Rafael de Oliveira Taveira (Taveira, Rafael)

Análise de Sentimentos usando Centralidade de Palavras baseada em Feromônios de Colônia de Formigas

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Análise de Sentimentos usando Centralidade de Palavras baseada em Feromônios de Colônia de Formigas

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: André Ricardo Backes

Uberlândia
2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

T232 Taveira, Rafael de Oliveira, 1991-
2025 Análise de Sentimentos Usando Centralidade de Palavras Baseada em Feromônios de Colônia de Formigas [recurso eletrônico] / Rafael de Oliveira Taveira. - 2025.

Orientador: André Ricardo Backes.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2025.518
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Backes, André Ricardo,1981-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. III. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

Dedico este trabalho à minha família, cuja presença e apoio incondicional são a razão de cada esforço aqui empenhado.

Agradecimentos

Primeiramente, a Deus, o autor da vida, que cooperou para que eu chegasse até aqui, me sustentando em graça e misericórdia, apesar de mim, me capacitando a compreender e desenvolver cada parte deste trabalho.

À minha família, em especial à minha esposa Leidiane e aos meus filhos Eloá e Théo, que gentilmente cederam grande parte do tempo que deveria ser dedicado a eles, permitindo minha dedicação à capacitação acadêmica. Eles suportaram com paciência cada momento desafiador e celebraram comigo cada pequena conquista. Sem dúvida, este é o maior presente que eu poderia ter recebido.

Aos meus pais, que sempre estiveram presentes e investiram incondicionalmente em meus estudos. Este trabalho carrega, sem dúvida, uma significativa parcela do esforço e dedicação de cada um deles, que não se detiveram para que eu chegasse até aqui.

Ao meu orientador, Prof. Dr. André Ricardo Backes, que me acolheu sob sua orientação, conduzindo cada etapa desta pesquisa de maneira leve, mas sempre com seriedade e objetividade. Suas dicas e conselhos foram fundamentais para a concepção, desenvolvimento e conclusão deste projeto.

Ao meu amigo Dr. André Geus, que me incentivou a retomar a carreira acadêmica, orientando-me com sabedoria e recomendando a parceria com o Prof. André. Um verdadeiro amigo é aquele que se alegra ao ver o outro alcançar novos patamares.

Ao SiDi, instituto onde tenho trabalhado nos últimos anos, que sempre me apoiou ao longo desta jornada. Agradeço não apenas pela concessão de horas para que eu pudesse assistir às aulas, mesmo durante o horário de trabalho, mas também pela contribuição significativa de conhecimento e pela oportunidade de estabelecer amizades duradouras.

A todos os professores, técnicos e demais funcionários da Faculdade de Computação (FACOM) da Universidade Federal de Uberlândia (UFU), que contribuíram significativamente para o conhecimento adquirido e aplicado neste trabalho.

"E tudo quanto fizerdes, fazei-o de todo o coração, como ao Senhor, e não aos homens."
(Colossenses 3:23)

Resumo

As redes sociais revolucionaram a forma como nos comunicamos e expressamos nossas opiniões. Plataformas como Facebook, Twitter e Instagram permitem que as pessoas compartilhem suas ideias, sentimentos e experiências em tempo real, criando um fluxo constante de informações. Essa transformação não só facilitou a conexão entre indivíduos de diferentes partes do mundo, mas também democratizou a produção e o consumo de conteúdo, dando voz a uma diversidade de perspectivas que antes eram menos acessíveis. Técnicas de análise de sentimentos e mineração de texto permitem que pesquisadores e empresas identifiquem tendências, padrões e emoções expressas nas postagens. Isso pode ser utilizado para diversas finalidades, desde o aprimoramento de estratégias de marketing até a compreensão de questões sociais e políticas. Esse projeto desenvolve uma metodologia inovadora para análise de sentimentos em textos, utilizando a centralidade de palavras baseada em feromônios de colônia de formigas (ACO). A metodologia envolve a mineração de dados, construção de redes complexas e aplicação do ACO para identificar a centralidade das palavras. Inspirado no comportamento das formigas, o modelo permite uma análise profunda e contextualizada dos sentimento, onde os feromônios que são depositados indicam conexões emocionais. Os resultados dos experimentos demonstram que a combinação de ACO com outras medidas de centralidade, como *Katz*, *PageRank*, *EigenVector* e *Degree Centrality*, melhora significativamente a precisão da análise de sentimentos. Os testes realizados em diversos *datasets* mostraram que o uso de feromônios não apenas aumenta a acurácia, mas também o F1-score. Esses resultados indicam que a metodologia proposta é eficaz na captura das nuances emocionais presentes nos textos. A aplicação prática deste modelo pode beneficiar diversas áreas, como marketing, atendimento ao cliente e análise de tendências sociais, proporcionando *insights* valiosos sobre o comportamento e as preferências dos usuários.

Palavras-chave: Análise de sentimentos, Redes complexas, Feromônios, Centralidade.

Abstract

Social networks have revolutionized the way we communicate and express our opinions. Platforms such as Facebook, Twitter and Instagram allow people to share their ideas, feelings and experiences in real time, creating a constant flow of information. This transformation has not only made it easier for individuals from different parts of the world to connect, but it has also democratized the production and consumption of content, giving voice to a diversity of perspectives that were previously less accessible. Sentiment analysis and text mining techniques allow researchers and companies to identify trends, patterns and emotions expressed in posts. This can be used for a variety of purposes, from improving marketing strategies to understanding social and political issues. This project develops an innovative methodology for analyzing sentiment in texts, using word centrality based on ant colony pheromones (ACO). The methodology involves data mining, building complex networks and applying ACO to identify word centrality. Inspired by the behavior of ants, the model allows a deep and contextualized analysis of feelings, where the pheromones that are deposited indicate emotional connections. The results of the experiments show that combining ACO with other centrality measures, such as Katz, PageRank, EigenVector and Degree Centrality, significantly improves the accuracy of sentiment analysis. Tests carried out on various datasets showed that the use of pheromones not only increases accuracy, but also the F1-Score. These results indicate that the proposed methodology is effective in capturing the emotional nuances present in texts. The practical application of this model can benefit various areas, such as marketing, customer service and social trend analysis, providing valuable insights into user behavior and preferences.

Keywords: Sentiment analysis, Complex networks, Pheromones, Centrality.

Lista de ilustrações

Lista de tabelas

Lista de siglas

AS Análise de Sentimentos

ACO Otimização por Colônia de Formigas

AG Algoritmo Genético

API Interface de Programação de Aplicações

MO Mineração de Opinião

PLN Processamento de Linguagem Natural

SVM Máquinas de Vetores de Suporte

Sumário

Capítulo 1

Introdução

O advento das redes sociais transformou radicalmente o modo como as pessoas se expressam e se comunicam, proporcionando um meio dinâmico e interativo para as pessoas compartilharem informações, opiniões e experiências. Além de sua influência social, as redes sociais têm se destacado como uma fonte inestimável de dados, gerando um volume significativo de informações que, quando analisadas corretamente, oferecem *insights* valiosos sobre comportamentos, tendências e opiniões (ROTA; NICODEMO, 2023).

O grande volume de dados gerados nas redes sociais representa uma mina de informações valiosas para análises. A aplicação de técnicas avançadas de processamento de línguagem natural e aprendizado de máquina permite a extração de padrões, detecção de tendências e até mesmo a compreensão de sentimentos coletivos em escala global (CHAKRABORTY; BHATTACHARYYA; BAG, 2020).

Nesse contexto, os dados em formato de texto livre têm se destacado como um ponto de interesse para os pesquisadores na área de Processamento de Linguagem Natural (PLN). Sistemas baseados em PLN são sistemas capazes de auxiliar na extração e análise automática de grandes volumes de dados textuais. Isso é feito através de técnicas de mineração de texto, que permitem identificar e extrair informações relevantes a partir de textos não estruturados. Um dos objetivos é compreender as opiniões dos usuários sobre diversos assuntos, produtos ou serviços. Essa análise é valiosa, pois fornece *insights* diretos sobre a percepção do público, o que pode ser usado para melhorar produtos e serviços, desenvolver estratégias de marketing mais eficazes, entre outras aplicações (ROTA; NICODEMO, 2023).

Uma área específica de conhecimento que surgiu a partir desses estudos é a Análise de Sentimentos (AS), também conhecida como Mineração de Opinião (MO). A AS é uma técnica de PLN que visa identificar e classificar as opiniões expressas em um trecho de texto, especialmente para determinar a atitude do escritor em relação a algum tópico ou a polaridade geral do texto (YADOLLAHI; SHAHRAKI; ZAIANE, 2017).

As duas expressões AS ou MO são intercambiáveis. Eles expressam um significado mútuo. No entanto, alguns pesquisadores afirmaram que AS e MO têm noções ligeira-

mente diferentes (TSYTSARAU; PALPANAS, 2012). A MO extrai e analisa a opinião das pessoas sobre uma entidade, enquanto a AS identifica o sentimento expresso em um texto e o analisa. Portanto, o objetivo da AS é encontrar opiniões, identificar os sentimentos que elas expressam e, em seguida, classificar sua polaridade. Para esse trabalho, consideraremos apenas o conceito de AS, desenvolvendo a análise das polaridades positivas e negativas de um texto.

Além das abordagens estatísticas tradicionais, o PLN tem se voltado para métodos baseados em redes complexas. A representação de textos como grafos, especialmente o modelo de co-ocorrência, captura de forma intuitiva relações semânticas e sintáticas entre palavras, difíceis de representar em modelos vetoriais. Nesse modelo, palavras são nós conectados por arestas quando co-ocorrem, permitindo identificar padrões característicos de cada classe (F. de Arruda et al., 2019).

