise quando associado ao algoritmo de *k-means*. No entanto, Syakur et al. (2018) reconhecem limitações importantes, como a subjetividade na identificação visual do cotovelo, especialmente em curvas que não apresentam inflexões nítidas. Ademais, a qualidade do agrupamento depende fortemente da escolha inicial dos centróides, o que pode afetar a acurácia do ponto de cotovelo encontrado.

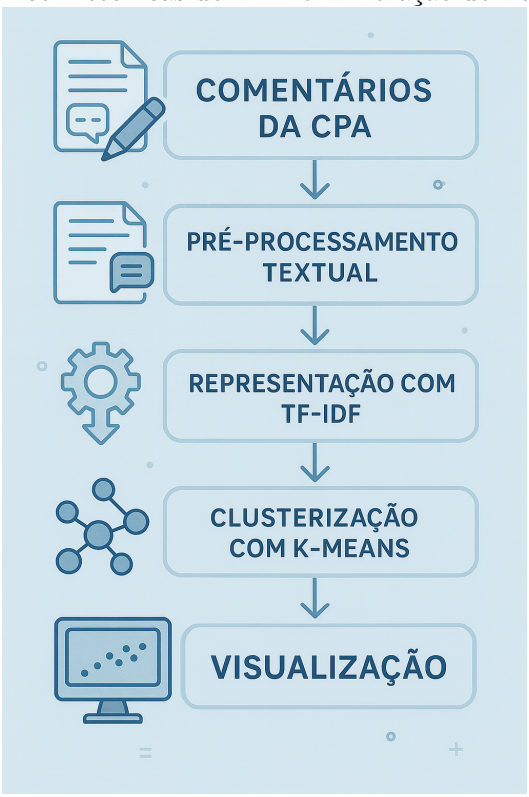
Dessa forma, o método do cotovelo, apesar de simples e amplamente difundido, deve ser utilizado com cautela, preferencialmente combinado com outros critérios de validação de *clusters*, a fim de garantir uma segmentação robusta e representativa dos dados analisados.

3 Procedimentos metodológicos

Este trabalho visa análise automatizada dos comentários qualitativos coletados pela CPA da Unifal-MG, referentes aos ciclos avaliativos dos períodos letivos de 2023.1, 2023.2, 2024.1 e 2024.2. O objetivo central consiste em aplicar técnicas de NLP e mineração de textos para identificar padrões semânticos e temas latentes expressos pela comunidade acadêmica.

As etapas metodológicas da análise estão sintetizadas no organograma a seguir, ver Figura 2, que ilustra o fluxo sequencial do processamento textual até a interpretação dos resultados.

Figura 2: Etapas da análise de comentários da CPA com técnicas de PLN e Mineração de Texto



Fonte: autor

Inicialmente, realizamos o pré-processamento dos comentários, envolvendo correção ortográfica, lematização, remoção de *stopwords* e normalização, de forma a garantir uniformidade lexical. Em seguida, os textos são convertidos em vetores numéricos por meio da representação TF-IDF, dando evidência as palavras que tornam o documento único. A redução de dimensionalidade via SVD é então aplicada, permitindo a identificação de estruturas semânticas ocultas, conforme os princípios da LSA.

Na etapa subsequente, os vetores reduzidos são submetidos ao algoritmo *k-means*, com o objetivo de agrupar semanticamente os comentários. A definição do número ótimo de grupos é guiada pela análise do coeficiente de silhoeta. Por fim, os resultados obtidos são visualizados e interpretados com base em nuvens de palavras e análise qualitativa dos agrupamentos.

3.1 Coleta e organização dos dados

Discentes, docentes e técnicos administrativos foram convidados a registrar sugestões, elogios,

críticas e reclamações, por meio de formulários disponibilizados pela CPA, resultando em um conjunto expressivo e diversificado de manifestações textuais.

Após a coleta, os dados foram organizados em 4 arquivos HTML e foram consolidados em um único arquivo no formato `.csv`, estruturado com as informações da Tabela 1:

Tabela 1: Estrutura dos dados analisados

Campo	**Descrição**
id	Identificação única de cada resposta
Semestre	Semestre em que as respostas foram coletadas (2023.1, 2023.2, 2024.1 e 2024.2)
Eixo	1 - Planejamento e Avaliação Institucional 2 - Desenvolvimento Institucional 3 - Políticas Acadêmicas 4 - Políticas de Gestão 5 - Infraestrutura
Resposta	Comentário realizado pelo membro da comunidade acadêmica

A leitura e o acesso ao conteúdo textual foram realizados por meio da biblioteca `pandas`[8], amplamente utilizada para análise de dados em *Python*, o que facilitou a preparação do *corpus* para as etapas subsequentes de pré-processamento linguístico e análise semântica.

3.2 Pré-processamento de dados

No presente estudo, adotamos um conjunto de etapas de pré-processamento textual com o objetivo de estruturar os dados linguísticos de maneira adequada às tarefas de análise semântica e modelagem computacional. Essas etapas, alinhadas às práticas consolidadas no NLP, visam à redução de ruídos textuais, à padronização dos dados e à melhoria da qualidade do *corpus*. As operações realizadas incluíram a filtragem dos dados, a remoção de valores ausentes, de caracteres especiais, de sinais de pontuação e de numerais, além da conversão de todos os textos para letras minúsculas, uniformizando a representação lexical.

