UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

Bacharelado em Sistemas de Informação

Rayane Cordeiro Barbosa

ANÁLISE DE DESEMPENHO DO MODELO CASSIOPEIA NA CLUSTERIZAÇÃO DE DADOS DA REDE SOCIAL X

Diamantina-MG

2025

Rayane Cordeiro Barbosa

ANÁLISE DE DESEMPENHO DO MODELO CASSIOPEIA NA CLUSTERIZAÇÃO DE DADOS DA REDE SOCIAL X

Trabalho de Conclusão de Curso apresentado a Faculdade de Ciências Exatas da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como parte dos requisitos para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Marcus Vinicius Carvalho Guelpeli.

Diamantina-MG

2025

AGRADECIMENTOS

Gostaria inicialmente de agradecer aos meus pais e meus irmãos, Jusciane e Luiz, pelo suporte e encorajamento constantes que me acompanharam ao longo de toda essa trajetória. Agradeço também meus primos, que sempre me apoiaram e incentivaram, bem como os meus amigos de infância e faculdade, que tornaram essa caminhada até aqui mais leve e significativa.

Agradeço imensamente aos docentes do curso pela transmissão de conhecimento e faço um agradecimento em especial ao meu orientador, professor Dr. Marcus Guelpeli, pela dedicação, paciência e orientações valiosas, que foram fundamentais para a realização deste trabalho. Por fim, e acima de tudo, agradeço a Deus por me conceder força, sabedoria e oportunidades que tornaram possível a concretização desta pesquisa.

RESUMO

O crescimento do volume de dados sociais não estruturados gerados em plataformas digitais, como o X, tem ampliado a necessidade de investigar métodos capazes de identificar padrões e extrair informações relevantes das publicações criadas nesta rede social, que tem como característica textos ruidosos, heterogêneos e com tamanhos diversos. Diante desse cenário, este estudo buscou avaliar o desempenho do modelo de clusterização Cassiopeia, proposto por Guelpeli (2012), verificando sua capacidade de formar agrupamentos semanticamente consistentes e úteis para a descoberta de conhecimento. Para isso, foi construído um *corpus* composto por 735 publicações relacionadas a instituições financeiras, coletadas por meio da técnica *Web Scraping* e submetidas às etapas de pré-processamento, processamento e pós-processamento. O desempenho foi avaliado por meio de métricas internas, a partir da realização de 200 simulações com o uso de uma lista de *Stop Words* como parâmetro de filtragem textual. Os resultados apresentaram valores entre 77,5% e 78,1%, indicando elevada coesão interna e boa separação entre os *clusters*. A inspeção manual de grupos selecionados confirmou a presença de padrões temáticos claros, evidenciando que o Cassiopeia é capaz de lidar com textos do X, produzindo agrupamentos significativos sem técnicas adicionais de enriquecimento textual ou rotulagem de dados. Conclui-se que o modelo apresenta potencial relevante para aplicações em mineração de textos provenientes do X.

Palavras-chave: Mineração de textos, clusterização, Cassiopeia, descoberta de conhecimento, X.

ABSTRACT

The growth in the volume of unstructured social data generated on digital platforms, such as X, has increased the need to investigate methods capable of identifying patterns and extracting relevant information from publications created on this social network, which is characterized by noisy, heterogeneous texts of varying lengths. Given this scenario, this study sought to evaluate the performance of the Cassiopeia clustering model, proposed by Guelpeli (2012), verifying its ability to form semantically consistent and useful clusters for knowledge discovery. To this end, a corpus was constructed consisting of 735 publications related to financial institutions, collected using the Web Scraping technique and subjected to pre-processing, processing, and post-processing stages. Performance was evaluated using internal metrics, based on 200 simulations using a Stop Words list as a textual filtering parameter. The results showed valuesbetween 77.5% and 78.1%, indicating high internal cohesion and good separation between clusters. Manual inspection of selected groups confirmed the presence of clear thematic patterns, demonstrating that Cassiopeia is capable of handling X texts, producing meaningful groupings without additional text enrichment techniques or data labeling. It is concluded that the model shows significant potential for applications in mining texts from X.

Keywords: Text mining, clustering, Cassiopeia, knowledge Discovery, X.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

A	Acoplamento
C	Coesão
CE	Caracteres e Espaços
PN	Palavras e Números
S	Coeficiente de Silhoutte
URLs	Uniform Resource Locator
VBA	Visual Basic for Applications

SUMÁRIO

1 INTRODUÇÃO

O crescimento acelerado das tecnologias digitais e o uso massivo das redes sociais transformaram profundamente a forma como informações são produzidas, disseminadas e consumidas. Plataformas como o X, tornaram-se importantes espaços de comunicação e expressão, gerando diariamente grandes volumes de dados textuais que refletem opiniões, comportamentos e tendências da sociedade contemporânea. Embora esse enorme fluxo de informações representem uma fonte valiosa de informação, sua natureza informal, ruidosa, não estruturada e por vezes limitada em extensão representa um desafio a modelos tradicionais de mineração de texto. Surge, portanto, a necessidade de investigar métodos capazes de lidar com essas características de maneira eficiente e consistente.

Esse cenário de crescente produção de dados sociais tem ampliado o interesse por técnicas que permitam identificar padrões temáticos e extrair conhecimento a partir de grandes bases textuais. Entretanto, abordagens tradicionais, como a utilização de modelos de aprendizado de máquina supervisionado e o uso de técnicas de incorporação de palavras, mostram-se custosas e por vezes pouco eficientes nesse contexto (Meddeb; Romdhane, 2022). A rotulação manual, necessária para modelos supervisionados, torna-se inviável frente ao alto volume de dados e demanda esforço humano significativo; já a incorporação de palavras, embora útil para lidar com escassez semântica, altera o conteúdo textual original, introduz custo computacional adicional e é sensível ao ruído característico dos textos sociais. Diante dessas limitações, surge a motivação para investigar se um modelo não supervisionado, como o Cassiopeia, pode oferecer uma alternativa mais simples, robusta e alinhada às propriedades dos dados do X.

Assim, a questão central desta pesquisa é: o modelo Cassiopeia apresenta desempenho adequado para clusterizar dados textuais não estruturados da plataforma X? Para responder a essa pergunta, parte-se da hipótese de que o Cassiopeia, ao utilizar uma lista de *Stop Words* para reduzir ruídos, é capaz de formar agrupamentos semanticamente coesos e gerar descoberta de conhecimento, ao observar o desempenho da métrica interna e as análises manuais nos *clusters* formados.

O estudo apresenta três contribuições principais: (i) a aplicação do modelo Cassiopeia no *corpus* criado com os dados não estruturados da rede social X; (ii) a avaliação do desempenho do modelo observando a medida harmônica Coeficiente de *Silhoutte*, ao longo

de 200 simulações, fazendo-se o uso de uma lista de *Stop Words*; (iii) e a análise da coerência temática e descoberta de conhecimento dos agrupamentos formados.

2 REFERENCIAL TEÓRICO

2.1 Big data e os dados sociais

O *Big Data* é gerado a partir de uma ampla gama de fontes, como conteúdos produzidos por usuários, mídias sociais, transações digitais e sensores inteligentes, caracterizando-se pelos cinco atributos centrais conhecidos como os 5Vs, Volume, Velocidade, Variedade, Veracidade e Valor, que refletem a complexidade e o potencial analítico desses dados (Memon *et al.*, 2024). De acordo com George, Haas e Pentland (2014), o *Big Data* compreende diferentes tipos de dados, os públicos, privados, de exaustão, de autoquantificação e os dados da comunidade, também chamados de dados sociais, que consistem principalmente em informações não estruturadas provenientes de interações sociais mediadas por tecnologia digital, como postagens e avaliações em mídias sociais. Tão importante quanto a origem dos dados são as metodologias e técnicas empregadas para analisá-los, dentre as quais destacam-se o processamento de linguagem natural, a análise de *cluster*, a mineração de texto e o aprendizado de máquina.

Com a ampla difusão da internet e o uso intenso das redes sociais, o X consolidou-se como uma das principais plataformas de comunicação em tempo real, onde milhões de usuários compartilham diariamente pensamentos, opiniões, sentimentos e notícias em suas publicações. Essa interação contínua gera uma quantidade imensa de informações que, quando reunidas, formam um retrato dinâmico do comportamento social e das tendências contemporâneas. Onde, cada postagem realizada em uma plataforma digital representa uma unidade de dado social, isto é, uma informação produzida voluntária ou involuntariamente pelos usuários, que se acumula em grandes bases de dados digitais (Memon *et al.*, 2024). De acordo com Basit e Albarry (2024), os dados não estruturados provenientes das mídias sociais são ricos em informações e constituídos por diversos formatos de conteúdo, como textos, imagens, vídeos e interações entre usuários, possuindo um grande potencial para a identificação de padrões, tendências e comportamentos sociais.