Existem vários algoritmos para percorrer e classificar uma rede complexa. Alguns dos mais conhecidos incluem Busca em Profundidade, Busca em Largura, Dijkstra e Kruskal (CHIRE; MAHMOOD; LIANG, 2025). Uma abordagem não convencional é a utilização da Otimização por Colônia de Formigas (do inglês *Ant Colony Optimization* - ACO), proposta por Dorigo (1992), que utiliza informação de feromônios virtuais para se mover dentro de um grafo. Este algoritmo é inspirado no comportamento das formigas na natureza, onde elas depositam feromônios ao longo do caminho para a fonte de alimento, permitindo que outras formigas sigam o mesmo caminho. No contexto das redes complexas, os "feromônios" são informações deixadas pelos caminhos mais bem-sucedidos, guiando a busca para soluções melhores. A principal vantagem deste algoritmo é sua capacidade de explorar o espaço de soluções de maneira eficiente, adaptando-se a mudanças e encontrando soluções robustas em ambientes dinâmicos.

Além disso, o ACO possui a vantagem de ser naturalmente paralelizável, o que significa que várias "formigas" (ou agentes) podem explorar diferentes partes da rede simultaneamente, aumentando a velocidade de convergência para uma solução ótima (PEAKE et al., 2022). O ACO também se destaca por sua flexibilidade, pois pode ser adaptada para resolver uma ampla variedade de problemas de otimização em redes complexas, como roteamento em redes de comunicação, planejamento de trajetórias e otimização de fluxos em sistemas de transporte (DORIGO; STÜTZLE, 2019). Essa flexibilidade é reforçada pela capacidade do algoritmo de balancear a exploração de novas soluções e a exploração de soluções conhecidas, o que ajuda a evitar que o processo de otimização fique preso em mínimos locais. No entanto, é importante notar que a eficiência do ACO pode depender da calibragem adequada de parâmetros, como a taxa de evaporação dos feromônios e a influência relativa da heurística de visibilidade, que precisam ser ajustados para o contexto específico da rede em estudo.

A identificação dos nós mais importantes em uma rede complexa é fundamental para entender sua estrutura e funcionamento. Para isso, são utilizadas diversas métricas de

centralidade, que quantificam a influência de cada nó no contexto da rede. Essas métricas podem ser calculadas com base em informações locais, como o número de conexões diretas de um nó, ou em informações globais, como sua posição em relação a todos os outros nós (SAXENA; IYENGAR, 2020). Ao combinar diferentes métricas, é possível obter uma visão mais completa e nuançada da estrutura e dinâmica da rede, permitindo identificar os nós mais influentes e os padrões de conexão mais relevantes.

Considerando o cenário onde os textos são representados por uma rede complexa sob um modelo de co-ocorrência, esse projeto tem como foco a utilização do sub produto de um ACO o feromônio, como uma medida de centralidade. A teoria por trás desta proposta é baseada na analogia com o comportamento das formigas, que utilizam feromônios para comunicar informações essenciais à sobrevivência da colônia. Analogamente, a aplicação de feromônios nas arestas de uma rede de palavras visa refletir a relevância e a frequência de associações emocionais. Esta abordagem dinâmica, ao capturar a intensidade e a interconexão dos sentimentos, visa identificar palavras-chave que desempenham papéis centrais na expressão emocional do texto, que pode por exemplo, ser utilizada na análise de sentimentos, classificando as polaridades expressas em um determinado texto.

1.1 Motivação

Ainda que muitas técnicas de análise de sentimentos tenham sido desenvolvidas ao longo da história, pouca literatura existe envolvendo algoritmos bio-inspirados para esse problema em questão. Em geral, há pesquisas utilizando-se do algoritmo de Boltzmann (PAN et al., 2004) e Redes Neurais (KIM; HOVY, 2014) e até mesmo envolvendo centralidade de palavras, porém utilizando-se de outras técnicas baseadas em Autovetor, Katz e PageRank (VILARINHO; RUIZ, 2018), assim como também técnicas modernas como *Machine Learning*, *Deep Learning* e *Large Language Models* (JIM et al., 2024). Assim, a utilização do ACO se torna um modelo pouco explorado para este problema.

A situação supracitada acontece pelo simples fato das técnicas atuais serem consolidadas, robustas e em grande parte, até mesmo mais simples de serem implementadas quando comparado a utilização do ACO.

Em geral quando utilizado um ACO o resultado importante é aquele definido pela função de avaliação do algoritmo. Ou seja, em um problema como o caixeiro viajante (*Travelling Salesman Problem - TSP*), roteamento de redes (*Network Design Problem*) ou roteamento de veiculos (*Vehicle Routing Problem - VRP*), o objetivo está sempre em descobrir a menor rota ou a mais eficiente, sendo que os dados gerados nesse processo são geralmente descartados.

Para isso, a utilização desse subproduto de dados chamado feromônio, pode ser também aproveitada para outras finalidades. Entende-se que o feromônio depositado pelas formigas atua como uma medida de centralidade útil, que captura aspectos diferentes dos

dados, podendo gerar novos *insights* sobre os resultados obtidos.

1.2 Objetivos

O método proposto busca preencher lacunas em abordagens tradicionais, tais como a análise de polaridade baseada em dicionários (PALTOGLOU; THELWALL, 2010), a contagem simples de termos positivos e negativos (AGARWAL; XU; LIU, 2011), técnicas de aprendizado de máquina supervisionado usando classificadores simples (KOTSIANTIS; KANELLOPOULOS; TAMPAKAS, 2007), a análise de frequência de termos (LEOPARDI; BISIO, 2015) e abordagens baseadas em regras heurísticas (ALVAREZ; VILARES, 2011). Estas técnicas, embora amplamente utilizadas, podem não ser suficientemente sensíveis para capturar as complexidades emocionais presentes em textos ou lidar com assuntos mais complexos e contextuais e até mesmo se limitar a evolução natural da línguagem.

A relevância desta pesquisa reside na sua contribuição para o avanço do campo de análise de sentimentos, apresentando uma metodologia inovadora que pode ser aplicada em diversos contextos, desde a análise de discursos políticos até a avaliação de sentimentos em mídias sociais. Além disso, a abordagem baseada em colônia de formigas e feromônios oferece uma perspectiva única para compreender a interação complexa entre palavras e emoções, enriquecendo a compreensão das nuances presentes na expressão humana através do texto.

O objetivo deste projeto é estabelecer um modelo de análise de sentimentos que incorpore um modelo natural baseado no comportamento das formigas. Formigas deixam rastros químicos, chamados feromônios, para se comunicarem e da mesma forma, podemos aplicar isso a palavras em um texto.

Cada palavra possui “feromônios” (marcas) que as unem emocionalmente com as palavras ao redor. Se uma palavra como “feliz” aparece frequentemente perto de “família” e “celebrar”, essas palavras podem ter uma conexão emocional forte. Então, ao seguir essas trilhas de “feromônios” nas palavras, podemos entender não apenas a intensidade do sentimento, mas também quais palavras são essenciais para expressar uma emoção específica.

Mais especificamente, tem-se por objetivos:

1. Investigar a eficiência de métodos de algoritmos inteligentes, para estabelecer um parâmetro de comparação para a técnica proposta.

2. Realizar, a partir do algoritmo de colônia de formigas aplicado a um conjunto de redes complexas, a coleta de uma tabela de feromônios que infere a centralidade das palavras a serem utilizadas na classificação de textos.

3. Analisar e avaliar os resultados das melhorias em relação ao estado da arte para o problema.

1.3 Hipótese

Este trabalho parte de duas hipóteses principais:

1. Algoritmos bio-inspirados podem obter resultados competitivos com algoritmos de centralidade do estado da arte em termos de métricas de avaliação como Acurácia e F1-score.

2. A utilização do feromônio resultante do algoritmo ACO como uma nova medida de centralidade de palavras pode produzir resultados competitivos, quando comparado a abordagens de centralidade tradicionais (como Katz e PageRank).

1.4 Contribuições

1. Um estudo avaliativo do algoritmo de otimização por colônia de formigas. O ACO foi aplicado a vários *datasets* possibilitando a utilização dos resultados em um analisador de sentimentos, cujos resultados foram promissores tanto individualmente quanto em conjunto com outras métricas.

2. Avaliação de outros quatro algoritmos de centralidade: Katz, PageRank, EigenVector, Degree Centrality, aplicados aos mesmos *datasets* afim de estabelecer um meio sólido de comparação.

3. Avaliação de várias combinações obtidas através da composição dos algoritmos de centralidade (ACO Katz, PageRank, EigenVector, Degree Centrality) por votação majoritária dos resultados, de modo a extrair melhores caracteristicas e maximizar o desempenho dos resultados.

4. A análise das matrizes de confusão geradas pelo analisador de sentimentos, assim como a discussão dos resultados e *insights* obtidos.

1.5 Organização da Dissertação

Esta dissertação está estruturada em seis capítulos. A seguir, é apresentada uma breve descrição de cada capítulo:

- ❑ No Capítulo 2 - Fundamentação Teórica: são apresentados e discutidos os principais conceitos que embasam este estudo, incluindo Mineração de Texto, Redes Complexas, Colônia de Formigas, Medidas de Centralidade e Análise de Sentimentos.

- No Capítulo 3 - Trabalhos Relacionados: são descritos os principais estudos e pesquisas que serviram de referência para este trabalho. Aqui, são analisados os métodos e abordagens semelhantes encontrados na literatura, destacando as lacunas e oportunidades que motivaram a condução deste estudo.