Em seguida, procedeu-se à remoção de stopwords, à identificação de palavras com erros de digitação e à exclusão de termos com frequência extremamente alta ou baixa, por apresentarem, respectivamente, pouca capacidade discriminativa ou irrelevância estatística. Por fim, aplicou-se a lematização, com a finalidade de reduzir as palavras às suas formas canônicas. Cada uma dessas etapas será detalhada nas seções subsequentes.

3.2.1 Limpeza dos dados

Durante a etapa de limpeza de dados, os comentários textuais foram inicialmente filtrados com base em dois critérios: o semestre e eixo. Essa filtragem visou organizar o *corpus* de forma estruturada e segmentada, permitindo uma análise mais precisa e contextualizada das manifestações da comunidade acadêmica em relação aos diferentes aspectos avaliados. A separação por semestre possibilita o exame temporal das avaliações, contribuindo para a identificação de padrões recorrentes e possíveis mudanças nas percepções ao longo do tempo. Já a segmentação por eixo (como infraestrutura, ensino, gestão, entre outros) permite associar os comentários a dimensões específicas do processo avaliativo, favorecendo análises direcionadas e a geração de relatórios temáticos mais coerentes com os objetivos da CPA.

Além da organização temática e temporal, a etapa de limpeza também incluiu a identificação e remoção de valores ausentes. Comentários vazios ou nulos foram excluídos do *corpus*, a fim de eliminar ruídos e preservar a integridade da análise textual.

3.2.2 Normalização

Inicialmente, procedeu-se à remoção de caracteres especiais, sinais de pontuação e números, utilizando a biblioteca `re`[9]. Essa biblioteca é amplamente utilizada para manipulação e filtragem de textos com expressões regulares, permitindo a definição precisa de padrões para substituição ou exclusão de trechos textuais. Em seguida, todas as palavras foram convertidas para **letras minúsculas**, com o objetivo de uniformizar o *corpus* e evitar duplicações semânticas oriundas da capitalização de termos.

Outro passo foi a remoção de *stopwords* — palavras de alta frequência e baixo valor semântico, como 'de', 'para', 'como' — que foi realizada com o auxílio da biblioteca `nltk`[10] (*Natural Language Toolkit*). Trata-se de uma biblioteca consolidada para aplicações de NLP, oferecendo recursos para tokenização, análise sintática, léxica

[8] A documentação da biblioteca `Pandas` está disponível no endereço https://pandas.pydata.org/docs/

[9] A documentação, em português, da biblioteca `re` está disponível no endereço https://docs.python.org/pt-br/3/library/re.html

[10] A documentação da biblioteca `nltk` está disponível no endereço https://www.nltk.org/

e semântica. Como os comentários estavam em língua portuguesa, utilizou-se o comando `stopwords.words('portuguese')`, que fornece uma lista pré-definida de palavras consideradas irrelevantes para tarefas de mineração de texto em nosso idioma.

Adicionalmente, aplicou-se a correção ortográfica automatizada por meio da biblioteca `pyspellchecker`[11], especializada na identificação e sugestão de correções para palavras com erros de digitação e possui suporte para o português. Essa ferramenta baseia-se em um modelo estatístico que compara palavras não reconhecidas com um dicionário predefinido e propõe alternativas mais prováveis. No entanto, durante sua aplicação, foram identificados falsos positivos em palavras válidas e contextualmente relevantes. Por exemplo, a palavra "Unifal"foi incorretamente sugerida como "animal"e "Planejamento"como "Planeamento". Para mitigar esse problema, implementou-se um filtro personalizado de exceções, no qual termos institucionalmente corretos e semanticamente significativos foram excluídos da verificação ortográfica, preservando a integridade dos dados.

A etapa seguinte, procedeu-se à lematização dos textos, utilizando a biblioteca `stanza`[12], desenvolvida pelo Stanford NLP Group. A lematização consiste na transformação de palavras flexionadas em suas formas canônicas (ou "lemas"), aplicada na redução da dimensionalidade e a agregação semântica do *corpus*. Embora existam bibliotecas mais rápidas para esse fim, a escolha pela `stanza` se deu em razão de sua alta acurácia na lematização em português, superando outras abordagens em consistência morfossintática, conforme verificado em estudos comparativos prévios.

Na sequência, removemos as 25% das palavras mais frequentes e as 25% das palavras menos frequentes do *corpus*, com o intuito de eliminar tanto os termos excessivamente genéricos quanto os extremamente raros, que poderiam introduzir ruído na representação semântica. Essa filtragem estatística contribuiu para refinar a base textual, realçando os termos de relevância intermediária, mais informativos para a etapa de análise semântica e clusterização dos comentários institucionais.

Por fim, após todas as etapas de normalização e filtragem lexical, os comentários que continham menos de três palavras foram eliminados. Essa decisão foi baseada na premissa de que comentários extremamente curtos apresentam baixa relevância semântica e tendem a não contribuir significativamente para a análise de padrões latentes ou para a construção de agrupamentos temáticos consistentes.

3.3 Vetorização

Para a representação dos comentários em um espaço vetorial adequado à análise semântica, foi utilizada a técnica de LSA, com base na SVD de uma matriz de termos ponderados por TF-IDF. Inicialmente, construiu-se a matriz TF-IDF a partir dos textos normalizados, por meio da função `TfidfVectorizer`, pertencente à biblioteca `scikit-learn`[13] que é uma das bibliotecas mais amplamente utilizadas em aplicações de aprendizado de máquina em *Python*, oferecendo implementações eficientes de algoritmos de classificação, regressão, agrupamento e redução de dimensionalidade. A função `TfidfVectorizer` transforma o *corpus* em uma matriz numérica esparsa, na qual cada linha representa um documento (neste caso, um comentário) e cada coluna representa um termo, ponderado pela frequência inversa nos documentos.