O estudo desses dados oferece múltiplas possibilidades de aplicação, desde a interpretação de sentimentos expressos em publicações até a previsão de tendências sociais, políticas ou mercadológicas. Assim, o X torna-se um campo fértil para a observação de

fenômenos sociais em larga escala, uma vez que cada interação dos usuários representa uma fonte potencial de conhecimento. Conforme ressaltam Bello-Orgaz, Jung e Camacho (2016), a análise do *Big Data Social* configura-se como uma área interdisciplinar que busca transformar o fluxo informacional das redes em conhecimento útil sobre o comportamento humano e as dinâmicas comunicacionais da sociedade conectada.

2.2 Mineração de texto

A mineração de texto em *Big Data Social* surge como uma ferramenta essencial para compreender o vasto volume de informações geradas nas interações digitais contemporâneas. Essa ferramenta, também conhecida pelo termo em inglês *Text Mining*, é um campo voltado à extração e análise dos textos ou frases provenientes dos dados não estruturados e advém da mineração de dados que é uma técnica voltada para a identificação de padrões em conjuntos de dados altamente organizados. Essa abordagem aplicada a textos consiste em empregar técnicas computacionais para navegar, organizar e identificar padrões em bases textuais, permitindo descobrir relações e significados implícitos nos dados.

A mineração de texto abrange diversas abordagens metodológicas que variam conforme o objetivo da análise e o tipo de dado trabalhado. Entre as principais, destaca-se a busca por palavra-chave, voltada à identificação de termos específicos em grandes volumes de texto; a classificação, que consiste em atribuir rótulos ou categorias predefinidas aos documentos; e a clusterização, que agrupa automaticamente textos com características semelhantes, revelando padrões temáticos. Ademais, temos a sumarização que busca gerar resumos automáticos a partir do conteúdo original; a extração que se concentra na identificação de informações relevantes, como entidades, nomes e relações; e o Processamento de Linguagem Natural que fornece a base computacional que permite compreender a estrutura e o significado da linguagem humana. Dessa forma, a mineração de texto consolida-se como uma ferramenta fundamental na era do *Big Data*, ao integrar linguística, estatística e inteligência artificial para transformar dados em conhecimento, tornando possível a análise e interpretação de grandes volumes de informações textuais de maneira sistemática e eficiente (Karami *et al.,* 2020).

A aplicação de técnicas de mineração de texto permite extrair padrões, percepções e informações estratégicas além de apoiar decisões gerenciais, a partir dos dados não estruturados, como aqueles encontrados em postagens de mídias sociais e documentos *web*

(Basit; Albarry, 2024). Algoritmos específicos podem ser utilizados para análise de sentimento, modelagem de tópicos e extração de conhecimento, afim de compreender por exemplo a posição do seu público-alvo sobre determinados temas, identificar as preferências dos clientes, apoiar na realização de campanhas mais assertivas, orientar como deve se posicionar ou publicar conteúdos para fortalecer sua marca, saber a opinião da sociedade sobre determinadas políticas ou medidas públicas e ainda descobrir oportunidades de inovação em produtos ou serviços, identificar áreas com maior engajamento da comunidade e antecipar tendências de comportamento do público. O conhecimento extraído destes grandes volumes de dados dispersos na internet oferece possibilidades altamente valiosas, possibilitando que organizações e governos ajustem suas estratégias de forma mais eficaz e direcionada.

2.2.1 Dados estruturados e não estruturados

A aplicação das técnicas de mineração de texto em dados não estruturados e a mineração de dados em dados estruturados, contribui para a ampliação da capacidade analítica, promovendo interpretações mais precisas e aprofundadas.

No contexto atual de crescimento exponencial das informações digitais, impulsionado principalmente pelas interações em plataformas como o X, compreender a natureza dos dados tornou-se essencial para a análise informacional e o desenvolvimento de sistemas inteligentes. As publicações, respostas e menções realizadas diariamente pelos usuários dessa rede social resultam em um grande volume de dados que podem conter diferentes formatos (áudio, vídeo, imagem, texto, metadados) e níveis de organização (dados estruturados, semiestruturados, não estruturados) em uma única publicação.

Quanto ao nível de organização eles podem ser estruturados, não estruturados e semiestruturados. Nesta pesquisa, temos como enfoque os dados estruturados para contextualização e os não estruturados. Os dados estruturados são aqueles que seguem uma formatação padronizada, como data, endereço e número de telefone, onde são armazenados de modo organizado, geralmente em bancos de dados relacionais, possibilitando sua manipulação e recuperação de forma eficiente.

Os dados não estruturados caracterizam-se pela ausência de um modelo fixo ou formato padronizado, o que impede seu processamento e compreensão direta pelos modelos tradicionais de dados estruturados (Zhang *et al.*, 2017). Uma parte expressiva desses dados

são provenientes de mídias sociais, sendo compostos por textos e/ou elementos ruidosos como *emoticons*, abreviações, *URLs*, *hashtags* e menções que embora ricos em significado, exigem técnicas específicas para sua interpretação e tratamento (Meddeb; Romdhane, 2022). A descoberta de conhecimento a partir desses dados demanda etapas de extração, processamento e análise, representando um desafio devido à sua natureza heterogênea, ao grande volume de informações e às questões de qualidade, privacidade e relevância dos dados (Bello-Orgaz; Jung; Camacho, 2016).

Apesar dos avanços na análise de dados estruturados, ainda persistem desafios relacionados ao tratamento de grandes volumes de dados não estruturados, cuja complexidade dificulta a interpretação e análise em larga escala, podendo, entretanto, ser explorados por meio de técnicas de mineração de texto que possibilitam a extração de conhecimento relevante (Memon *et al.*, 2024; Ingole; Bhoir; Vidhate, 2018; Yordanova; Stefanova, 2017).

2.2.1.1 Plataforma X

O X, uma das plataformas de *microblogging* mais utilizadas no mundo, permite que seus usuários publiquem mensagens curtas e longas em tempo real sobre temas pessoais e sociais. As interações contínuas entre milhões de usuários geram um volume expressivo de dados não estruturados, compostos por textos, símbolos e outros elementos comunicativos. Esse fluxo informacional heterogêneo configura-se como uma fonte valiosa para pesquisas em mineração de texto.

De acordo com dados recentes do Statista, o X apresenta uma expressiva presença global, contabilizando cerca de 557 milhões de usuários ativos mensais em fevereiro de 2025, posicionando-se entre as principais redes sociais mais populares do mundo (Dixon, 2025). No Brasil, a plataforma reúne 16 milhões de usuários, situando-se entre os dez maiores públicos de publicidade e destacando-se pelo tempo médio mensal de uso de mais de 4 horas por usuário ativo no aplicativo móvel em celulares *android* segundo o relatório da DataReportal, o que evidencia o seu potencial como plataforma geradora de dados sociais para pesquisas em larga escala (Kemp, 2025).

Os dados desta rede têm sido amplamente explorados por pesquisadores para abordar diversas questões, como a análise de sentimentos/opiniões, a análise de contexto e a análise de tópicos, também chamada de extração, modelagem ou detecção de tópicos.

2.2.1.2 Web Scraping

O avanço das tecnologias digitais e o aumento expressivo do conteúdo disponível na internet impulsionaram o desenvolvimento de métodos voltados à coleta automatizada de informações. Nesse contexto, destaca-se o *Web Scraping* (Raspagem da *Web*), técnica essencial para a obtenção de dados provenientes de ambientes online de forma automatizada usando *softwares* ou *bots*. O *Web Scraping* na plataforma X, consiste na extração automatizada de informações públicas disponíveis nas postagens, perfis e interações dos usuários, visando coletar grandes volumes de dados para análise de forma sistemática e escalável. Segundo Shahade *et al.* (2023), o *Web Scraping* consiste em um processo sistemático de extração de dados de páginas *web*, no qual informações disponíveis publicamente são capturadas e transformadas de um formato não estruturado para um formato estruturado, os quais podem posteriormente ser organizados em planilhas ou bases de dados.