- No Capítulo 4 - Metodologia: são detalhadas as etapas metodológicas adotadas na pesquisa, incluindo a Mineração de Dados, a construção e utilização de Redes Complexas em conjunto com o ACO, os processos de pós-processamento e as técnicas de classificação e avaliação dos resultados obtidos.

- No Capítulo 5 - Experimentos e Análise dos Resultados: são apresentados os experimentos realizados para testar a eficácia das abordagens propostas. Neste capítulo, são discutidos os resultados das classificações realizadas utilizando o ACO e outras medidas de centralidade, além de uma análise comparativa entre essas técnicas.

- No Capítulo 6 - Conclusão: são apresentadas as conclusões deste estudo, destacando as principais contribuições teóricas e práticas. Também são discutidas as limitações do trabalho, propostas para pesquisas futuras e as contribuições em produção bibliográfica.

Capítulo 2

Fundamentação Teórica

Este capítulo aborda os principais conceitos e técnicas que fundamentam o presente projeto. Em particular, o Pré-processamento de textos, construção da Rede Complexa e ACO, como também a conceitos de AS.

2.1 Mineração de Texto

A era digital trouxe uma explosão de dados textuais. De e-mails e postagens nas redes sociais a artigos de pesquisa e notícias, o texto é onipresente em nosso mundo moderno. Este vasto tesouro de informações tem um enorme potencial para revelar *insights* valiosos, mas sua natureza não estruturada representa um desafio significativo. As técnicas tradicionais de análise de dados não estão preparadas para lidar com as complexidades da línguagem humana, tornando difícil extrair padrões significativos de grandes volumes de texto. De acordo com (ARANHA; PASSOS, 2006), mineração de texto não é um mecanismo de busca, no qual sabe-se o que necessita ser buscado, ou mesmo mineração de dados, no qual o dado pode ser genérico. Também não tem por objetivo imitar o comportamento humano como um *chatbot*, mas sim identificar padrões e anomalias presentes em textos.

Mineração de textos é um conjunto de métodos avançados utilizados para explorar, organizar, identificar e extrair informações valiosas a partir de grandes bases de dados textuais. Considerada uma extensão da área de *Data Mining*, essa técnica foca especificamente na análise de textos, possibilitando a descoberta de padrões e *insights* que seriam difíceis de identificar manualmente.

2.1.1 Coleta de Dados

Para coletar dados para mineração de texto, é preciso primeiro definir a fonte de dados, estas que podem ser de diversos locais e formatos. A *web* é um excelente ambiente para coleta de dados, pois contém diversas informações em formato de texto advindas das

redes sociais, jornais eletrônicos, fóruns, blogs, entre outros. Ambientes fora da internet também são fontes ricas, como bancos de dados de ONGs e órgãos governamentais.

A coleta de dados é o processo de reunir informações de diversas fontes para análise. No contexto da mineração de texto, isso envolve a extração de grandes volumes de texto para identificar padrões, tendências e *insights*. A razão para realizar a coleta de dados é múltipla: pode-se buscar entender comportamentos de usuários, analisar sentimentos em relação a um produto ou evento, ou até mesmo prever tendências futuras com base em dados históricos. A coleta de dados é essencial para alimentar algoritmos de *machine learning* e PLN, que dependem de grandes quantidades de dados para serem treinados e gerar resultados precisos.

O Twitter (renomeado para X em 24 de julho de 2023) é uma plataforma de mídia social que permite que os usuários publiquem mensagens curtas, conhecidas como *tweets*. Os *tweets* são públicos por padrão, o que os torna uma fonte valiosa de dados para a mineração de texto. Outro exemplo é o Kaggle, um site de competições e comunidades de dados que oferece uma variedade de conjuntos de dados para mineração de texto. Um desses conjuntos de dados é o descrito em (PONTES, 2023), que contém cerca de 800k tweets em português. A fim de superar limitações existentes na Interface de Programação de Aplicações (API) da plataforma (como quantidade de tweets e período), a utilização de bases de texto como esta fornecida na plataforma Kaggle simplifica o processo de coleta, minimizando assim esforço e salvaguardando recursos para os objetivos propostos.

Além do Twitter e do Kaggle, outras fontes valiosas para a coleta de dados textuais incluem o Reddit e o Google News. O Reddit é uma plataforma de fóruns onde os usuários discutem uma ampla variedade de tópicos, desde tecnologia até hobbies específicos. Através da API do Reddit, é possível extrair comentários e postagens de subreddits específicos, permitindo uma análise detalhada de comunidades e interesses particulares. Já o Google News oferece uma vasta quantidade de artigos de notícias de diversas fontes ao redor do mundo. Utilizando técnicas de web scraping ou uma API específica, é possível coletar textos de notícias para análises de tendências, eventos globais e estudos de mídia. Essas fontes adicionais ampliam as possibilidades de análise e enriquecem os *insights* obtidos a partir dos dados textuais.

O GitHub também é uma excelente fonte para a coleta de dados textuais, especialmente para análises relacionadas a desenvolvimento de software e colaboração em projetos de código aberto. No GitHub, é possível acessar uma vasta quantidade de repositórios que contêm não apenas código, mas também documentação, issues, pull requests e comentários de desenvolvedores. Esses textos podem ser analisados para entender padrões de colaboração, identificar problemas comuns enfrentados por desenvolvedores, ou até mesmo para estudar a evolução de projetos de software ao longo do tempo. Além disso, o GitHub oferece uma API robusta que permite a extração de dados de maneira estruturada e eficiente, proporcionando insights valiosos sobre práticas de desenvolvimento,

tendências tecnológicas e a dinâmica das comunidades de código aberto.

Por fim, existem diversas bases de dados públicas que foram extraídas e disponibilizadas na internet, facilitando o acesso a informações valiosas para análises. Exemplos incluem o *Common Crawl*, que oferece um vasto repositório de dados da *web* coletados regularmente, e o *Enron Email Dataset*, que contém milhões de emails de executivos da *Enron*, amplamente utilizado para pesquisas em processamento de linguagem natural. Outros exemplos notáveis são os datasets relacionados a eventos específicos, como o "Deflategate 2015", que contém discussões e tweets sobre um escândalo da NFL, o "GOP Debate", que reúne dados de debates políticos, e o "Coachella", que inclui tweets e posts sobre o famoso festival de música. Algumas das bases mencionadas aqui, serão apresentadas com mais detalhes nos experimentos propostos para esse projeto.

2.1.2 Balanceamento de Dados

Após a obtenção dos dados é necessário equilibrá-los. Se observarmos a quantidade de textos positivos e negativos em um *dataset*, dificilmente veremos uma igualdade, normalmente existirá uma diferença significativa entre eles. Essa diferença é prevista, visto que não se tem controle sobre os dados obtidos e previamente classificados. Segundo (JAPKOWICZ; STEPHEN, 2002), o desequilíbrio de classe é descrito como uma grande discrepância entre duas classes da variável-alvo, onde uma classe é representada por muitas instâncias, enquanto a outra é representada apenas por um pequeno número de instâncias. O desequilíbrio de classes leva um algoritmo de aprendizado de máquina a classificar erroneamente as instâncias da classe minoritária. Mesmo que o algoritmo apresente alta precisão, ele pode simplesmente classificar todas as instâncias como pertencentes à classe majoritária em um conjunto de dados desequilibrado. Isso acontece porque o algoritmo aprende mais com a classe majoritária devido ao seu grande número de amostras, enquanto reconhece menos a classe minoritária por ser menos representada (JOHNSON; KHOSHGOFTAAR, 2019). A solução deste problema consiste em redimensionar o conjunto de dados promovendo o equilibrio dessas classes. Superamostragem e subamostragem são exemplos citados por (LING; LI, 1998) que consistem respectivamente na adição ou remoção de amostragens. As métricas utilizadas podem variar, levando em conta a relevancia de cada amostra. Uma abordagem foi proposta por (BATISTA et al., 2004), um método de subamostragem baseado na detecção de *Tomek Links*. Um par de instâncias é um *Tomek Link* se elas são vizinhas mais próximas, mas pertencem a classes diferentes. Se duas amostras são um *Tomek Link*, significa que pelo menos uma delas é um ruído ou ambas estão na fronteira dos clusters. A reamostragem pode então ser feita removendo apenas as amostras pertencentes às classes marjoritárias ou removendo todos os pares *Tomek Links* do conjunto de dados. Outra abordagem é o algoritmo de *random undersampling* que envolve a remoção aleatória de exemplos da classe majoritária até que o número de exemplos em ambas as classes seja equilibrado (PRUSA et al., 2015). O ba-

lanceamento das classes pode melhorar a performance do modelo em termos de precisão para classes minoritárias, mas a perda de muitos dados pode impactar a capacidade geral do modelo de capturar nuances nos dados.

2.1.3 Pré-processamento de Texto

Dados coletados, geralmente são desestruturados e apresentam ruídos, imperfeições e inconsistências. Isso dificulta a aplicação de algoritmos de análise de texto, pois podem levar a resultados errados ou de baixa qualidade. A etapa de pré-processamento é essencial para a mineração de texto. Ela tem como objetivo preparar os dados para a etapa de classificação, eliminando informações irrelevantes ou alterando o formato dos dados para facilitar o processamento. O pré-processamento é composto por várias tarefas, que variam de acordo com as características dos dados. Por exemplo, dados coletados de redes sociais podem precisar de limpeza para remover ruídos e erros, enquanto dados de artigos científicos podem precisar de normalização para padronizar o vocabulário.