Com a matriz TF-IDF construída, aplicou-se a técnica de redução de dimensionalidade por meio da função `TruncatedSVD`, também fornecida pela biblioteca `scikit-learn`. Essa função executa a SVD truncada, permitindo a extração das componentes semânticas latentes mais relevantes da matriz de termos. Diferente de uma decomposição completa, o `TruncatedSVD` é otimizado para matrizes esparsas e se mostra ideal para aplicações de LSA em grandes volumes de texto.

Para garantir a preservação da informação semântica mais significativa, foi implementado um código que determina automaticamente o número de componentes a ser mantido na decomposição, com base na variância explicada acumulada. O critério adotado foi que a soma das variâncias explicadas pelas componentes selecionadas fosse de, no mínimo, 90%. Esse procedimento assegura um equilíbrio entre a redução de dimensionalidade e a retenção do conteúdo semântico essencial, permitindo que os dados projetados no novo espaço vetorial reflitam com maior clareza os padrões latentes presentes nos comentários analisados.

[11] Documentação oficial disponível no endereço https://pypi.org/project/pyspellchecker/

[12] Documentação disponível no endereço https://stanfordnlp.github.io/stanza/

[13] A documentação oficial está diponível no endereço https://scikit-learn.org/stable/

3.4 Clusterização

O processo de agrupamento foi realizado a partir da matriz TF-IDF previamente reduzida por meio da SVD. Utilizamos o algoritmo *k-means*, implementado pela função `KMeans` da biblioteca `scikit-learn`, adotando-se a distância euclidiana como métrica de similaridade entre os vetores.

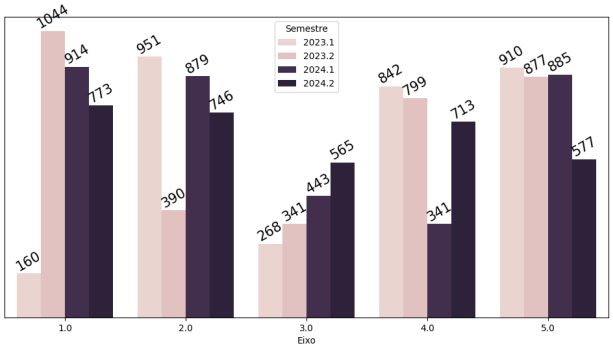
Embora as referências apresentadas neste trabalho recomendem a normalização prévia dos dados antes da aplicação do *k-means*, optamos por não normalizá-los, considerando que a matriz resultante da redução dimensional via SVD já apresenta os vetores projetados em uma mesma base métrica e com distribuição ajustada, o que mitiga variações indesejadas de escala.

Para determinar o número ideal de clusters, inicialmente consideramos o método do cotovelo e o coeficiente de silhoeta. No entanto, conforme será discutido sobre os resultados, o método do cotovelo não forneceu uma avaliação satisfatória da qualidade dos agrupamentos para o nosso conjunto de dados. Diante disso, optamos por utilizar o dendrograma, que se mostrou mais adequado para identificar a estrutura hierárquica dos dados e auxiliar na definição do número apropriado de clusters.

4 Resultados e Discussão

A base de dados apresenta o total de 13418 comentários. O gráfico da Figura 3 apresenta a distribuição por semestre e eixo dos dados coletados.

Figura 3: Distribuição dos comentários por eixo e semestre



Fonte: o autor

O semestre com maior número de manifestações foi 2024.1, com 3462 registros, seguido por 2023.2 (3451), 2024.2 (3374) e 2023.1 (3131). A variação na quantidade de comentários entre os semestres pode refletir diferentes níveis de participação da comunidade acadêmica ao longo do tempo.

No recorte por eixo temático, o Eixo 5 (Infraestrutura) se destacou com o maior volume de comentários (3249), o que evidencia o papel central das condições físicas da universidade nas percepções da comunidade. Em seguida, os Eixos 2 (Desenvolvimento Institucional) e 1 (Planejamento e Avaliação Institucional) apresentaram, respectivamente, 2966 e 2891 manifestações, indicando atenção significativa às dimensões estratégicas e sociais da instituição. O Eixo 4 (Políticas de Gestão) totalizou 2695 comentários, enquanto o Eixo 3 (Políticas Acadêmicas) obteve o menor número de registros (1617), o que pode sugerir menor mobilização em torno de temas relacionados ao ensino, à pesquisa, à extensão e ao atendimento discente. Essa distribuição revela que os aspectos estruturais e institucionais despertam maior interesse da comunidade universitária, servindo como indicativo de prioridades percebidas e direcionamento para ações futuras de melhoria institucional.

Considerando que a filtragem dos dados por semestre e eixo gera uma quantidade significativa de combinações possíveis para análise, optamos por discutir, sem perda de generalidade, os comentários registrados no eixo 5 do semestre 2024.2, que reúne 577 manifestações e oferece um recorte suficientemente expressivo para a apresentação e interpretação dos resultados deste estudo. Este filtro está sendo escolhido por ser o mais recente e num dos eixos de maior engajamento no total dos dados.