O *Web Scraping* integra-se às etapas iniciais de processos analíticos mais amplos, como a mineração de texto, servindo como instrumento fundamental para a coleta automatizada de conteúdo gerado pelos usuários, por exemplo. O método permite extrair grandes quantidades de dados de forma eficiente, convertendo-os em formatos estruturados que favorecem a aplicação de técnicas de aprendizado de máquina e de processamento de linguagem natural. Essa técnica é amplamente utilizada para subsidiar estudos em diferentes áreas. Shahade *et al.* (2023) ressalta que, ao ser automatizado, o processo torna-se menos suscetível a erros humanos e mais ágil, garantindo maior precisão na obtenção dos dados. Além disso, a técnica possibilita a transformação de informações dispersas como as do X, em conjuntos organizados e padronizados, contribuindo para que organizações possam obter vantagem competitiva no mercado com a análise desses dados.

Dessa forma, o *Web Scraping* apresenta-se como uma etapa importante dentro do fluxo de coleta de dados em larga escala em ambientes digitais, ao contribuir para a estruturação dos dados em pesquisas acadêmicas e mercadológicas.

2.2.1.3 Corpus

A criação de um *corpus* pode ser realizada por meio da coleta de dados textuais públicos não estruturados utilizando a técnica de *Web Scraping*. Esse processo possibilita a formação de grandes *corporas*, compostos por postagens e comentários dos usuários, extraídos automaticamente das plataformas digitais.

O conceito de *corpus* ou *corpora* está diretamente associado à constituição de bases textuais utilizadas para análise linguística e mineração de textos. Sardinha (2004) explica que a Linguística de *Corpus* tem como foco extrair e examinar conjuntos de textos. A extensão de um *corpus*, segundo Sardinha (2004), pode ser avaliada por três dimensões: o número de palavras, que indica a representatividade lexical; o número de textos que correspondem ao objetivo da *corpora*; e o número de tipos textuais que se refere a diversidade da língua. A partir dessas dimensões, é possível perceber tanto a relevância do *corpus* como fonte de dados reais da linguagem, quanto a importância de investigar a frequência de ocorrência de elementos linguísticos para estimar padrões e probabilidades.

Com o *corpus* criado, podemos aplicar 3 macroetapas para obter informações sobre a *corpora*. Inicialmente temos o processo de pré-processamento, que consiste na limpeza e preparação dos textos para serem processados e analisados. Nessa etapa, pode ser realizado a remoção de ruídos, a padronização de caracteres, a tokenização (separação do texto em palavras ou unidades menores), a eliminação de palavras irrelevantes (*Stop Words*), a redução das palavras à sua forma base por meio de lematização ou *stemming* e a eliminação de arquivos que estão em dissonância com o objetivo da pesquisa. Após esse processo, ocorre o processamento, etapa em que os textos tratados são analisados por meio de modelos ou sistemas específicos que possibilitam por meio de alguma técnica de mineração de texto extrair informações úteis que antes estavam ocultas. Por fim, realiza-se o pós-processamento, que envolve a organização, interpretação e visualização dos resultados obtidos, possibilitando identificar tendências, relações e padrões presentes naquele *corpus*.

A criação de um *corpus* constitui-se uma etapa fundamental para qualquer investigação baseada em mineração de textos, pois é nesse momento que se forma a base de dados textual que servirá como fonte para a extração de conhecimento. Essa construção exige planejamento e rigor metodológico, garantindo que o material coletado seja adequado às finalidades analíticas do estudo.

2.3 Métodos de clusterização

A clusterização ou agrupamento, é uma técnica central na mineração de texto, que pode ser aplicada em *corporas*, tendo como objetivo organizar objetos ou documentos desse *corpus* em grupos (*clusters*) que compartilham características semelhantes, ao mesmo tempo, em que se diferenciam de outros grupos (Arora; Deepali; Varshney, 2016). Ou seja, quanto

mais semelhantes forem os elementos dentro de um mesmo *cluster* e mais diferentes forem os *clusters* entre si, mais bem definido será o agrupamento obtido. Essa técnica é classificada como aprendizado não supervisionado, uma vez que não requer conhecimento prévio sobre os rótulos ou categorias, permitindo que os dados sejam organizados com base em sua similaridade intrínseca (Bello-Orgaz; Jung; Camacho, 2016).

Os algoritmos de clusterização podem ser classificados em duas categorias principais: os métodos particionais (não-hierárquicos) e os hierárquicos (partição hierárquica). Segundo Rejito, Atthariq, Abdullah (2021) e Guelpeli (2012) o algoritmo *K-Means* é um método de agrupamento particional que divide o conjunto de dados em k agrupamentos distintos, sendo necessário definir previamente o número de *clusters* (k) antes do início do processo. O algoritmo inicia com a seleção aleatória de k centroides iniciais, que representam os centros provisórios de cada grupo. Em seguida, cada instância do conjunto de dados é atribuída ao *cluster* cujo centroide esteja mais próximo, geralmente com base na distância euclidiana. Após essa etapa, os centroides são recalculados a partir da média dos valores das observações pertencentes a cada agrupamento. Esse processo é iterativo e se repete até que os centroides se estabilizem, ou seja, quando não há mais alterações significativas na composição dos *clusters*, indicando a convergência do algoritmo. Os algoritmos de agrupamento hierárquico baseiam-se na construção de uma estrutura hierárquica de relações entre os dados, permitindo visualizar o processo de formação dos grupos em diferentes níveis de similaridade. Esses algoritmos são classificados em duas abordagens principais: aglomerativa e divisiva. No método aglomerativo, o processo inicia considerando cada elemento como um *cluster* individual, que é progressivamente unido a outros com maior grau de similaridade, formando agrupamentos cada vez maiores até atingir um critério de parada estabelecido (Bibi *et al.*, 2022; Guelpeli, 2012). Já o método divisivo adota a estratégia oposta: parte de um único agrupamento contendo todos os elementos e realiza divisões sucessivas com base em medidas de dissimilaridade, gerando grupos menores até que determinada condição de término seja alcançada (Bibi *et al.*, 2022; Guelpeli, 2012). Essa abordagem hierárquica possibilita uma representação em forma de dendrograma, que evidencia as relações de proximidade entre os dados e a estrutura de agrupamento formada ao longo do processo (Guelpeli, 2012; Rezende; Marcacini; Moura, 2011).

Estudos aplicados demonstram que a clusterização permite organizar grandes corpora de textos acadêmicos ou de mídias sociais em grupos temáticos, facilitando a descoberta de padrões, a identificação de tendências e o suporte à tomada de decisão em

ambientes acadêmicos, corporativos e governamentais. Assim, a clusterização configura-se como uma ferramenta estratégica na mineração de texto, capaz de revelar relações e informações ocultas dos documentos da base textual, apoiando análises quantitativas e qualitativas em diversos contextos de pesquisa e aplicação prática.

2.3.1 Métricas para avaliação dos clusters

Ao empregar a técnica de clusterização no *corpus*, são formados os agrupamentos textuais que refletem similaridades semânticas entre os documentos analisados. A partir desses agrupamentos gerados pelo modelo em questão utilizado, o próximo passo consiste na realização de análises que permitam compreender a qualidade dos *clusters* obtidos. Essa etapa é fundamental para validar a eficácia do modelo escolhido.

A análise dos agrupamentos textuais gerados por algoritmos de clusterização, como o modelo Cassiopeia proposto por Guelpeli (2012), pode ser realizada por meio de métricas internas (não supervisionadas), que avaliam o quão eficientemente os documentos foram agrupados, sendo essenciais para validar a eficácia do método.

2.3.1.1 Métricas internas

As métricas internas avaliam a qualidade dos *clusters* exclusivamente com base nas informações contidas nos próprios agrupamentos, sem recorrer a qualquer referência externa ou classificação prévia. As medidas utilizadas neste contexto são Coesão, Acoplamento e Coeficiente de *Silhouette*.