É importante ressaltar que textos são dados não estruturados, então é necessário estruturá-los. O processamento de texto começa dividindo-o em partes menores. Isso é possível porque a línguagem é composta de partes que se combinam para formar um todo. Depois de dividir o texto em partes léxicas, como palavras ou termos, essas partes são ligadas umas às outras por meio da sintaxe. A sintaxe é a estrutura do texto, que indica como as partes se relacionam, é a gramática das relações entre as palavras. Ela estuda como as palavras se combinam para formar frases e períodos, estabelecendo a ordem das palavras, a relação entre elas e o sentido que elas expressam (ARANHA; PASSOS, 2006).

2.1.3.1 Tokenização

A tokenização é o processo de quebrar um texto em unidades menores, chamadas de *tokens*. Os *tokens* podem ser palavras, frases ou até mesmo caracteres individuais.

A tokenização é uma etapa essencial no processamento de texto, pois permite que os computadores processem o texto de forma eficiente. O *token* podem ser usados para representar o texto em um formato que seja mais fácil de compreender e manipular pelos algoritmos de processamento de texto.

Alguns dos tipos de tokenização mais comuns incluem:

- ❏ Tokenização léxica: Divide o texto em palavras ou frases.
- ❏ Tokenização de caracteres: Divide o texto em caracteres individuais.
- ❏ Tokenização de N-gramas: Divide o texto em sequências de N *tokens*. É utilizado prefixos numéricos para representar a quantidade de *tokens*. Por exemplo, para um, dois, três ou quatro *tokens*, pode-se referir como unigramas, bigramas, trigramas e quadrigramas respectivamente.

A Tabela 1 apresenta um exemplo de tokenização para a seguinte frase: "O gato comeu o rato". A escolha do tipo de tokenização mais adequado dependerá do objetivo da aplicação. Por exemplo, segundo (PALMER, 2010), a tokenização léxica é geralmente usada para análise de sentimento, enquanto a tokenização de caracteres é geralmente usada para processamento de línguagem natural. N-gramas são úteis para capturar a relação entre palavras adjacentes, são utilizados para prever a próxima palavra. Essa previsão pode ser dada por n conjuntos de palavras, os bigramas por exemplo, são os pares consecutivos possiveis extraidos de um determinado texto.

Tabela 1 – Exemplo de tokenização

Tipo de tokenização	***tokens***
Léxico	O, gato, comeu, o, rato
Caracteres	O, g, a, t, o, , c, o, m, e, u, , o, r, a, t, o
N-gramas (bigramas)	o gato, gato comeu, comeu o, o rato

2.1.3.2 Remoção de *stopword*

A remoção de *stopword* é uma técnica para limpeza de palavras que possuem um elevado número de ocorrências em uma sentença ou documento. Geralmente são compostas de preposições, artigos, conjunções, adjetivos e advérbios que são consideradas irrelevantes para o entendimento, diminuindo assim o tamanho das estruturas e facilitando o processamento (BARION; LAGO, 2008).

Na prática, é utilizado uma lista de palavras consideradas irrelevantes no idioma em que se deseja trabalhar, e então quando houver a ocorrência de uma palavra existente na lista, esse termo será eliminado, como demonstrado na Tabela 2.

Tabela 2 – Exemplo de remoção de *stopwords*

	Frase
Original	Pedro usou o capacete quando andou de bicicleta
sem *stopwords*	Pedro usou capacete andou bicicleta

Ao tratar bases provenientes da internet, é importante acrescentar alguns outros filtros a fim de obter uma melhor base de dados para busca e indexação. Tais filtros podem ser:

- ❏ Remoção de links.
- ❏ Remoção de caracteres não alfabéticos.
- ❏ Remoção de pontuação.
- ❏ Remoção de acentuação.
- ❏ Remoção de repetições de letras (muito utilizado nas redes sociais para criar enfâse. Por exemplo: "Oláaa").

- ❏ Remoção de citações (Geralmente antecedidas por # ou @).
- ❏ Remoção de caracteres maiúsculos (substituindo pelo relativo minúsculo).

2.1.3.3 Lematização

A lematização é um processo de redução de uma palavra a sua forma básica, independentemente de sua conjugação ou flexão. O lema é a forma básica de uma palavra, que representa seu significado essencial (PLISSON; LAVRAC; MLADENIC, 2004). Em geral altera-se conjugação verbal para o infinitivo caso seja um verbo e altera-se os substantantivos e os adjetivos para o singular masculino, como demonstrado na Tabela 3.

Tabela 3 – Exemplo de lematização

Palavras	**Lema**
Viajando, Viajarei, Viajou, Viajem	Viajar
Gato, Gatuno, Gatinho	Gato

2.1.3.4 Stematização

A stematização é um processo de limpeza de palavras, que consiste em remover as desinências de gênero, número, tempo verbal e grau. O objetivo da stematização é reduzir o número de termos a serem indexados nas estruturas de indexação, facilitando a busca de informações. Os vocábulos processados são representados pelo seu radical. Técnicamente, busca-se por uma sub-palavra única e não ambígua que represente um vocábulo e suas variações, mais conhecida como radical (ALVARES, 2005). Utilizar este método é benéfico para identificar semelhanças morfológicas entre as palavras de um documento, além de simplificar a busca por termos equivalentes. A stematização é exemplificada pela Tabela 4.

Tabela 4 – Exemplo de stematização

Palavras	**Radical**
Viajando, Viajarei, Viajou, Viajem	Viaj
Gato, Gatuno, Gatinho	Gat

2.2 Redes Complexas

Segundo (ESTRADA, 2016), um grafo é uma estrutura matemática composta de objetos chamados vértices, conectados por arestas. A principal diferença entre um grafo e uma rede complexa é que um grafo é uma estrutura estática e simplificada, usada para modelar e analisar as propriedades estruturais de uma rede, enquanto as redes complexas incorporam a dinâmica e a complexidade das interações reais entre os elementos (SILVA; SANTOS et al., 2024).

Redes complexas são grafos que apresentam propriedades emergentes que não são facilmente previsíveis a partir de suas propriedades locais (Escalabilidade, Heterogeneidade e Auto-organização) e podem ser encontradas em uma ampla variedade de sistemas, incluindo redes sociais, redes biológicas e redes de transporte. São sistemas formados por um grande número de elementos interconectados, onde as conexões entre os elementos não seguem um padrão regular ou previsível. As arestas podem ser direcionadas ou não direcionadas, e podem representar relacionamentos de vários tipos, como amizade, parentesco ou conectividade. A complexidade dessas redes surge das interações dinâmicas e não-lineares entre os elementos. As redes complexas diferem significativamente dos grafos tradicionais devido às suas características estatísticas únicas.

Enquanto os grafos, como os modelos de Erdős e Rényi (ERDőS; RéNYI, 1959), são estruturas matemáticas estáticas que representam conexões entre vértices de forma aleatória, as redes complexas incorporam propriedades emergentes e padrões de conectividade que refletem a complexidade das interações reais. Muitas redes do mundo real, por exemplo, seguem uma distribuição de lei de potência, caracterizando-as como redes de livre escala (*scale-free networks*), um conceito amplamente explorado em (BARABÁSI, 2016). Essas redes exibem distribuições de grau não triviais, alta modularidade e a presença de hubs, que são vértices altamente conectados (BARABÁSI, 2016). Tais características são fundamentais para a análise e compreensão das dinâmicas de sistemas reais, como redes sociais, biológicas e tecnológicas. Estudos têm demonstrado que essas propriedades estatísticas são essenciais para capturar a topologia e a funcionalidade das redes complexas, diferenciando-as dos grafos tradicionais (RIBAS, 2021). Como exemplo, na Figura 1 e 2, vemos um grafo representando as pontes de *Königsberg* e uma rede complexa representando interações das proteínas da levedura.

Alguns termos podem ser elucidados da seguinte forma:

- Vértices: Os vértices em um grafo podem representar objetos de qualquer tipo, como pessoas, cidades ou genes.

- Arestas: As arestas em um grafo representam relacionamentos de qualquer tipo, como amizade, parentesco ou conectividade.

- Grafos direcionados e não direcionados: Em um grafo direcionado, as arestas têm direções. Isso significa que um vértice pode estar conectado a outro vértice, mas não o contrário. Em um grafo não direcionado, as arestas não têm direções. Isso significa que um vértice pode estar conectado a outro vértice da mesma forma que o outro vértice está conectado a ele.

- Escalabilidade: As propriedades das redes complexas podem crescer exponencialmente com o tamanho da rede. As redes complexas são geralmente caracterizadas

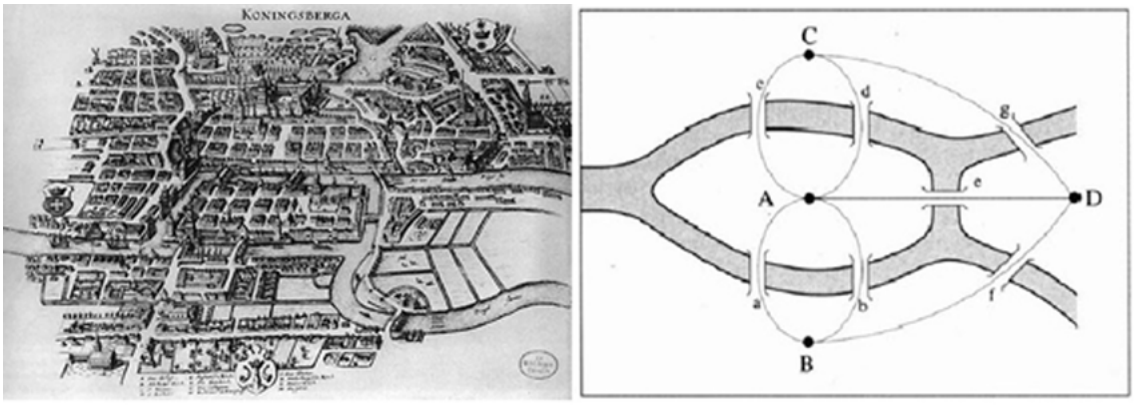

Figura 1 – Pontes de *Königsberg*.