Considerando o filtro aplicado, verificamos que o dicionário resultante da base textual contém 4030 palavras distintas, o que indica uma diversidade lexical significativa. Contudo, verificamos a presença de entradas não informativas, como os sinais '-', 'xxxx' e '—', que, apesar de não possuírem valor semântico relevante, foram computadas como palavras válidas. Para promover a padronização e a qualificação do vocabulário, foi realizado um processo de normalização textual em múltiplas etapas.

A primeira etapa consistiu na remoção de numerais, pontuações e caracteres especiais, com exceção do hífen (-), cuja manutenção se justificou pela sua relevância semântica em construções como 'pode-se', cuja integridade seria comprometida caso fosse transformada em 'podese'. Também foi realizada a conversão de todos os caracteres para letras minúsculas. Essa etapa inicial resultou na redução do vocabulário para 2861 palavras distintas.

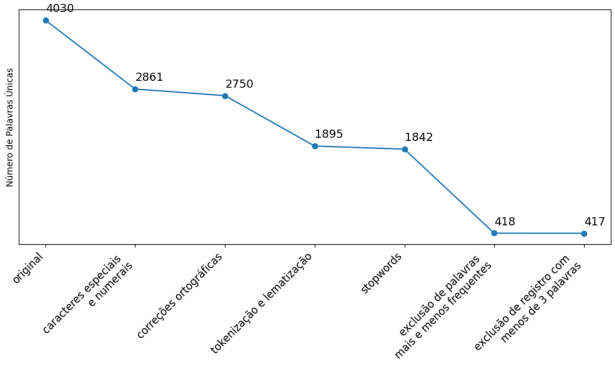
Ainda na fase de normalização, procedeu-se à padronização de termos recorrentes, como a

substituição da forma 'unifal' e 'unifal mg' por 'unifalmg', assegurando uniformidade institucional no *corpus*. Também foram corrigidos erros ortográficos, resultando em um novo total de 2750 palavras. A etapa seguinte consistiu na tokenização e lematização o que possibilitou a compactação do vocabulário para 1895 termos. Na sequência, foram removidas as palavras irrelevantes (*stopwords*), reduzindo o dicionário para 1842 termos.

Posteriormente, foi realizada uma análise de frequência, com o objetivo de identificar termos que apresentavam baixa ou excessiva recorrência. Foram selecionadas para exclusão as 25% palavras mais frequentes (32 termos) e as 25% menos frequentes (1392 termos), por contribuírem com pouco conteúdo informativo ou por representarem ocorrências pontuais no *corpus*. A exclusão dessas classes resultou em um vocabulário com 418 palavras.

Por fim, observou-se a presença de comentários compostos por uma ou duas palavras, muitos dos quais eram registros vazios, compostos por sinais ('-'), ou expressões genéricas como 'ok' e 'tudo ok'. Considerando a baixa densidade semântica desses registros, optou-se por sua exclusão, sem prejuízo à representatividade da análise. Como resultado, o número de comentários válidos foi reduzido para 316, e o dicionário final passou a conter 417 palavras, representando um *corpus* limpo, padronizado e adequado para as etapas subsequentes de análise semântica e interpretativa. A Figura 4 apresenta o gráfico da variação do número de palavras do dicionário durante o pré-processamento.

Figura 4: Variação do número de palavras do dicionário



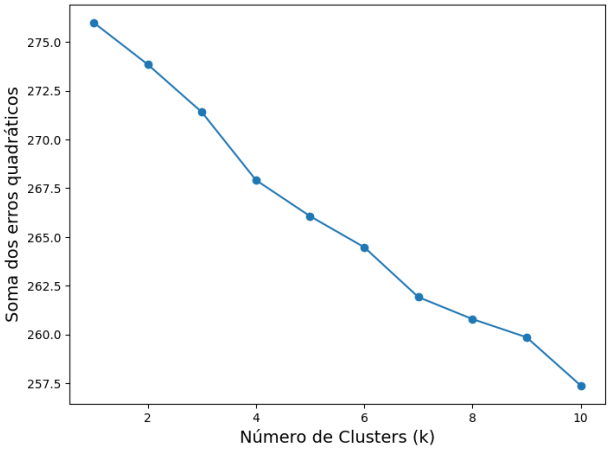
Fonte: o autor

A remoção de registros com até duas palavras pouco impactou o dicionário (417 palavras), mas resultou na eliminação de comentários com baixa densidade informacional, como "ok" e "tudo ok", restringindo a base a 316 respostas substancialmente mais relevantes para análise. O gráfico sintetiza com clareza o efeito cumulativo das etapas de limpeza e refinamento lexical, fundamentais para garantir qualidade semântica ao *corpus* analisado.

Para representar vetorialmente os comentários, foi construída uma matriz TF-IDF com dimensões 316×415, refletindo os 362 documentos e 415 termos distintos obtidos após a normalização textual. Essa matriz foi reduzida por meio da técnica de SVD, com a definição automatizada do número de componentes necessário para preservar, no mínimo, 90% da variância dos dados. O critério foi atendido com 175 componentes, resultando em uma matriz reduzida de dimensões 362×175, cuja variância explicada atingiu 90,05%. Essa redução permitiu manter a estrutura semântica essencial do *corpus*, ao mesmo tempo em que tornou as análises mais eficientes e menos suscetíveis à redundância e ao ruído.