A Coesão (C) calcula quanto os textos de um mesmo grupo são parecidos entre si, ou seja, quanto mais semelhantes os documentos deste *cluster*, maior a coesão. Quanto menor a distância entre os pontos dentro de um *cluster*, maior a coesão, devido o fato deles estarem próximos ao centroide. Na Equação 1, $Sim(P_i, P_j)$ representa a similaridade entre os textos i e j dentro do agrupamento P, n é o número total de textos em P, e P_i e P_j são documentos pertencentes ao mesmo agrupamento.

$$C = \frac{\Sigma_{i>j} \mathrm{Sim}(P_i, P_j)}{\frac{n(n-1)}{2}} \quad (1)$$

Já o Acoplamento (A) mede o quanto os textos de um grupo estão isolados um do outro, quanto menor o valor resultante da fórmula do acoplamento, mais distintos são os grupos entre si. Na Equação 2, *C* representa o centroide de um agrupamento *P*, *Sim* (C_i, C_j), é a similaridade entre o centroide C_i do agrupamento *P* e o centroide C_j de outro agrupamento P_i e n_a é o número total de agrupamentos considerados em *P*.

$$A = \frac{\Sigma_{i>j} \mathrm{Sim}(C_i, C_j)}{\frac{n_a(n_a-1)}{2}} \quad (2)$$

A medida harmônica da Coesão e do Acoplamento é o Coeficiente de *Silhouette*, que mensura simultaneamente a coesão interna e o acoplamento externo dos *clusters*. O coeficiente de *Silhouette* (S) é então definido como na Equação 3, variando entre 0 e 1, sendo *a(i)* a distância média entre o elemento *i-ésimo* e os demais do mesmo grupo, *b(i)* a menor distância entre o elemento *i-ésimo* e qualquer outro grupo ao qual ele não pertence, e *max* representa o maior valor entre *a(i)* e *b(i)*.

$$S = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (3)$$

Valores próximos de 1 indicam que o elemento está bem alocado em seu *cluster*, ou seja, ele está bem definido e homogêneo, ao mesmo tempo que adequadamente separados de outros grupos, enquanto valores próximos de zero sugerem que o elemento está na fronteira entre dois *clusters* ou possível alocação incorreta, sendo o ideal, *clusters* com alta coesão e baixo acoplamento na avaliação da medida harmônica (Coeficiente de *Silhouette*).

2.3.2 Cassiopeia

Com o *corpus* formado pelos dados sociais não estruturados da plataforma X, obtidos por meio da técnica de *Web Scraping*, aplicou-se um modelo de mineração de texto baseado na técnica de clusterização, para agrupar estes documentos textuais. Neste trabalho, foi utilizado o modelo clusterizador denominado Cassiopeia.

O modelo proposto por Guelpeli (2012), em sua tese apresenta um sistema de agrupamento de textos estruturado em três macroetapas: pré-processamento, processamento e

pós-processamento, cada uma composta por procedimentos específicos que visam superar desafios clássicos como alta dimensionalidade, esparsidade dos dados e dependência de domínio e idioma. O funcionamento do modelo inicia-se com a etapa de pré-processamento, na qual os textos-fonte são submetidos a uma preparação. Primeiramente, realiza-se a conversão de todos os caracteres para minúsculas (*case folding*) e a remoção de elementos não textuais, como figuras, tabelas e marcações, garantindo um formato compatível para o processamento computacional. Em seguida, aplica-se a sumarização automática, que consiste na extração das sentenças mais informativas de cada documento, reduzindo significativamente o número de palavras e, consequentemente, a dimensionalidade do espaço amostral. Essa redução é controlada por um grau de compressão, definido pelo usuário, que determina o percentual de sentenças a ser mantido no sumário.

A independência do idioma no modelo Cassiopeia é alcançada principalmente pela estratégia de manter as *Stop Words* durante o pré-processamento. Como a sumarização automática reduz o texto ao seu conteúdo mais relevante, a presença de *Stop Words* não prejudica a qualidade do agrupamento, tornando desnecessária a remoção manual dessas palavras, que normalmente exigiria listas específicas para cada idioma. Dessa forma, o modelo pode ser aplicado a textos em diferentes línguas sem adaptações, pois a seleção das palavras mais significativas ocorre de maneira automática e baseada na frequência relativa dentro de cada documento, independentemente do idioma utilizado.

Após o pré-processamento, o modelo avança para a etapa de processamento, que se inicia com a identificação e seleção dos atributos de cada texto. Para isso, calcula-se a frequência relativa de cada palavra, normalizando a frequência absoluta pelo número total de palavras do documento. As palavras são então ordenadas de forma decrescente, ou seja, da mais frequente para a menos frequente, pois esse ordenamento facilita a identificação das palavras que mais contribuem para o significado do texto e permite a aplicação eficiente do método de corte de Luhn. O Cassiopeia propõe um novo método de corte de Luhn, denominado corte médio, que difere dos cortes superior e inferior tradicionais. Nesse método, calcula-se a frequência média das palavras do documento e identifica-se a palavra cuja frequência mais se aproxima desse valor. A partir desse ponto central, selecionam-se as 24 palavras anteriores (à esquerda) e as 25 posteriores (à direita), formando um vetor de 50 palavras que representa o centroide semântico do documento. Esse vetor reduzido é fundamental para a eficiência computacional do modelo, pois diminui drasticamente a dimensionalidade e a esparsidade dos dados, facilitando o cálculo de similaridade entre textos.

Com os centroides definidos, o modelo emprega um método hierárquico aglomerativo para o agrupamento dos textos. Inicialmente, cada documento é considerado um agrupamento isolado. O algoritmo *Cliques* é utilizado para medir a similaridade entre os centroides, baseada no número de palavras comuns entre eles (Guelpeli, 2012). Os agrupamentos mais similares são fundidos iterativamente, e a cada fusão, um novo centroide é criado, composto pelas 25 palavras mais frequentes de cada agrupamento original, totalizando novamente 50 palavras. Esse processo de fusão e reagrupamento prossegue até que não haja mais alterações nos centroides, ou seja, até a estabilização dos agrupamentos. O método hierárquico aglomerativo adotado junto com a implementação da técnica *top-down* permite a formação de uma estrutura geral-específica, na qual agrupamentos mais gerais são progressivamente refinados em subagrupamentos mais específicos, refletindo diferentes níveis de similaridade temática entre os textos.

Na etapa de pós-processamento, o modelo organiza os textos agrupados e sumarizados em uma estrutura hierárquica, na qual cada agrupamento contém os textos-fonte e seus respectivos sumários. Essa organização facilita a recuperação de informações, permitindo consultas tanto a documentos mais gerais quanto a documentos mais específicos, de acordo com o interesse do usuário. A estrutura hierárquica resultante é especialmente útil para atenuar a sobrecarga informacional, pois apresenta os textos de forma condensada e agrupada por similaridade semântica. O clusterizador Cassiopeia, ao integrar sumarização automática, o novo método proposto de seleção estatística de atributos baseada no corte de *Luhn*, representação vetorial reduzida e agrupamento hierárquico aglomerativo, resulta em agrupamentos de alta coesão semântica, independência do domínio e do idioma, representando um avanço significativo para a área de mineração de textos ao agrupar de forma não supervisionada grandes coleções textuais.