Fonte: Soto (2013)

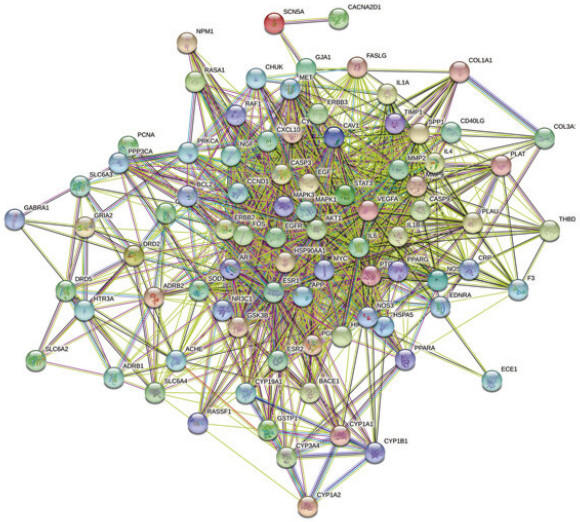

Figura 2 – Interações de Proteínas

Fonte: Liu et al. (2021)

por um alto grau de conectividade entre os seus vértices. Quanto maior o tamanho da rede, maior o número de vértices e de conexões entre eles.

- Heterogeneidade: Os vértices e arestas em redes complexas podem ter propriedades diferentes, tais como grau (alto numero de conexões), peso (importância do vértice) e direção (indica o fluxo da informação ou recurso).

- Auto-organização: As redes complexas são geralmente caracterizadas por um alto grau de conectividade e interação entre os seus vértices. Elas podem se organizar espontaneamente sem a necessidade de um controle externo.

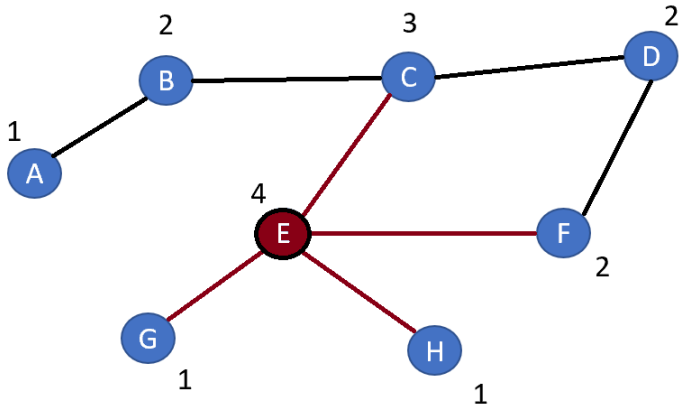

Figura 3 – Centralidade de vértices.

Fonte: O Autor

Segundo (KNöBEL, 2004), embora não seja uma definição precisa, uma rede complexa pode ser vista como uma rede que cresce de forma natural, ou seja, sua estrutura evolui de forma aleatória e constante, sem um plano de construção predefinido. Além disso, redes complexas são dinâmicas, com arestas e vértices sendo criados e removidos constantemente. Esse dinamismo garante a robustez da rede, pois ela pode se adaptar a mudanças no ambiente.

As redes complexas têm sido aplicadas nas mais diversas áreas, para a resolução dos mais variados tipos de problemas. Alguns exemplos são: Detecção de comunidades (GUL et al., 2022), avaliação da qualidade de textos (OLIVA et al., 2021), interações biológicas (PALUKURI; PATIL; MARCOTTE, 2023) e propagação de doenças (LI et al., 2023).

2.3 Medidas de Centralidade

De acordo com (SILVA; ZHAO, 2016), a centralidade de vértice é uma medida importante para a compreensão de redes complexas pois quantifica a importância de um vértice em uma rede, levando em consideração a conectividade do vértice com outros vértices da rede. Podemos inferir que um vértice que contém um grande número de arestas vinculadas é mais importante ou dito central para essa rede. Esses vértices podem ser chamados de *hubs*, que por exemplo, em um cenário de redes sociais seriam equivalentes as grandes empresas e influenciadores cujo conteúdo é capaz de atingir um maior número de pessoas.

Vemos, por exemplo, na Figura 3, o vértice mais central é o vértice E, que possui arestas com outros 4 vértices, enquanto A, G e H são os menos centrais por possuirem apenas 1 aresta vinculada.

Existem diversas formas de se medir a centralidade de um vértice. No entanto, podemos classifica-las de duas formas gerais:

- ❏ Centralidade local: Leva em consideração a conectividade dos vértices mais próximos, ou seja, dos indivíduos em que o vértice possui uma conexão direta.
- ❏ Centralidade global: Leva em consideração não apenas a conexão direta entre os vértices mais a conectividade de todos os vértices da rede.

Deste modo, podemos inferir que um vértice é globalmente central quando possui uma posição importante e/ou estratégica em seu contexto. Por outro lado, um vértice é localmente central quando a preocupação é relativa apenas aos vértices adjacentes.

Os algoritmos de centralidade são ferramentas fundamentais na análise de redes complexas, sendo utilizadas para identificar os vértices mais importantes nessa estrutura. Entre esses algoritmos, destacaremos quatro deles a seguir.

2.3.1 Katz Centrality

O Katz Centrality mede a influência de um vértice levando em consideração a importância dos seus vizinhos. Proposto em 1953 por Leo Katz (KATZ, 1953), essa medida avalia a influência de um vértice em uma rede, levando em consideração tanto as conexões diretas quanto as indiretas, ponderadas por um fator de atenuação sendo descrito pela equação (1).

$$C_K(i) = \sum_{j \neq i} \alpha^{d_{ij}} \cdot A_{ij} \tag{1}$$

onde d_{ij} é a distância entre os vértices i e j, A_{ij} é o valor da entrada ij da matriz de adjacência A, e α é o fator de atenuação ($0 < \alpha < 1$).

A Katz Centrality se distingue de outras medidas de centralidade porque considera o impacto cumulativo de estar conectado a outros vértices influentes na rede, além de valorizar as conexões mais próximas.

2.3.2 PageRank

O PageRank, inicialmente desenvolvido pelo Google em 1998 (BRIN; PAGE, 1998), atribui um valor de importância para cada vértice, refletindo a sua relevância. Essa atribuição leva em consideração a relevância dos vértices que apontam para ele e pode ser descrita pela equação (2).

$$PR(i) = \frac{1-\beta}{N} + \beta \sum_{j \in M(i)} \frac{PR(j)}{L(j)} \tag{2}$$

onde N é o número total de vértices na rede, β é o fator de amortecimento (normalmente 0.85), $M(i)$ é o conjunto de vértices que linkam para o vértice i, $L(j)$ é o número de links saindo do vértice j e $PR(j)$ é o PageRank do vértice j.

2.3.3 Eigenvector Centrality

O Eigenvector Centrality foi originalmente introduzido por Philip Bonacich em 1972 (BONACICH, 1972). É outro método que atribui uma pontuação a cada vértice, considerando não apenas a quantidade de conexões, mas também a centralidade dos vértices conectados, criando uma hierarquia de influência. Esse conceito tornou-se fundamental na análise de redes sociais e influenciou o desenvolvimento de algoritmos como o PageRank. A Eigenvector Centrality de um vértice i é descrita pela equação (3).

$$\mathbf{C}(i) = \frac{1}{\lambda} \sum_{j \in \mathrm{N}(i)} \mathbf{C}(j) \tag{3}$$

onde $\mathbf{C}(i)$ é o valor de centralidade do vértice i, $\mathrm{N}(i)$ é o conjunto de vizinhos do vértice i e λ é o maior valor próprio da matriz de adjacência A.

2.3.4 Degree Centrality

O Degree Centrality foi amplamente discutido e formalizado por Linton Freeman em 1979 (FREEMAN, 1979). É a medida mais simples, baseando-se no número de arestas que um vértice possui, o que pode indicar sua posição central em redes onde as conexões diretas são predominantes. Para um grafo não direcionado, é descrito pela equação (4).

$$C_D(i) = \deg(i) \tag{4}$$

onde i é simplesmente o número de arestas conectadas a esse vértice. Para um grafo direcionado, a equação (5) considera o número de arestas de entrada e saída.

$$C_D(i) = \text{in-deg}(i) + \text{out-deg}(i) \tag{5}$$

2.4 Colônia de Formigas

O algoritmo de otimização baseado em colônia de formigas (ACO, do inglês *ant colony optimization algorithm*) é uma heurística baseada em probabilidade, criada para solução de problemas computacionais que envolvem procura de caminhos. A proposta teve início nos anos 90, e sua concepção está diretamente ligada à observação do comportamento das formigas reais. O conceito original foi proposto por Marco Dorigo, um cientista da computação belga, que se inspirou no comportamento das formigas para desenvolver um método de otimização enquanto pesquisava na Universidade Livre de Bruxelas. Ele ficou intrigado com a eficiência em que as formigas encontram o caminho mais curto entre a colônia e uma fonte de alimento. As formigas depositam um feromônio químico ao longo do caminho que percorrem para encontrar alimento. Quanto mais formigas passam por

Figura 4 – Fluxograma de ACO

Fonte: Bedi e Singh (2013)

um caminho, mais forte o feromônio fica, o que atrai mais formigas para aquele caminho e influencia diretamente no comportamento coletivo.