Para o processo de agrupamento, com foco na aplicação do algoritmo *k-means*, testamos valores de k variando de 2 a 10 clusters, uma vez que números mais elevados poderiam gerar agrupamentos excessivamente diluídos, dificultando a análise e interpretação semântica dos grupos formados. Inicialmente, aplicamos o método do cotovelo com o intuito de identificar o ponto de inflexão na curva da soma dos erros quadráticos. No entanto, conforme evidencia a Figura 5, não se observou uma estabilização clara da curva, o que compromete a aplicabilidade desse critério para os dados em questão — um padrão que se repetiu mesmo em outros filtros aplicados ao *corpus*.

Figura 5: Gráfico de Cotovelo

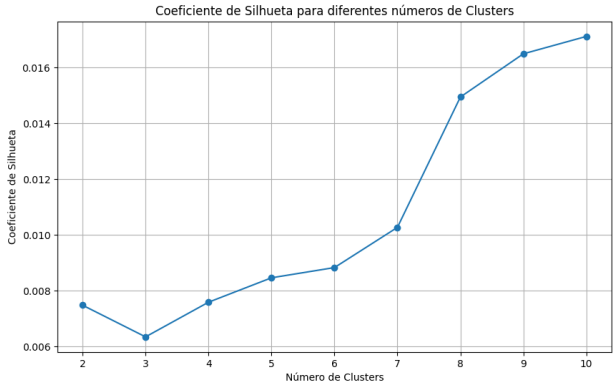


Fonte: o autor

Em seguida, analisamos o Coeficiente de silhoeta, ver Figura 6, para os valores de k dentro do intervalo definido. Observou-se um valor máximo em $k = 10$, sugerindo esse como o número mais

adequado de *clusters*.

Figura 6: Gráfico do Coeficiente de Silheta

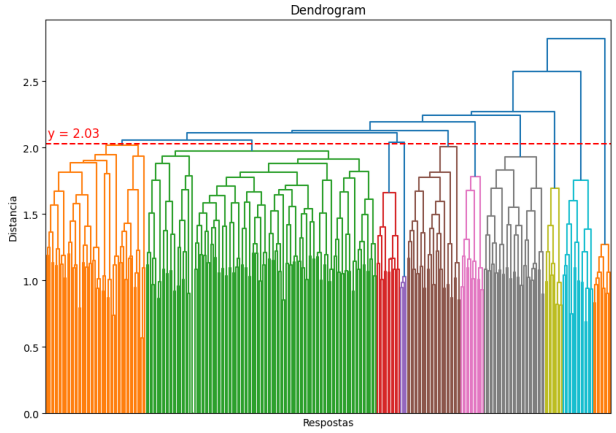


Fonte: o autor

Entretanto, testamos o coeficiente de silhoeta para todas as dimensões e não foi indicado um número adequado de clusters.

Para complementar a análise, foi gerado um dendrograma por meio de clusterização hierárquica. Nesse tipo de representação, a distância vertical entre as ramificações indica o grau de dissimilaridade entre os agrupamentos: cortes horizontais em níveis mais altos implicam em maior distinção entre os grupos. Observa-se que, ao se adotar um corte na altura aproximada de 2,03, são formados dez agrupamentos principais — coincidindo com a sugestão do Coeficiente de silhoeta.

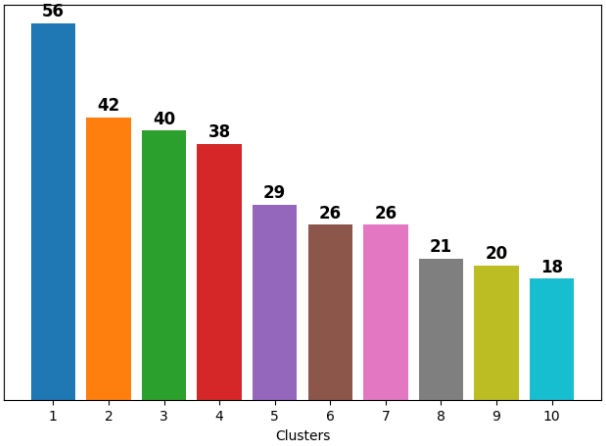
Figura 7: Dendograma



Com a aplicação do dendograma definimos que a melhor escolha para o número de clusters é $k = 10$ para a etapa de análise interpretativa dos agrupamentos.

Ao aplicar o algoritmo *k-means*, configuramos a segmentação em 10 *clusters*, cuja distribuição está representada na Figura 8, cujos clusters foram numerados em ordem decrescente de frequência.

Figura 8: Distribuição dos Clusters



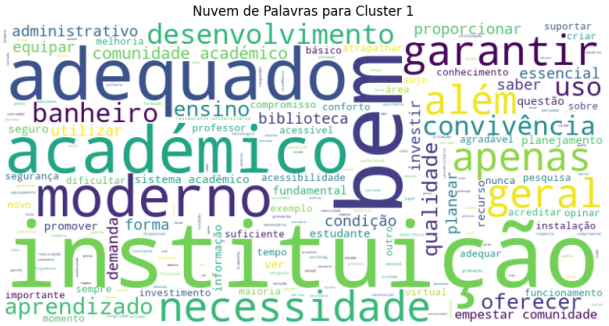
Fonte: o autor

Os resultados obtidos pela clusterização podem ser utilizados para identificar e resolver as demandas relacionadas aos respectivos *clusters*. Salientamos que a forma como os dados foram estruturados preserva o anonimato dos respondentes. Deste modo, apresentamos três exemplos de comentários dos dois *clusters* de maior frequência e suas respectivas nuvens de palavras.