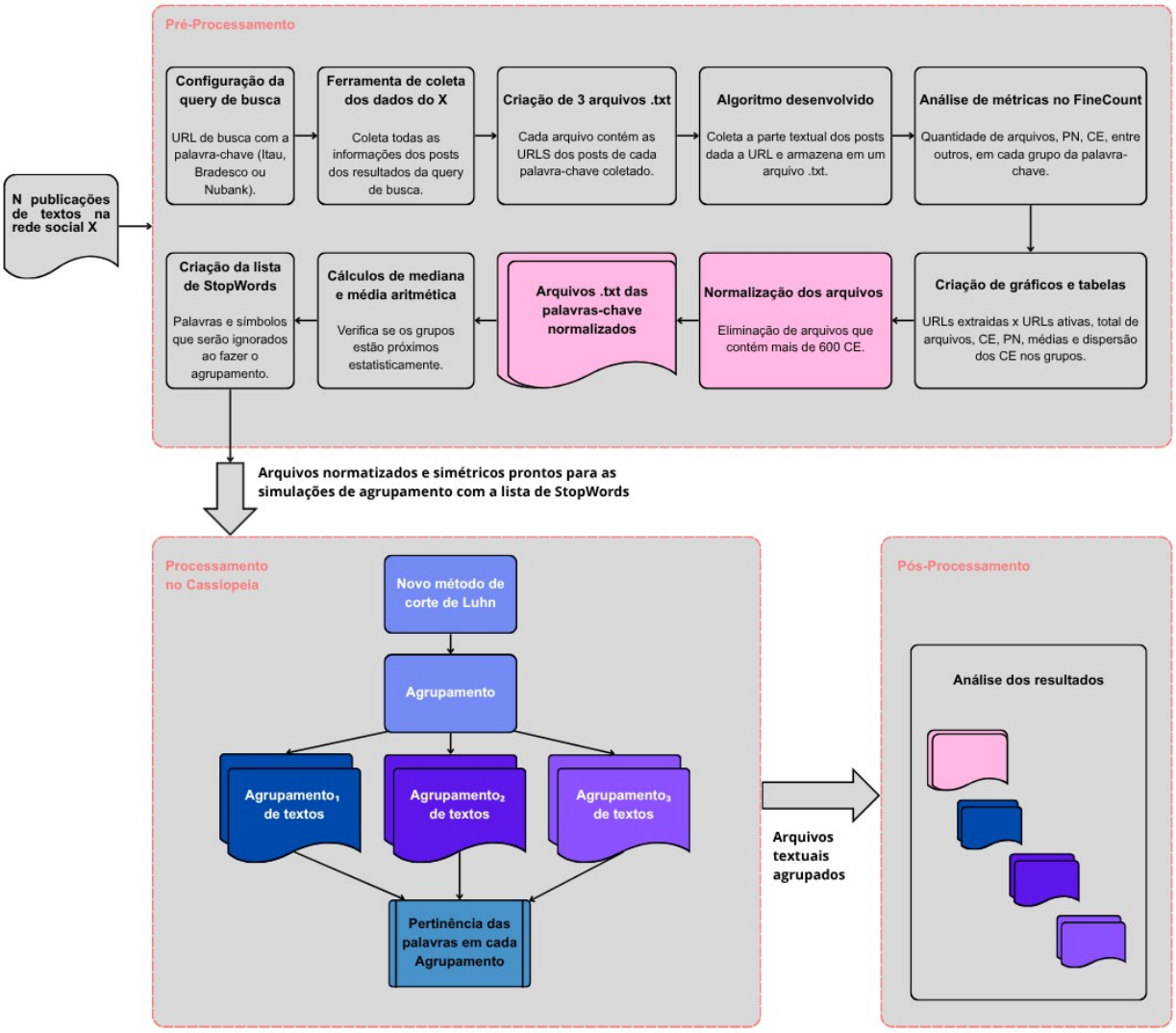
3 METODOLOGIA

A metodologia desta pesquisa, visa descobrir as métricas de desempenho internas do modelo Cassiopeia quando aplicado a dados da rede social X, além de verificar se há efetivamente a correlação destes números com os conteúdos dos *clusters*. A coleta dos dados foi iniciada em julho de 2024, observando o critério de pertencimento ao tema Bancos. A rede social X foi escolhida para a realização da raspagem (também conhecido como *Web Scraping*, técnica utilizada para a extração de dados na *Web*) devido o alto engajamento e à grande

quantidade de informações não estruturadas, que apresentam um vasto potencial para a descoberta de conhecimento. O trabalho tem como objetivo a criação de um *corpus* com dados sobre instituições financeiras do Brasil afim de averiguar a performance do modelo Cassiopeia ao agrupar estes documentos coletados do X. O *corpus* foi delimitado a este tema devido as características semelhantes que podem ser correlacionadas nos agrupamentos e por estarem englobados em um único tema geral (Bancos).

A estrutura da pesquisa, ilustrada na Figura 1, está dividida em três macroetapas (pré-processamento, processamento e pós processamento) que serão explicadas detalhadamente a seguir.

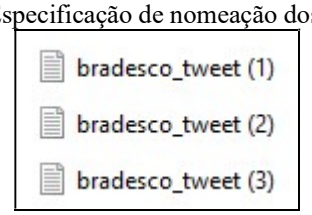
Figura 1. Diagrama da metodologia da pesquisa.



Fonte: Elaborado pelos autores.

A primeira etapa dessa metodologia foi definir os parâmetros de consulta da *URL* de busca com as palavras-chave Bradesco, Itau e Nubank, onde ao final foram extraídas 2.973 *URLs* da rede social X provenientes das publicações com estes termos. Para realizar essa extração, ao longo da pesquisa foi utilizado duas ferramentas na versão gratuita para a coleta dos dados. Inicialmente foi usado o *lobstr.io* (extraía os dados e gerava ao final um .csv), posteriormente devido a limitações na quantidade de dados que poderiam ser baixados gratuitamente foi continuada a coleta com o *phantombuster*, que exibia todas as informações dos posts incluindo as *URLs*. Entretanto, a parte textual do conteúdo do post não era coletada de forma completa, apenas parte dela e também havia restrições na quantidade de dados que podiam fazer *downloads*. Para sanar esse problema foi desenvolvido um algoritmo (https://github.com/cordeirorayane/Text-Discovery-in-X) utilizando *VBA* e *Selenium*, que dado o arquivo .txt que contém todas as *URLs* extraídas da palavra-chave, ele lê a *URL*, coleta a parte textual do post e o armazena em um arquivo .txt, seguindo a especificação de nomeação dos arquivos da seguinte forma: palavra-chave + _ + *tweet* + (número sequencial), como evidenciado na Figura 2.

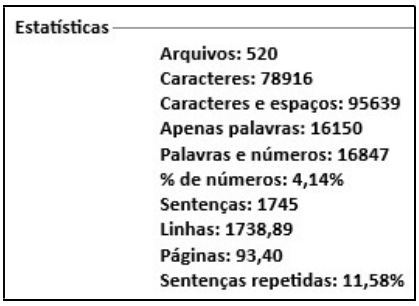
Figura 2. Especificação de nomeação dos arquivos.



Fonte: Elaborado pelos autores.

Nessa fase de Web *Scraping* foi desconsiderado o aprovisionamento de qualquer tipo de mídia (áudio, foto, *gif*), sendo relevante para este estudo apenas textos. Devido a fatores como exclusão de contas ou publicações, entre outros erros, houve uma perda de 15,14% das *URLs* sob o valor inicial. Efetivado essa fase, os arquivos .txt com os conteúdos textuais foram submetidos a análises de texto e contagem de palavras no *FineCount* (*software* profissional de contagem de palavras e análise de texto. Disponível em: https://www.finecount.eu) na versão 3.0 gratuita, vide Figura 3.

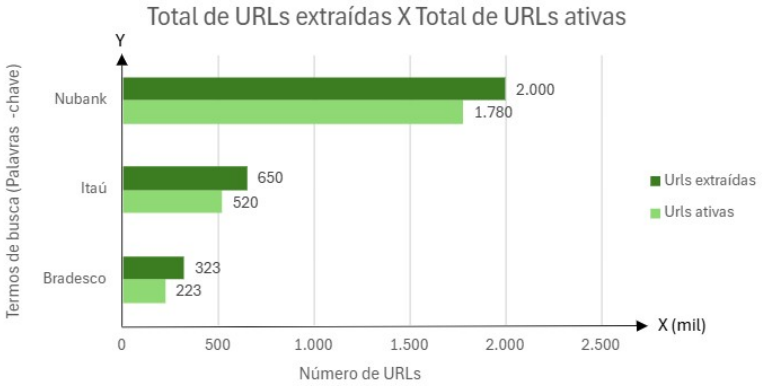
Figura 3. Métricas de análise observadas no FineCount 3.0 dos arquivos da palavra-chave Itau



Fonte: Elaborado pelos autores.

A imagem da Figura 4, faz o comparativo de quantas *URLs* foram extraídas por meio das ferramentas citadas anteriormente e quantas estavam realmente ativas quando foram submetidas no algoritmo criado pelo autor após o desbloqueio da rede social X no Brasil.

Figura 4. Comparação do total de *URLs* extraídas e ativas em cada termo de busca.



Fonte: Elaborado pelos autores.

O X possui diferentes limites de caracteres para contas gratuitas (280) e assinantes premium (até 25.000). No processo de contagem de caracteres é considerado o espaço como um caractere, diminuindo no limite disponível total. Para compreender se os arquivos textuais dos grupos dos termos de busca (Nubank, Bradesco e Itau), estão próximos do limite de caracteres estipulado pelo X e a quantidade de palavras e números em cada arquivo, foi calculado a média aritmética de cada um dos termos observando o total de quantidade de Caracteres e Espaços (CE), Palavras e Números (PN) em relação a quantidade de arquivos, conforme Tabela 1.

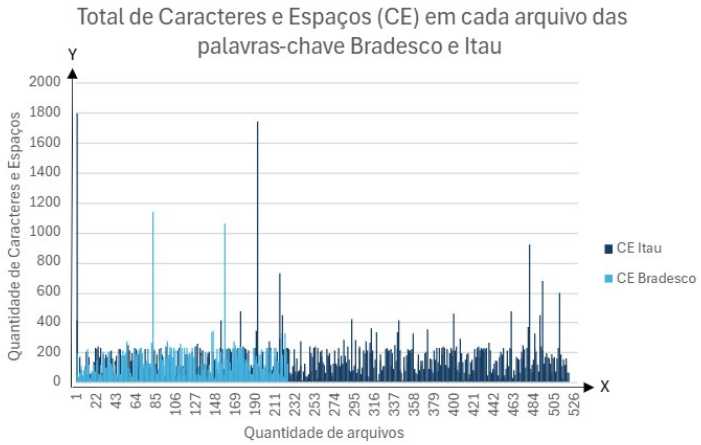
Tabela 1. Cálculos realizados nos arquivos textuais coletados.