O algoritmo começa com uma população de soluções aleatórias para o problema de otimização. Cada solução é representada por uma formiga. As formigas então exploram o espaço de busca, avaliando diferentes soluções. À medida que exploram, elas depositam um feromônio virtual ao longo do caminho que percorrem. Quanto melhor a solução que uma formiga encontra, mais feromônio será depositado neste caminho pelas formigas que convergirão para ele.

Após um certo período de tempo, as formigas começam a seguir os caminhos com mais feromônio, pois esses caminhos são mais propensos a levar a boas soluções. O algoritmo ACO continua iterando até que uma solução ótima seja encontrada, ou até que um critério de parada seja atingido. Segundo (BEDI; SINGH, 2013), o algoritmo pode ser descrito em forma de fluxograma, como mostrado na Figura 4.

A comunicação entre as formigas é feita indiretamente pelo depósito de feromônios durante o caminho, de modo a favorecer o melhor caminho a ser escolhido. Esse depósito

possui um valor inicial, que é um dos parâmetros do algoritmo e que é constante durante o primeiro ciclo, independentemente da quantidade de formigas. Após o primeiro ciclo, quando as formigas já percorreram o caminho, a quantidade de feromônios será proporcional ao valor da função objetivo, sendo possível obter esse cálculo de diversas formas. No entanto, esse valor necessita ser constante, ou seja, a cada ciclo será depositado o mesmo valor de feromônio no caminho utilizado pelas formigas para chegar ao ápice do resultado da função objetivo.

Segundo (ALOISE D., 2002), a evaporação do feromônio evita o seu acumulo exagerado e ilimitado, o que levaria a uma estagnação do processo de busca dentro do espaço de solução. Essa subtração, usualmente linear, é uma estratégia de reforço negativo, contribuindo para a escolha adequada entre as opções disponíveis para cada formiga em suas próximas iterações.

Dois fatores são imprecindiveis para o cálculo da probabilidade, o fator heurístico e o fator associativo do feromônio. Segundo (MULLEN et al., 2009), o fator heurístico é calculado antes mesmo do primeiro ciclo e permanece inalterado no andamento do algoritmo, enquanto o fator relacionado ao ferormônio é atualizado a cada iteração. O fator heurístico não é trivial para todos os problemas propostos, a sua escolha é um dos fatores que determinam o sucesso da abordagem utilizando ACO. Não se trata de uma escolha direta, pode ser relacionado a apenas uma variável, ou várias, o que vai depender da necessidade implícita de cada problema.

A probabilidade de uma formiga escolher a aresta é descrita na Equação (6).

$$p_{ij}(t) = \frac{\tau_{ij}(t)}{\sum_{k \in N(j)} \tau_{jk}(t) + \beta} \tag{6}$$

Onde a aresta (i, j) no passo t é representado por $p_{ij}(t)$. A quantidade de feromônio na aresta (i, j) no passo t é representada por $\tau_{ij}(t)$ e $N(j)$ é o conjunto de vizinhos do vértice j. O termo constante β pode ser usado para controlar o equilíbrio entre utilizar novos caminhos e utilizar caminhos conhecidos. Um valor alto de β incentiva a exploração, enquanto um valor baixo de β incentiva a estagnação. Segundo (DORIGO; BLUM, 2005), um valor de β entre 1 e 10 pode ser adequado para a maioria dos problemas. A probabilidade é proporcional à quantidade de feromônio na aresta. Quanto mais feromônio uma aresta tiver, maior a probabilidade de ser escolhida por uma formiga.

A atualização do feromônio é descrita pela Equação (7).

$$\tau_{ij}(t+1) = \tau_{ij}(t) + \alpha \cdot q_{ij}(t) \tag{7}$$

Onde $\tau_{ij}(t+1)$ indica a quantidade de feromônio na aresta (i, j) no passo $t+1$. A quantidade de feromônio na aresta (i, j) no passo t é representado por $\tau_{ij}(t)$. O parâmetro que controla a taxa de evaporação do feromônio é dado por α e $q_{ij}(t)$ é uma medida da qualidade da solução que passa pela aresta (i, j) no passo t.

A solução é descrita pela Equação (8).

$$f(s) = \sum_{(i,j)\in s} q_{ij}(t) \tag{8}$$

Onde uma medida da qualidade da aresta (i, j) na solução s é representada por $q_{ij}(t)$. A solução é a soma das avaliações das arestas na solução.

O número de formigas é descrito pela Equação (9).

$$n = \frac{N}{m} \tag{9}$$

Onde n controla o numero de formigas que irão percorrer a rede. Um número maior de formigas geralmente resulta em uma melhor qualidade da solução, mas também pode aumentar o tempo de execução do algoritmo. Sendo assim, temos n que é o número de formigas, N que é o número de vértices no problema e m que é o número de formigas por vértice.

A taxa de evaporação do ferormônio é representada pela Equação (10).

$$q(t) = q(t-1) * (1-\rho) \tag{10}$$

Onde $q(t-1)$ é a quantidade de feromônio na aresta no tempo $t-1$ e ρ é a taxa de evaporação, que é um valor entre 0 e 1 que representa a quantidade de feromônio que evapora em cada iteração do algoritmo.

2.5 Análise de Sentimentos

A AS foca em identificar opiniões, sentimentos, atitudes e emoções expressas em textos que mencionam entidades específicas, como eventos, produtos e serviços. Seu objetivo é classificar automaticamente o sentimento central contido no conteúdo analisado, um processo que exige o uso de técnicas de PLN e aprendizado de máquina para que o sistema possa interpretar as nuances das expressões humanas (LIU; ZHANG, 2012). Assim, a aplicação dessas técnicas permite a compreensão em larga escala de como as pessoas expressam suas opiniões sobre determinados tópicos.

Essa análise apresenta diversos desafios, como a detecção de ironia, a filtragem de spam e a identificação de opiniões tendenciosas. Sistemas automatizados são desenvolvidos para identificar o tom subjacente das discussões, geralmente classificando as opiniões em sentimentos positivos, negativos, podendo incluir a neutralidade como uma categoria adicional. Essa classificação ajuda a entender como os usuários percebem diferentes assuntos e interações no ambiente digital.

Um dos métodos centrais da AS é a classificação binária, representada pela Equação (11).

$$F : D \times C \rightarrow 0, 1 \tag{11}$$

Onde cada texto é categorizado como positivo ou negativo. O valor $F(d_i, c_j)$ é igual a 1 se o documento d_i pertence à classe c_j, e 0 caso contrário. Nesse modelo, d_i representa um documento dentro de um conjunto D, e c_j representa uma classe de sentimento dentro de C (FERREIRA, 2014). Este tipo de formalização é fundamental para o desenvolvimento de algoritmos capazes de realizar a classificação automática de sentimentos.

Portanto, a AS não é apenas um simples problema de classificação binária, mas envolve lidar com uma série de complicações linguísticas, como o sarcasmo e a subjetividade, que tornam o processo desafiador. Essa aprendizagem precisa ser continuamente refinada para melhorar a precisão e o entendimento de textos complexos.

Existem diversas abordagens para realizar a AS. Segundo (THAKKAR, 2013), a escolha da abordagem mais adequada depende da natureza da tarefa e do tipo de conhecimento que se deseja obter. Uma abordagem comum é a contagem de termos categorizados, onde as palavras são previamente classificadas como positivas, negativas ou neutras, permitindo que o sentimento de um texto seja determinado pela contagem dessas palavras (THAKKAR, 2013). Por outro lado, o aprendizado de máquina tem se destacado como uma abordagem eficaz na análise de sentimentos (JURAFSKY; MARTIN, 2024), a seleção de características é crucial nesse contexto. Algoritmos como Máquinas de Vetores de Suporte (SVM) e Naive Bayes são frequentemente aplicados, com resultados de precisão variando de acordo com a qualidade das características selecionadas. Além disso, métodos como o TF-IDF (*term frequency–inverse document frequency*), que contam a frequência de palavras no texto, podem melhorar a relevância de termos específicos em um determinado documento.

A abordagem baseada em modelos de linguagem também ganha destaque ao classificar o sentimento de textos com base na frequência de n-gramas. No entanto, a simples presença de unigramas, ou seja, *tokens* individuais que representam cada palavra válida presente no texto, pode ser mais útil em alguns cenários, como na classificação de blogs de filmes (JURAFSKY; MARTIN, 2024). Essa abordagem também pode ser complementada com a análise semântica, que considera o significado das palavras e frases em contextos específicos. A desambiguação semântica é essencial para garantir que termos com múltiplos sentidos sejam corretamente interpretados no contexto correto (JURAFSKY; MARTIN, 2024), especialmente em casos onde as frases carregam complexidade adicional, como a identificação de sarcasmo ou ironia.

Além disso, as abordagens de visualização, que utilizam grafos para representar redes de sentimentos, têm se mostrado promissoras no contexto das redes sociais. A análise de redes sociais, que conecta tweets com base em menções ou seguidores, pode revelar padrões de influência e opiniões predominantes (GIACHANOU; CRESTANI, 2016). A formação de grafos rotulados com *tokens*, como palavras, emojis ou *hashtags*, permite que algoritmos de análise de sentimentos identifiquem a polaridade das conexões entre tweets, determinando o sentimento geral da rede (AIELLO et al., 2013). Esses métodos visuais

oferecem uma maneira intuitiva de compreender a dinâmica dos sentimentos dentro de grandes volumes de dados.