***Cluster* 1**:

1. "Ainda estou em fase de adaptação à instituição, pois sou novo aqui. No momento, não me sinto suficientemente preparado para comentar sobre o assunto com propriedade".
2. "A estrutura física da Universisade, diante de um cenário nacional, é impecável. Há um cuidado intenso com os espaços físicos, mantendo-os sempre apropriados para uso comunitário. Salas de aulas e banheiros com espaços suficientes para o bom uso.".
3. "Toda e escola oferece condições excelentes para que os alunos tenham uma formação de qualidade".

Figura 9: Nuvem de palavras do *cluster* 1

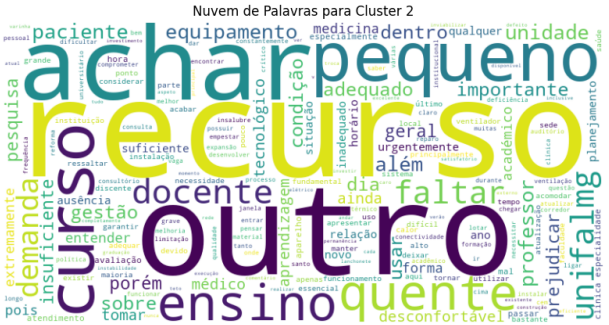


Fonte: o autor

***Cluster* 2**:

1. "A minha questão principal sobre a estrutura física é a falta de vagas para automóveis. Não tem vaga suficiente para os docentes".
2. "Falta de recursos tecnológicos atualizados. Principalmente nas salas de informática".
3. "No geral, a infra-estrutura física é favorável para a realização das atividades. Porém, faltam espaços e recursos para melhor preservação de documentos (como, para o CEDOC). Faltam, também, espaços para atendimentos aos discentes e espaços de sociabilidade entre eles (considero um último um fator muito importante para a instituição considerar)".

Figura 10: Nuvem de palavras do *cluster* 2

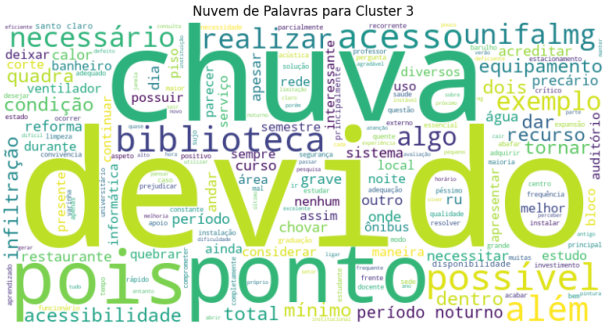


Fonte: o autor

A partir destes resultados, e com o auxílio de ferramentas de inteligência artificial, por exemplo, podemos identificar os assuntos mais recorrentes de cada *cluster*.

Seguem as nuvens de palavras dos demais *clusters*.

Figura 11: Nuvem de palavras do *cluster* 3



Fonte: o autor

Figura 12: Nuvem de palavras do *cluster* 4

Fonte: o autor

Figura 13: Nuvem de palavras do *cluster* 5

Fonte: o autor

Figura 14: Nuvem de palavras do *cluster* 6

Fonte: o autor

Figura 15: Nuvem de palavras do *cluster* 7

Fonte: o autor

Figura 16: Nuvem de palavras do *cluster* 8

Fonte: o autor

Figura 17: Nuvem de palavras do *cluster* 9

Fonte: o autor

Figura 18: Nuvem de palavras do *cluster* 10

Fonte: o autor

5 Considerações Finais

Nosso objetivo principal foi realizar a análise semântica automatizada dos comentários qualitativos submetidos à CPA da Unifal-MG, com a aplicação de técnicas de NLP, LSA e clusterização por meio do algoritmo *K-means*. O problema investigado está relacionado à dificuldade de sistematizar e interpretar, de forma objetiva e escalável, os grandes volumes de comentários coletados nos formulários institucionais de autoavaliação. Esses dados, geralmente analisados de maneira manual e subjetiva, apresentam limitações quanto à padronização e à identificação de padrões significativos, o que pode comprometer a efetividade das ações institucionais. Diante disso, o estudo motiva-se pela necessidade de desenvolver abordagens automatizadas e orientadas por dados que ampliem a capacidade analítica da CPA, promovendo a extração de temas recorrentes e percepções latentes expressas pela comunidade acadêmica, de modo a embasar decisões.

Para alcançar os objetivos propostos, a pesquisa foi conduzida por meio de etapas metodológicas articuladas, iniciando-se com o pré-processamento textual dos comentários, que incluiu correção ortográfica, lematização, remoção de *stopwords* e filtragem lexical. Essa preparação permitiu a padronização linguística e a redução de ruídos, viabilizando a construção da matriz TF-IDF, utilizada para representar a importância relativa dos termos no *corpus*. A seguir, aplicou-se a SVD, que possibilitou a redução da dimensionalidade e a projeção dos textos em um espaço semântico vetorial mais compacto e informativo.

Com base nessa representação, foi implementado o algoritmo *K-means*, cuja configuração com 10 *clusters* foi definida a partir da análise do coeficiente de silhoeta e de um dendrograma interpretativo. Os resultados obtidos permitem a identificação de assuntos recorrentes em cada agrupamento com o auxílio de nuvens de palavras com a aplicação de outras ferramentas com inteligência artificial, por exemplo. Os resultados reforçam a potencialidade das técnicas aplicadas na sistematização de dados qualitativos e na produção de evidências para subsidiar ações institucionais mais estratégicas e fundamentadas.