Termos de pesquisa	Total de arquivos	Total de CE	Total de PN	Média de CE	Média de PN
Nubank	1780	191.987	35.510	107,857	19,949
Bradesco	223	39.073	6.969	175,215	31,251
Itau	520	95.639	16.847	183,921	32,398

Fonte: Elaborado pelos autores.

Após observação dos resultados, devido a alta discrepância dos valores da Nubank, o mesmo foi desconsiderado para este estudo devido os seguintes fatores: média de PN e CE do termo Nubank estarem baixas, evidenciando que os arquivos textuais apesar de serem muitos em quantidade, eram pobres em conteúdo textual; alto poder computacional em relação a combinação de arquivos até chegar na estabilização dos agrupamentos; Bradesco e Itau possuir médias próximas e estarem perto do limite de 280 caracteres estipulados para contas gratuitas no X. Em seguida, foi criado o gráfico da Figura 5, para observação de como está a dispersão da quantidade de CE nos arquivos dos grupos Bradesco e Itau.

Figura 5. Dispersão dos caracteres e espaços nos arquivos das palavras-chave Bradesco e Itau.



Fonte: Elaborado pelos autores.

Devido a presença de arquivos com altos valores de CE tanto no grupo Itau quanto no grupo Bradesco, na fase de normalização, um critério importante definido para esta etapa de pré-processamento foi a definição de um limite de 600 caracteres (incluindo espaço). Haja visto, que se uma base textual é composta por alguns arquivos que são muito grandes em detrimento da maioria que são curtos, estes podem ser mal alocados devido o seu extenso conteúdo informativo e estipulando um valor atenderia tanto as contas que possuem acesso as postagens curtas quanto as longas, à priori.

Após a eliminação dos 8 arquivos (2 do Bradesco e 6 do Itau) que ultrapassaram o limite de caracteres estipulado durante o processo de normalização, a base de dados textuais (*corpus*) antes composta de 743 arquivos provenientes da *URLs* ativa, passou a ser composta de 735 arquivos, conforme Figura 6, onde a pasta Bradesco contêm 221 arquivos,

provenientes das 223 *URLs* ativas do total de 323 extraídas e o Itau 514 arquivos, provenientes das 520 *URLs* ativas do total de 650 extraídas.

Figura 6. Quantidade atual de arquivos do *Corpus* da pesquisa.

Fonte: Elaborado pelos autores.

Com o *corpus* do trabalho finalizado, foram realizados cálculos de mediana e média aritmética das PN (palavras e números) para os 735 dados coletados do X, a fim de verificar se os grupos estão próximos estatisticamente para serem agrupados por um modelo clusterizador. Os resultados obtidos estão evidenciados na Tabela 2, onde foi possível verificar que tanto o Bradesco quanto o Itau possuem a mesma mediana das PN e para a média aritmética das PN a diferença entre eles é de menos 0,75.

Tabela 2. Cálculos de mediana e média aritmética realizados nos arquivos textuais dos termos.

Termos de pesquisa	Mediana das PN	Média aritmética das PN
Bradesco	28	29,488
Itau	28	30,221

Fonte: Elaborado pelos autores.

Utilizando neste cálculo o PN e não CE, é possível verificar se o conteúdo textual e numérico dos arquivos dos termos de busca é expressivo para que o modelo consiga clusterizar os arquivos achando correlações entre eles. Exemplo: Se a minha média aritmética das PN do Bradesco for 7 e do Itau 30, pode ser que não haja conteúdo suficiente para criar correlações entre os arquivos dos 2 grupos, ou seja, os agrupamentos podem ser formados por arquivos de um único grupo ou podem ser alocados em vários *clusters* devido a escassez de conteúdo que não gera tanta similaridade.

Após a aferição dos resultados da Tabela 2 que evidenciou a representatividade de forma justa dos grupos, ou seja, sem que houvesse altas distorções ou predominância de um grupo sobre o outro, foi iniciada as 200 simulações de clusterização nos dados do *corpus* do trabalho por meio do modelo Cassiopeia fazendo uso da lista de *Stop Words*. Este arquivo foi composto por termos comuns que são repetitivos como preposições, conjunções e artigos, as palavras-chave Itau e Bradesco que são muito frequentes, as *hashtags*, os sinais de interrogação e exclamação, risadas e abreviações.

4 RESULTADOS

Com o *corpus* devidamente normalizado e os conjuntos de dados bem representados, serão apresentados neste capítulo os resultados das simulações e as análises sobre o conteúdo alocado dentro dos *clusters* gerados pelo modelo Cassiopeia.

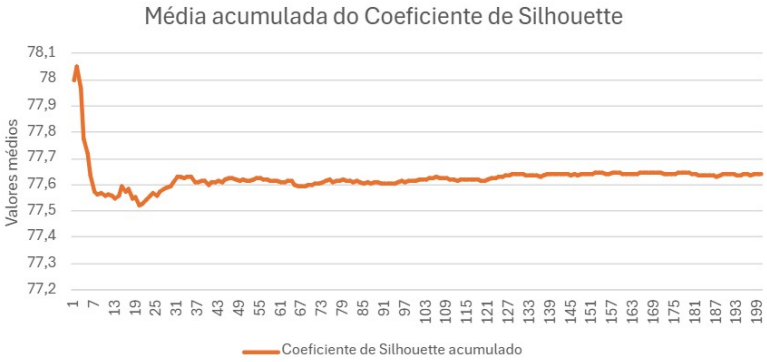
Na subseção 4.1, os 735 arquivos do *corpus* são submetidos ao modelo para a realização de simulações com o Coeficiente de *Silhouette*. Com os valores médios acumulados da medida interna obtidos ao processar todas as simulações, é analisado nos gráficos, os percentuais obtidos no eixo y e a linha de tendência no eixo x, afim de verificar o quão eficientemente o modelo agrupa, de forma não supervisionada, os arquivos coletados na plataforma X.

Na sequência, conforme descrito na subseção 4.2, são feitas as análises de alguns *clusters* gerados pelo modelo. Os grupos foram inspecionados de forma manual e aleatória, com o objetivo de verificar se os documentos alocados naquele grupo apresentavam correlação temática e permitiam uma descoberta de conhecimento. Para fins de ilustração e aprofundamento analítico, seis desses agrupamentos foram selecionados, abordando a quantidade de arquivos que compõem aquele *cluster*, o conteúdo textual de cada arquivo, a qual termo de busca eles se referem, o eixo temático e os termos mais frequentes.

4.1 Métrica interna

O resultado apresentado na Figura 7, que varia entre 77,5% e 78,1%, reflete os valores médios obtidos da métrica interna do Coeficiente de *Silhouette*, ao longo das 200 simulações que foram executadas com os dados do *corpus* fazendo-se o uso da lista de *Stop Words*. Nas primeiras simulações observa-se uma queda na medida harmônica interna, seguida pela estabilização dos valores ao longo das interações, evidenciada pela formação de uma linha de tendência até o final das simulações.

Figura 7. Resultados obtidos pelo Cassiopeia, usando a medida Coeficiente de *Silhouette*.



Fonte: Elaborado pelos autores.

4.2 Composição dos *clusters* e nuvem de palavras

Os seis *clusters* selecionados de forma aleatória e manual para análise estão demonstrados nas Figuras 8, 9, 10, 11, 12 e 13, que serão apresentadas a seguir. Cada figura exibe, à sua esquerda, os arquivos que integram o respectivo *cluster* acompanhados de seus conteúdos textuais, permitindo visualizar de forma direta os trechos que fundamentam a similaridade identificada pelo modelo. À direita, apresenta-se a nuvem de palavras gerada a partir desses mesmos arquivos, destacando os termos mais frequentes que contribuíram para a formação de cada agrupamento. Essa disposição visual foi utilizada para facilitar a compreensão da coerência interna dos *clusters*, evidenciando tanto a proximidade lexical entre os documentos quanto os principais elementos temáticos presentes em cada conjunto analisado. Os *clusters* analisados foram: 66, 70, 124, 208, 234 e 278.