2.6 Métricas de Desempenho

A matriz de confusão foi empregada para comparar os resultados das diferentes abordagens de centralidade. Ela oferece uma visão detalhada dos verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, como exemplificado pela Tabela 5. Essa matriz permite uma análise detalhada das classificações corretas e incorretas para cada classe (positivo e negativo), proporcionando *insights* sobre o comportamento de cada método em diferentes cenários. A partir dessas comparações, as métricas de acurácia, precisão, revocação e F1-score são calculadas para cada uma das abordagens, permitindo uma avaliação abrangente do desempenho (VILARINHO; RUIZ, 2018).

A acurácia representa a proporção de classificações corretas em relação ao total de casos analisados, sendo uma medida geral de desempenho. A precisão reflete a proporção de verdadeiros positivos em relação ao total de positivos preditos, destacando a capacidade do modelo em evitar falsos positivos. A revocação, por sua vez, indica a proporção de verdadeiros positivos identificados em relação ao total de positivos reais, enfatizando a capacidade do modelo em detectar a classe positiva. O F1-score harmoniza precisão e revocação, oferecendo uma métrica balanceada para avaliar o desempenho quando existe um impasse entre essas duas medidas (VILARINHO; RUIZ, 2018).

O cálculo de cada métrica pode ser formulado pelas Equações (12),(13),(14) e (15) a seguir:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{13}$$

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{14}$$

$$\text{F1-score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{15}$$

Onde os termos TP, TN, FP e FN são os quatro possíveis resultados de uma classificação binária, derivados da matriz de confusão:

- TP (*True Positive* ou Verdadeiro Positivo): O número de instâncias positivas que foram corretamente classificadas como positivas.
- TN (*True Negative* ou Verdadeiro Negativo): O número de instâncias negativas que foram corretamente classificadas como negativas.

- FP (*False Positive* ou Falso Positivo): O número de instâncias negativas que foram incorretamente classificadas como positivas. Também é conhecido como erro do Tipo I.
- FN (*False Negative* ou Falso Negativo): O número de instâncias positivas que foram incorretamente classificadas como negativas. Também é conhecido como erro do Tipo II.

Tabela 5 – Exemplo de matriz de confusão

	Predito Positivo	**Predito Negativo**
Verdadeiro Positivo	50	10
Verdadeiro Negativo	5	35

Neste exemplo da Tabela 5, 50 casos foram corretamente classificados como positivos, representando os Verdadeiros Positivos (TP). Em contrapartida, 10 casos que deveriam ter sido classificados como positivos foram erroneamente classificados como negativos, sendo os Falsos Negativos (FN). Além disso, 5 casos negativos foram incorretamente classificados como positivos, conhecidos como Falsos Positivos (FP). Por fim, 35 casos foram corretamente classificados como negativos, representando os Verdadeiros Negativos (TN).

2.6.1 Avaliação Ponderada

A avaliação de um classificador pode ser sensível ao desbalanceamento de classes no conjunto de dados. Um modelo que simplesmente favorece a classe majoritária pode alcançar uma alta acurácia, mas falhar em identificar corretamente as instâncias da classe minoritária. Para mitigar esse efeito e obter uma medida de performance mais fidedigna, o cálculo de precisão e revocação neste trabalho será realizado de forma ponderada para cada rótulo (positivo e negativo).

Este método garante que a importância de cada classe na métrica final seja proporcional ao seu número de amostras. Assim, obtém-se uma visão mais equilibrada do desempenho do modelo em todo o espectro de dados. O cálculo considera o número de amostras de cada rótulo (N) e a precisão ou revocação (P) associada a esse rótulo. A fórmula utilizada é representada pela Equação (16):

$$\frac{(N_0 * P_0) + (N_1 * P_1)}{N_0 + N_1} \tag{16}$$

Isso significa que, em vez de simplesmente calcular a média da precisão ou revocação para todos os rótulos, cada rótulo contribui para a métrica final de acordo com o número de amostras que possui. Desse modo, rótulos com mais amostras têm maior peso no cálculo final, resultando em uma avaliação mais representativa do desempenho do modelo para o conjunto de dados como um todo.

Capítulo 3

Trabalhos Relacionados

Neste capítulo são discutidos alguns dos principais trabalhos recentes e relacionados com este projeto. Tais trabalhos serviram de base para o desenvolvimento de novas metodologias para a medição de centralidade e análise de sentimentos.

No trabalho de (WATTS, 2007) é discutido como a análise de redes complexas pode ser usada para entender a propagação de informação. O trabalho mostra que as redes complexas podem capturar a estrutura social e os padrões de interação entre os indivíduos. É discutido a respeito dos modelos tradicionais de propagação de informação, que geralmente assumem que a informação se espalha de forma uniforme pela população. O trabalho discorre que esses modelos são inadequados para descrever a propagação de informação em redes complexas, pois as redes complexas são frequentemente caracterizadas por *clusters* e ligações fortes entre indivíduos próximos. Por fim, o trabalho apresenta um modelo de propagação de informação baseado em redes complexas. O modelo assume que a informação se espalha de um indivíduo para outro com uma probabilidade que depende da proximidade e da força da ligação entre os dois indivíduos. É utilizado esse modelo para analisar a propagação de informação em uma variedade de redes, incluindo redes sociais, redes econômicas e redes biológicas. O modelo é capaz de explicar uma variedade de fenômenos observados na propagação de informação, incluindo a disseminação de rumores, a difusão de inovações e a propagação de doenças.

No artigo de (KAPLAN; HAENLEIN, 2010) é abordado de maneira concisa os dados coletados em redes sociais. É abordado os possíveis locais de coleta de dados (posts, comentários e outras formas de conteúdos gerados pelos usuários), o processamento destes dados, incluindo as etapas de filtragem, normalização e codificação. Também aborda a respeito dos tipos de análises a qual os dados podem ser submetidos, tais como análise de tópicos, sentimentos e opnião. Alguns dos desafios citados no artigo incluem a quantidade de dados a ser coletada, a qualidade desses dados e a relevância. Ainda sim, mesmo com todos os desafios mencionados, os autores concluem que os benefícios superam os desafios pois os *insights* coletados possuem extremo valor para tomadas de decisão.

(ADEDOYIN-OLOWE; GABER; STAHL, 2014) discute em sua pesquisa diferentes

técnicas de mineração de dados que foram usadas para minerar diversos aspectos de redes sociais ao longo dos anos. O trabalho apresenta uma nova técnica de mineração de dados, chamada TRCM (*Transaction-based Rule Change Mining*) que é uma metodologia de análise temporal de dados que é usada para identificar tendências, padrões e mudanças nas associações entre itens em um conjunto de dados ao longo do tempo. É uma variação da ARM (*association rule mining*) que leva em consideração a dimensão temporal dos dados. Essa abordagem é baseada na ideia de que as regras que governam as relações entre itens em um conjunto de dados podem mudar ao longo do tempo. Ao comparar as regras que são descobertas em diferentes períodos de tempo, é possível identificar tendências, padrões e mudanças nas relações entre os itens.

O trabalho de (KAUR; MAHAJAN; KAUR, 2016) foca na seleção de características em análise de sentimentos. A técnica apresentada combina SVM com ACO, sendo nomeado de SVM-ACO. A idéia consiste na melhora do desempenho do classificador SVM ajustando as características com base nas rotas encontradas pelas formigas. Esse tipo de abordagem é eficaz pois um vetor de grande dimensão, contendo palavras que não necessáriamente colaboram com aquele contexto, pode diminuir a acurácia do modelo de classificação. Em contra partida, um vetor filtrado apenas com informações relevantes para aquele contexto maximiza a acurácia do modelo.

A abordagem de (GOEL; PRAKASH, 2016) explora o uso de algoritmos de inteligência coletiva, como ACO e Algoritmo Genético (AG), para a análise de sentimentos em comunidades online. O objetivo é melhorar a eficiência e a precisão da análise de sentimentos ao otimizar a seleção de características. Os dados são compostos por postagens de fóruns e comentários de usuários em plataformas online. O trabalho descreve o uso de vetores de feromônio positivos e negativos para rastrear o desenvolvimento do sistema de colônia de formigas. A cada nova palavra desejada, a formiga é inicializada para buscá-la, e o caminho percorrido é verificado com os vetores de feromônio. O vetor que mais se encaixa indica a polaridade da palavra, ajudando a determinar se o sentimento é positivo ou negativo. Os resultados indicam que a aplicação de ACO e GA pode melhorar significativamente a qualidade da análise de sentimentos em comparação com técnicas tradicionais, oferecendo um modelo mais robusto e eficiente.

O artigo de (AGGARWAL; CHHABRA, 2017) explora como a otimização por colônia de formigas (ACO) pode melhorar a análise de sentimentos em tweets. O ACO é usado para selecionar características textuais relevantes, tais como a frequência de palavras, emojis e palavras chaves, atribuindo pesos a essas características para otimizar a seleção das mais relevantes. Essas características otimizadas são então classificadas usando algoritmos de aprendizado supervisionado, como Support Vector Machine (SVM) e Naïve Bayes. O estudo conclui que o uso de ACO para otimização das características melhora a precisão dos modelos de classificação, especialmente quando combinado com SVM, em comparação ao Naïve Bayes.