Nesse contexto, a pesquisa apresenta contribuições relevantes em diferentes dimensões. Do ponto de vista metodológico, propõe e valida um modelo de análise automatizada de dados textuais que integra técnicas de NLP, LSA e clusterização por *K-means*, com potencial de aplicação em outras instituições de ensino superior que enfrentam desafios semelhantes na sistematização de manifestações qualitativas. Em termos institucionais, oferece à CPA da Unifal-MG uma ferramenta robusta para a interpretação das per-

cepções da comunidade acadêmica, ampliando a capacidade analítica da comissão e fortalecendo sua atuação como agente estratégico no processo de autoavaliação.

Apesar dos resultados alcançados, a pesquisa apresenta limitações que devem ser consideradas. A definição do número ideal de clusters envolveu certo grau de subjetividade, já que o método do cotovelo não apresentou ponto de inflexão claro, exigindo a combinação do coeficiente de silhoeta com a análise de um dendrograma. Além disso, embora os agrupamentos tenham sido gerados de forma automatizada, a interpretação temática de cada *cluster* pode exigir um julgamento humano, o que introduziria vieses. Persistem ainda possibilidades de ruídos linguísticos residuais, mesmo após o pré-processamento textual, dada a natureza espontânea dos comentários. Soma-se a isso uma limitação de ordem computacional, mesmo após a aplicação de um filtro que reduziu a análise a 4,3% do total de comentários, o tempo de execução do algoritmo foi de aproximadamente 20 minutos. Esse tempo elevado, frente a uma amostra reduzida, indica alta complexidade computacional, podendo dificultar análises em larga escala, limitar a replicação em sistemas institucionais com restrições de processamento e inviabilizar aplicações que exijam respostas em tempo real.

Diante dessas limitações e dos avanços metodológicos alcançados, abrem-se possibilidades promissoras para trabalhos futuros. Uma primeira direção consiste em ampliar a análise para a totalidade dos comentários coletados, abrangendo todos os eixos temáticos e períodos avaliativos, a fim de proporcionar uma visão mais abrangente e representativa das percepções institucionais. Além disso, recomenda-se a investigação de algoritmos alternativos de agrupamento e métodos hierárquicos. Também se destaca a relevância de incorporar a análise de polaridade — distinguindo entre manifestações positivas, neutras e negativas — como recurso complementar à interpretação temática dos grupos, oferecendo à gestão institucional uma visão mais precisa do tom geral das manifestações. Também indicamos o desenvolvimento de metodologias para identificar assuntos recorrentes de cada *cluster* a partir das nuvens de palavras geradas pelo algoritmo *k-means*, podendo, assim, ajudar na fundamentação de tomadas de decisões da gestão institucional amparadas nas necessidades da comunidade acadêmica. Outra proposta envolve o desenvolvimento de uma ferramenta visual interativa, que permita aos membros da CPA explorar os resultados de forma dinâmica, com filtros por eixo, semestre, polaridade e tema, favorecendo o uso prático dos achados. Por fim, sugere-se a replicação da metodologia em outras instituições de ensino superior, com o intuito de testar sua generalização e identificar padrões discursivos comuns ou específicos a diferentes contextos acadêmicos.

Enfim, espero que os resultados aqui apresentados inspirem novos estudos e sirvam de apoio para uma universidade cada vez mais sensível às demandas da sua comunidade.

Referências

Amaral, João Alberto da Silva. 2021. *Um modelo para seleção de tópicos relevantes em documentos aplicados a compliance*. Recife: Universidade de Pernambuco, Escola Politécnica de Pernambuco. Dissertação (mestrado em engenharia de computação). https://sucupira-legado.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=10968297. Acesso em: 10 maio 2025.

Bholowalia, Payal & Arvind Kumar. 2014. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications* 105(9). 17–24.

Brasil. 2004. Lei nº 10.861, de 14 de abril de 2004 - institui o sistema nacional de avaliação da educação superior (sinaes). Acesso em: 16 fev. 2025. https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm.

Brasil. 2017. Decreto nº 9.235, de 15 de dezembro de 2017 - regulação, supervisão e avaliação de instituições de educação superior. Acesso em: 16 fev. 2025. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/decreto/d9235.htm.

Caseli, Helena de Medeiros, Maria das Graças Volpe Nunes & Christiane Marquezan Pagano. 2024. Introdução e metodologia para o desenvolvimento de projetos. Em Helena de Medeiros Caseli & Maria das Graças Volpe Nunes (eds.), *Processamento de Linguagem Natural: fundamentos e aplicações*, 9–15, 286. S.l.: BPLN 2nd edn. https://brasileiraspln.com/livro-pln/2a-edicao/. Acesso em: 02 abr. 2025.

Castro, Rute Nogueira Silveira de. 2006. *Descoberta de relacionamentos entre padrões de*

software utilizando semântica latente. Fortaleza: Universidade Federal do Ceará. Dissertação (mestrado em ciência da computação). https://repositorio.ufc.br/handle/riufc/18647. Acesso em: 15 maio 2025.

Finatto, Maria José Bocorny, Helena M. Caseli, Lucelene Lopes & Amanda Rassi. 2024. Sequência de caracteres e palavras: morfologia e morfossintaxe. Em Helena M. Caseli, Maria das Graças Volpe Nunes & Adriana Pagano (eds.), *Processamento de linguagem natural: fundamentos e avanços*, 67–99. São Paulo: GLOBE 2nd edn. https://brasileiraspln.com/livro-pln/2a-edicao/. Acesso em: 2 abr. 2025.