O *cluster* 66 é constituído por três arquivos textuais da mesma palavra-chave (Itau), cujos conteúdos apresentam forte convergência temática relacionada ao mercado financeiro, com ênfase em análises de ações, desempenho setorial e comentários emitidos pelo Itaú BBA. A Figura 8 reúne os trechos textuais presentes nesses arquivos, acompanhados da nuvem de palavras que evidencia os termos mais frequentes, entre eles: ações, Itaú, BBA, alta, setor, mineradora, projeções, grafista e referências a empresas como Vivo, TIM e Aura Minerals.

Figura 8. Quantidade de arquivos no *cluster* 66 e os termos mais frequentes do agrupamento.

Fonte: Elaborado pelos autores.

O *cluster* 70 é composto por dois arquivos textuais de dois grupos diferentes (Itau e Bradesco), ambos relacionados a ocorrências de instabilidade em serviços bancários, especialmente situações em que sistemas ou operações ficaram “fora do ar”. A Figura 9 apresenta os trechos desses arquivos juntamente com a nuvem de palavras que evidencia os termos mais recorrentes, entre eles: Fora, ar, Pix, Itaú, Bradesco, Nubank, Caixa e Santander.

Figura 9. Quantidade de arquivos no *cluster* 70 e os termos mais frequentes do agrupamento.

Fonte: Elaborado pelos autores.

O *cluster* 124 é composto por três arquivos textuais da mesma palavra-chave (Itau), todos relacionados ao show da Madonna no Rio de Janeiro. A Figura 10 apresenta os trechos desses documentos, acompanhados da nuvem de palavras que evidencia os termos mais frequentes, entre eles: Madonna, Rio, Governo, Show, Não, Pago, Prefeitura, Heineken e Itaú.

Figura 10. Quantidade de arquivos no *cluster* 124 e os termos mais frequentes do agrupamento.

Fonte: Elaborado pelos autores.

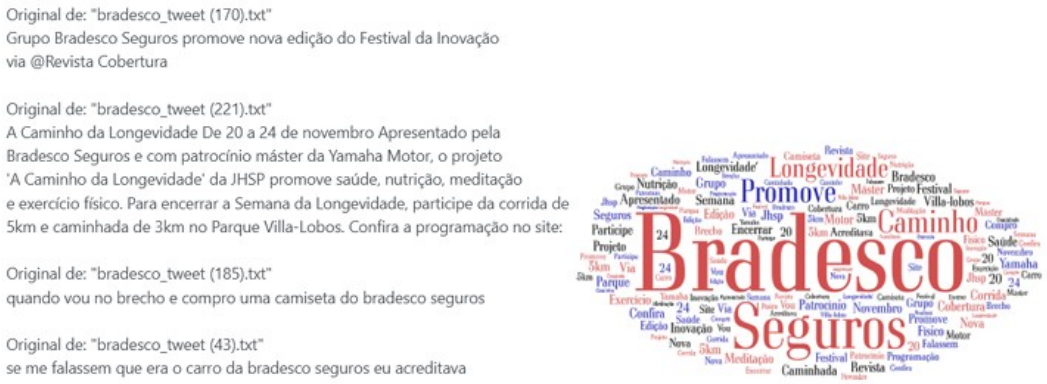
O *cluster* 208 é composto por três arquivos textuais da mesma palavra-chave (Itau), que apresentam forte convergência temática em torno da participação da artista Juliette no Rock in Rio, em ações promocionais vinculadas ao Itaú e ao TikTok. A Figura 11 reúne os trechos desses documentos e a respectiva nuvem de palavras, na qual se destacam termos como RockinRio, Juliette, Itaú, TikTok, Arena, Parceria, Domingo, Palco e 11.

Figura 11. Quantidade de arquivos no Cluster 208 e os termos mais frequentes do agrupamento.

Fonte: Elaborado pelos autores.

O *cluster* 234 é formado por quatro arquivos textuais vinculados à palavra-chave Bradesco, cujos conteúdos apresentam forte convergência temática centrada nas atividades da Bradesco Seguros, especialmente em eventos promocionais. A Figura 12 exibe os trechos presentes nos arquivos do *cluster* e a nuvem de palavras gerada a partir deles, evidenciando a predominância de termos como Bradesco, Seguros, Caminho, Longevidade, Promove, Festival, participação, saúde e bem-estar.

Figura 12. Quantidade de arquivos no Cluster 234 e os termos mais frequentes do agrupamento.



Fonte: Elaborado pelos autores.

O *cluster* 278 é composto por três arquivos textuais da palavra-chave Itau, cujos conteúdos apresentam elevada convergência temática sobre iniciativas de educação promovidas pela Fundação Itaú. A Figura 13 reúne os conteúdos textuais dos documentos agrupados no *cluster* e exibe a nuvem de palavras correspondente, destacando termos como Fundação, Itaú, Plataforma, Cursos, Gratuitos, Educação, professores e transformação.

Figura 13. Quantidade de arquivos no Cluster 278 e os termos mais frequentes do agrupamento.

Fonte: Elaborado pelos autores.

5 DISCUSSÃO DOS RESULTADOS

Serão apresentadas a seguir as discussões sobre as análises realizadas nos percentuais do Coeficiente de *Silhouette*, que refere-se ao desempenho interno do algoritmo ao realizar os agrupamentos; os pontos positivos e negativos de se utilizar uma lista de *Stop Words* no processo de simulações; a coesão dos elementos alocados nos grupos; o panorama geral sobre a utilização do Cassiopeia em textos longos e curtos; e a correlação existente entre os documentos que compõem os *clusters* junto com a descoberta de conhecimento antes implícita nos dados não estruturados.

5.1 Métrica interna

Os resultados obtidos da métrica interna Coeficiente de *Silhouette*, com os percentuais situados entre 77,5% e 78,1% nas 200 simulações, indicam uma alta coesão interna e boa separação entre os *clusters* do ponto de vista estrutural durante todas as interações, sem apresentar alterações bruscas. Isto é comprovado pela análise feita de forma manual e aleatória, que mostraram a correlação existente entre os documentos alocados e a descoberta de conhecimento proporcionada.

Ao utilizar a lista de *Stop Words* na metodologia desta pesquisa, o modelo fica dependente do idioma. Houve uma percepção que quanto melhor é o documento com as *Stop Words*, mais coeso os elementos textuais alocados nos grupos ficam. Nas simulações realizadas com a lista, há um aumento do número de *clusters* formados devido a quantidade de palavras que estão sendo ignoradas e o pouco conteúdo textual relevante que sobra dos dados que se traduzem em eixos temáticos (categorias dos *clusters*). Foi incorporado as palavras-chaves (Itau e Bradesco) na lista, por serem termos gerais e frequentes que acabam atrapalhando o algoritmo que utiliza a frequência de palavras ao fazer o agrupamento. A diminuição das ocorrências das palavras comuns é notada ao observar os termos mais frequentes do conjunto de todos os grupos gerados na guia do aplicativo que contém esta funcionalidade no modelo.

Apesar de o Cassiopeia ser sensível a textos curtos e coloquiais, o mesmo apresentou resultados significativos, proporcionando descoberta de conhecimento em dados da rede social X. O que demonstra a sua robustez tanto para textos longos quanto curtos, seja com a utilização de uma lista na geração das simulações ou com a retirada de todos os ruídos (palavras comuns, *links*, *hashtags*, abreviações, textos com idioma diversos, entre outros) na fase de normalização do *corpus*.

5.2 Composição dos *clusters* e nuvem de palavras

Com a formação de todos os agrupamentos realizados pelo Cassiopeia, tendo em vista, os resultados apresentados na seção 4.2, serão apresentados de forma mais detalhada e relacionada a formação dos grupos 66, 70, 124, 208, 234 e 278.

O *cluster* 66, ilustrado na Figura 8, reúne 3 arquivos do termo de busca Itau. O modelo clusterizador agrupou esses documentos devido à alta repetição de palavras como Ações, Itaú e BBA, indicando que todos tratam de análises de movimentos acionários baseadas nas projeções do Itaú BBA. A principal descoberta de conhecimento que pode ser extraída destes documentos é a convergência temática entre os textos, onde todos reforçam a influência das avaliações do Itaú BBA na interpretação do mercado financeiro.

O *cluster* 70, mostrado na Figura 9, contém 2 arquivos associados às palavras-chave Itau e Bradesco. Apesar de coletados por buscas diferentes, a repetição de termos como fora do ar, Pix e nomes de instituições financeiras levou o algoritmo a agrupá-los. A descoberta central é a identificação de uma narrativa comum sobre indisponibilidades digitais envolvendo múltiplos bancos, revelando possíveis padrões de percepção pública ou episódios paralelos de instabilidade nos serviços financeiros.

O *cluster* 124, ilustrado na Figura 10, abrange 3 arquivos obtidos pelo termo Itau. A alta frequência de palavras como Madonna, Show e Itaú evidencia um eixo temático voltado para o show da Madonna, contendo informações específicas como quem financiou o evento e possível doação de ingressos vip para o show.

O *cluster* 208, apresentado na Figura 11, agrupa 3 arquivos relacionados ao termo Itau. A repetição intensa de RockinRio, Juliette, Itaú e TikTok evidencia comunicações voltadas à presença de marcas em grandes eventos e ao uso de artistas (como Juliette) na estratégia de branding.

O *cluster* 234, exibido na Figura 12, contém 4 arquivos coletados a partir da palavra-chave Bradesco. Termos como Bradesco, Seguros, Caminho e Longevidade aparecem com alta frequência, revelando um núcleo temático sobre iniciativas do Bradesco Seguros em inovação e longevidade e engajamento da comunidade à marca. A descoberta de conhecimento destaca o protagonismo da empresa na comunicação voltada a ações sociais e envolvimento com o público.

O *cluster* 278, ilustrado na Figura 13, reúne 3 arquivos associados ao termo de busca Itau. A repetição de Fundação, Itaú, Plataforma e Cursos aponta para um eixo temático centrado em educação gratuita e divulgação da plataforma da Fundação Itaú. A descoberta mostra a consolidação pública da fundação como agente de transformação e inovação educacional.

Essas informações antes dispersas na rede social X, agora agrupadas e provendo descoberta de conhecimento ao analisar os agrupamentos formados, podem orientar

organizações de diversas formas. As análises do *cluster* 66 mostram que conteúdos relacionados a ações e projeções do Itaú BBA atraem interesse consistente, o que pode levar instituições financeiras a desenvolver serviços premium para clientes que buscam análises mais sofisticadas de investimentos. No caso do *cluster* 70, a recorrência de relatos sobre sistemas fora do ar pode ser usada por bancos para aprimorar a comunicação de incidentes e investir em resiliência tecnológica. Já o *cluster* 124 revela que eventos culturais de grande porte, como o show da Madonna, geram alto engajamento quando atrelados a benefícios gratuitos, incentivando marcas a reforçar estratégias de patrocínio cultural. O *cluster* 208 evidencia o impacto de artistas influentes em campanhas de *branding*, mostrando como celebridades podem ampliar a visibilidade de marcas em eventos como o *Rock in Rio*. O *cluster* 234 demonstra que iniciativas sociais e a presença da marca na internet podem fortalecer a imagem da empresa ao gerar reconhecimento e destaque além de promover interação com a comunidade. Por fim, o *cluster* 278 indica que plataformas educacionais gratuitas associadas a grandes instituições é bastante divulgada em redes sociais, o que eleva a presença e notoriedade da organização nestes meios digitais.

6 CONCLUSÃO

A presente pesquisa teve como propósito avaliar, por meio de métricas internas e análises manuais, o desempenho do modelo Cassiopeia na clusterização de dados provenientes da plataforma X, tomando como referência a hipótese de que o algoritmo seria capaz de gerar *clusters* distintos entre si, coesos internamente e capazes de revelar conhecimento fazendo uso de uma lista de *Stop Words*.

Os resultados obtidos evidenciaram o bom desempenho do modelo. A métrica interna apresentou valores elevados, variando entre 77,5% e 78,1%, o que indicou forte coesão dos elementos dentro dos *clusters* e clara distinção entre os grupos formados. A análise manual corroborou esses achados ao revelar agrupamentos temáticos consistentes, capazes de sintetizar eixos recorrentes presentes nos dados sociais, como instabilidade bancária, mercado financeiro, ações promocionais, iniciativas culturais e conteúdos educacionais. Esses resultados demonstraram que a hipótese inicialmente formulada foi confirmada.

Do ponto de vista interpretativo, os padrões observados mostraram que o Cassiopeia conseguiu lidar de forma significativa com as características desafiadoras dos

dados do X (textos curtos/longos e ruídos linguísticos), o que permitiu ao modelo organizar conteúdos dispersos e oferecer subsídios para a descoberta de conhecimento. O estudo demonstrou que, mesmo sem o uso de rotulagem de dados ou técnicas de incorporação de palavras, a aplicação do Cassiopeia na clusterização dos textos permitiu identificar eixos temáticos relevantes, potencialmente úteis para apoiar a tomada de decisão em organizações.

Entretanto, a utilização da lista de *Stop Words* tornou o processo dependente do idioma, configurando uma limitação do estudo. Além disso, o *corpus* permaneceu restrito a dois termos de busca, o que pode ter reduzido a diversidade temática dos agrupamentos.

Diante desses aspectos, recomenda-se que trabalhos futuros ampliem o *corpus*, incorporando mais bancos (palavras-chave) e textos com tamanhos variados, além de realizar novas simulações sem o uso de *Stop Words*, ao realizar de forma mais rigorosa o pré-processamento da corpora para a remoção de todos os ruídos. Tais iniciativas poderão fortalecer a independência linguística do modelo, permitir análises mais robustas e ampliar o entendimento sobre o comportamento do Cassiopeia em diferentes domínios da mineração de textos, contribuindo para aplicações práticas em cenários de grande volume de dados sociais não estruturados.

REFERÊNCIAS

ARORA, P.; DEEPALI, D.; VARSHNEY, S. **Analysis of k-means and k-medoids algorithm for big data.** Procedia Computer Science, v. 78, p. 507-512, 2016.

BASIT, M. Q.; ALBARRY, S. **Web mining algorithms for social media analytic**. International Journal of Computing Programming and Database Management, v. 5, n. 1, p. 52–54, 1 jan. 2024.

BELLO-ORGAZ, G.; JUNG, J. J.; CAMACHO, D. **Social big data: Recent achievements and new challenges**. Information Fusion, v. 28, n. 1, p. 45–59, mar. 2016.

BIBI, M. *et al*. **A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis**. Pattern Recognition Letters, v. 158, p. 80-86, 2022.

DIXON, S. J. **Most Popular Social Networks Worldwide as of February 2025, by Number of Monthly Active Users**. Disponível em: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users>. Acesso em: 10 out. 2025.

GEORGE, G.; HAAS, M. R.; PENTLAND, A. **Big Data and Management**. Academy of Management Journal, v. 57, n. 2, p. 321–326, abr. 2014.

GUELPELI, M. V. C. **Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização**. Niterói: Tese (Doutorado em Computação) - Universidade Federal Fluminense, 2012.

INGOLE, P.; BHOIR, S.; VIDHATE, A. V. **Hybrid Model for Text Classification**. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 1 mar. 2018.

KARAMI, A. *et al*. **Twitter and research: A systematic literature review through text mining**. IEEE Access, v. 8, p. 67698-67717, 2020.

KEMP, S. **X Users, Stats, Data & Trends for 2025**. Disponível em: <https://datareportal.com/essential-x-stats?utm_source=DataReportal&utm_medium=Country_Article_Hyperlink&utm_campaign=Digital_2025&utm_term=Brazil&utm_term=Brazil&utm_content=X_Stats_Link>. Acesso em: 10 out. 2025.

MEDDEB, A.; ROMDHANE, L. B. **Using topic modeling and word embedding for topic extraction in twitter**. Procedia Computer Science, v. 207, p. 790-799, 2022.

MEMON, A. B. *et al*. **A corpus-based real-time text classification and tagging approach for social data**. Frontiers in Computer Science, v. 6, p. 1294985, 2024.

REJITO, J.; ATTHARIQ, A.; ABDULLAH, A. S. **Application of text mining employing k-means algorithms for clustering tweets of Tokopedia**. In: Journal of Physics: Conference Series. IOP Publishing, 2021. p. 012019.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. **O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento**. Revista de Sistemas de Informação da FSMA, n. 7, p. 7-21, 2011.

SARDINHA, T. B. **Linguística de corpus**. Editora Manole Ltda, 2004.

SHAHADE, A. K. *et al*. **Multi-lingual opinion mining for social media discourses: an approach using deep learning based hybrid fine-tuned smith algorithm with adam optimizer**. International Journal of Information Management Data Insights, v. 3, n. 2, p. 100182, 2023.

YORDANOVA, S.; STEFANOVA, K. **Knowledge Discovery from Unstructured Data using Sentiment Analysis**. Ikonomiceski i Sotsialni Alternativi, n. 1, p. 13-27, 2017.

ZHANG, C. *et al*. **Study of Knowledge Discovery in Social Network Data Mining**. Journal of Computers, v. 28, n. 5, p. 105-115, 2017.