Gustriansyah, Riki, N Suhandi & F Antony. 2020. Clustering optimization in rfm analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science* 18(1). 470–477. doi:10.11591/ijeecs.v18.i1.pp470-477.

Madureira, Brielen. 2024a. Avaliação de tecnologias de linguagem. Em Helena de Medeiros Caseli & Maria das Graças Volpe Nunes (eds.), *Processamento de Linguagem Natural: fundamentos e aplicações*, 273–290. S.l.: BPLN 2nd edn. https://brasileiraspln.com/livro-pln/2a-edicao/. Acesso em: 02 abr. 2025.

Madureira, Brielen. 2024b. Diálogo e interatividade. Em Helena de Medeiros Caseli & Maria das Graças Volpe Nunes (eds.), *Processamento de Linguagem Natural: fundamentos e aplicações*, 397–424. S.l.: BPLN 2nd edn. https://brasileiraspln.com/livro-pln/2a-edicao/. Acesso em: 02 abr. 2025.

Marcolin, Carla Bonato, Fernanda da Silva Momo, João Luiz Becker & Ariel Behr. 2019. Mineração de texto para análise de discurso: temáticas e argumentos da decisão de voto de deputados durante a votação do impeachment. *Revista Alcance* 26(1). 4–12. https://periodicos.univali.br/index.php/ra/article/view/13339.

Moreira, Viviane P. 2024. Recuperação de informação. Em Helena de Medeiros Caseli & Maria das Graças Volpe Nunes (eds.), *Processamento de Linguagem Natural: fundamentos e aplicações*, 457–474. S.l.: BPLN 2nd edn. https://brasileiraspln.com/livro-pln/2a-edicao/. Acesso em: 02 abr. 2025.

Paes, Aline & Cláudia Freitas. 2024. Chatgpt, maritalk e outros agentes de conversação. Em Helena de Medeiros Caseli & Maria das Graças Volpe Nunes (eds.), *Processamento de Linguagem Natural: fundamentos e aplicações*, 425–455. S.l.: BPLN 2nd edn. https://brasileiraspln.com/livro-pln/2a-edicao/. Acesso em: 02 abr. 2025.

Pilatti, Rodrigo. 2023. Segmentação comportamental de utilizadores de cartão de crédito utilizando o algoritmo de máquina não supervisionado k-means. Trabalho de Conclusão de Curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. https://bdta.abcd.usp.br/directbitstream/650b7a2c-c19e-49e2-9b5c-70b6ebef2272/Rodrigo%20Pilatti.pdf.

Rezende, João Marcos de. 2019. *Um sistema de mineração em patentes para tendências tecnológicas e análise de similaridaderi*. Serra: Instituto Federal do Espírito Santo. Dissertação (mestrado em engenharia de controle e automação).

Ribeiro, Lívia de Andrade. 2024. *Proposição de Índice de qualidade de Água subterrânea (iqasub) para aplicação em áreas com potencial minerário*. Belo Horizonte: Universidade Federal de Minas Gerais, Escola de Engenharia. Dissertação (mestrado em saneamento, meio ambiente e recursos hídricos). https://repositorio.ufmg.br/handle/1843/77418. Acessado em 15 mai. 2025.

Rousseeuw, Peter J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20. 53–65.

Scikit-learn. 2023. TfidfVectorizer – scikit-learn 1.7.0 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. Acesso em: 10 jun. 2025.

Seebregts, Caio. 2022. Utilização de algoritmo k-means para definição de domínios litológicos.

Seno, Eloize Rossi Marques, Daniela Claro, Laila Mota & Jessica Rodrigues. 2024. Semântica distribucional. Em Helena de Medeiros Caseli & Maria das Graças Volpe Nunes (eds.), *Processamento de Linguagem Natural: uma introdução prática*, 191–206. São Paulo: BPLN 2nd edn. https://brasileiraspln.com/livro-pln/2a-edicao/. Acesso em: 2 abr. 2025.

Silva, Wylliam Eduardo Alves. 2021. Análise de cluster aplicada aos dados de preços de combustíveis na cidade de campina grande - pb. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia. https://dspace.bc.uepb.edu.br/xmlui/handle/123456789/25640.

Silveira, T. B. N., H. S. Lopes, A. E. Lazzaretti, D. P. Araújo & C. F. Valério. 2017. Classificação de contexto para processamento da linguagem natural baseado em representação vetorial de palavras e no agrupamento por k-means. Em *Anais do IX Computer on the Beach*, 502–510. UTFPR.

Soares, Arthur Maurício Thomaz. 2024. Democracia descomplicada: utilizando processamento de linguagem natural para classificar propostas de parlamentares na câmara dos deputados.

Syakur, M. A., B. K. Khotimah, E. M. S. Rochman & B. D. Satoto. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering* 336. 012017. doi:10.1088/1757-899X/336/1/012017. https://iopscience.iop.org/article/10.1088/1757-899X/336/1/012017.

Universidade Federal de Alfenas. 2018. Resolução nº 24, de 7 de maio de 2018 - regimento interno da comissão própria de avaliação da unifal-mg. Acesso em: 10 jun. 2025. https://www.unifal-mg.edu.br/portal/wp-content/uploads/sites/52/2019/02/015-2016_consuni.pdf.