A abordagem de (VILARINHO; RUIZ, 2018) apresenta um método para análise de sentimentos do Twitter (TSA) que usa métricas de centralidade global para avaliar sentimentos positivos ou negativos expressos em microblogs. A técnica proposta mede a importância de uma sentença para um dado gráfico de sentimento G calculando seu coeficiente SentiElection. SentiElection é um conjunto de três medidas de centralidade global: Índice de *Katz*, *Eigenvector centrality* e *PageRank*. Os resultados são comparados a um modelo anterior baseado em similaridade de contenção e métricas de similaridade baseadas em máximo subgrafo comum.

Posteriormente, o trabalho de (AHMAD et al., 2019) aplicou a técnica de otimização por colônia de formigas (ACO) e k-vizinhos mais próximos (KNN), nomeado ACO-KNN, para resolver o problema de seleção de características em análise de sentimentos. O ACO é usado para encontrar um subconjunto ótimo de características que maximiza a precisão do modelo de classificação enquanto o KNN funciona como um classificador para avaliar os subconjuntos de características candidatas. O foco da abordagem está na geração de um novo *dataset* a partir do *dataset* fornecido, o qual possui a *bag of words* filtradas por caracteristicas, tornando o processo posterior de análise de sentimentos mais preciso.

O artigo de (SALAVATI; ABDOLLAHPOURI, 2019), no universo do marketing viral, propõe novos métodos baseados na otimização por colônia de formigas (ACO) para identificar os nós mais influentes em redes sociais. É proposto dois métodos, o *Profit Maximization Ant Colony Optimization* (PMACO), que foca em maximizar o lucro dos nós selecionados, e o *Influence Maximization with Optimization Ant Colony Optimization* (IMOACO) que além de maximizar o lucro, considera a similaridade entre os nós, reduzindo a redundância e aumentando a taxa de penetração do processo de disseminação. Ambas as abordagens reduzem a complexidade computacional eliminando nós menos influentes e selecionam nós dissimilares para aumentar a cobertura da rede. Os autores concluem que as abordagens são eficientes e reduzem significamente a redundância ao considerar a similaridade de nós mais distantes, cobrindo assim toda a rede.

O trabalho de (CHAKRABORTY; BHATTACHARYYA; BAG, 2020) fornece uma visão geral da evolução da análise de sentimentos discutindo as diferentes técnicas de análise de sentimentos, bem como o processo de captura de dados de mídias sociais. O trabalho também examina as técnicas de comunalização de dados do usuário, que permitem que os pesquisadores combinem dados de diferentes usuários para obter *insights* mais abrangentes. Por fim, o trabalho apresenta uma análise dos dados coletados. Os dados mostram que a análise de sentimentos pode ser usada para entender uma ampla gama de tópicos, incluindo produtos, serviços, eventos e política.

O trabalho de (HAMDI, 2022) também utiliza uma abordagem mista de ACO e SVM, no qual o autor intitula *Affirmative Ant Colony Optimization Based Support Vector Machine* (AACOSVM). A proposta consiste em otimizar os parâmetros da máquina de vetores de suporte (SVM) em duas etapas: transição de estado e atualização de estado. Essa

técnica permite que as formigas utilizem um caminho de feromônio falso para escolher parâmetros, motivando as formigas a criar subconjuntos com o menor número de erros de classificação. O autor afirma que a abordagem supera os classificadores existentes, como o *Entropy-Based Classifier* (EBC) e o *Enhanced Feature Attention Network* (EFAN), em termos de F-Measure e precisão, ao analisar sentimentos em avaliações de produtos da Amazon.

A revisão sistemática de (VEMULA; RATHEE, 2023) apresenta uma análise de três algoritmos bio-inspirados utilizados na análise de sentimentos: *Firefly Algorithm* (FFA), inspirado no comportamento de vaga-lumes, *Cuckoo Search Algorithm* (CSA), baseado no comportamento de parasitas de ninhos de cucos e *Chicken Swarm Optimization Algorithm* (CSOA) que é inspirado na hierarquia social e comportamento de busca de alimento das galinhas. Esses métodos são aplicados para transformar textos em vetores de características e classificar sentimentos em positivo ou negativo. Esse tipo de abordagem ressalta as caracteristicas significativas de um texto, logo, ajudam a melhorar a precisão e eficiência da analise de sentimentos.

Não são muitas as abordagens que utilizam explicitamente ACO para análise de sentimentos. Em geral, foi mais comum observar o ACO como um otimizador de outros algoritmos, como, por exemplo, Naïve Bayes, SVM e KNN, onde a maior contribuição do ACO estava na seleção de características. O trabalho de (GOEL; PRAKASH, 2016) foi o que mais incentivou uma abordagem em que o feromônio fosse de fato utilizado para a classificação, embora através de uma heurística complexa, porém justificável, visto que a rede é atualizada dinamicamente, o que possibilita um novo vetor de feromônio para cada nova palavra adicionada. Os demais trabalhos, em geral, abordam técnicas de análise de sentimentos, mineração de dados e análise de redes complexas que de fato colaboram efetivamente com a construção desse trabalho.

É importante contextualizar esta pesquisa frente aos avanços mais recentes na área, notavelmente a ascensão dos modelos de linguagem massivos (*Large Language Models*) baseados na arquitetura *Transformer*, como o BERT (*Bidirectional Encoder Representations from Transformers*) e os modelos GPT (*Generative Pre-trained Transformer*). Esses modelos alcançaram o estado da arte em diversas tarefas de PLN, incluindo a análise de sentimentos, devido à sua capacidade de capturar nuances contextuais complexas em grandes volumes de texto. No entanto, uma característica proeminente desses modelos é sua natureza de caixa-preta (*black box*), onde os mecanismos de decisão internos são de difícil interpretação. Nesse contexto, a presente dissertação se posiciona de forma complementar: em vez de competir diretamente em performance pura, o método aqui proposto, baseado em feromônios de ACO sobre redes complexas, oferece uma vantagem significativa em interpretabilidade. A análise das trilhas de feromônio permite visualizar e compreender quais caminhos semânticos foram mais relevantes para a classificação, fornecendo uma justificativa transparente para o resultado.

Capítulo 4

Metodologia de Pesquisa

A metodologia deste projeto é composta por quatro etapas principais. A primeira etapa é a mineração de dados, onde os *datasets* serão adquiridos, balanceados, pré-processados e disponibilizados para a criação da rede complexa. A segunda etapa consiste na execução do ACO sobre os bigramas que compõe a rede. A terceira etapa é o pós processamento, são filtrados os vértices que foram percorridos e seus respectivos valores de feromônios. Nesta etapa também são aplicadas algumas técnicas de PLN, como lematização e stematização. Por fim, a última etapa consiste na classificação e avaliação dos resultados, incluindo a comparação com outros métodos propostos e da literatura.

4.1 Mineração de dados

Para esse estudo utilizaram-se de cinco *datasets* pré classificados como positivos e negativos, sendo quatro deles em inglês e apenas um em português, compostos de *tweets* filtrados, obtidos pela API pública do twitter. Não será considerada a polaridade neutra, pois, além de se tratar de uma abordagem controversa, alguns dos *datasets* obtidos não possuiam dados dessa polaridade. A divisão dos dados consiste em 80% de cada conjunto de dados para treino e 20% para a classificação. A seguir, uma breve apresentação de cada *dataset* utilizado:

- US *Airline* 2015[1]: Esse conjunto de dados contém 14.640 *tweets* coletados entre 16 e 24 de fevereiro de 2015 e abrange 6 companhias aéreas dos Estados Unidos. Essas opiniões variam de elogios a reclamações sobre diferentes aspectos das companhias. Inicialmente, os dados foram classificados como positivos, negativos e neutros. Após a exclusão de 3.103 *tweets* neutros, o conjunto final incluiu 9.171 *tweets* negativos e 2.362 positivos.

[1] Disponível em: <https://github.com/ArchismanKarmakar/US-Airlines-2015-dataset-sentiment-analysis> Acesso em 24 de agosto de 2024

- Coachella 2015[2]: Este estudo de análise de sentimentos examinou as reações ao *Coachella Valley Music and Arts Festival* de 2015. Foram coletadas postagens que discutiam os melhores shows e o evento como um todo. A coleção resultante contém 2.362 documentos classificados como positivos e 635 como negativos.

- Deflategate 2015[3]: Antes do Super *Bowl* de 2015, houveram muitas conversas sobre bolas de futebol murchas e se os *Patriots* trapacearam contra os *Colts* mediante a esmagadora vitória de 45 a 7. Este conjunto de dados analisa o sentimento do Twitter em dias importantes durante o escândalo para avaliar o sentimento público sobre todo o ocorrido. A base contém 1.236 postagens de sentimento positivo e 5.033 postagens de sentimento negativos.

- GOP Debate 2015[4]: Foram levantados dezenas de milhares de *tweets* sobre o debate do Partido Republicano em Ohio, no início de agosto de 2015. Colaboradores fizeram uma análise de sentimentos e uma categorização de dados levando em conta se o tweet era relevante, o candidato mencionado, o assunto mencionado e qual era o sentimento de um determinado tweet. Ao remover as mensagens não relevantes do conjunto de dados carregado, o *dataset* resultante possui um total de 2.236 *tweets* classificados como positivo e 8.492 classificados como negativos.

- *Tweets with Theme*[5]: O *Portuguese Tweets for Sentiment Analysis* contém cerca de 800k *tweets* em português disponíveis para uso em análises. *Tweets with themes* é um sub conjunto recolhido utilizando cerca de 100 termos políticos, juntamente com emoticons positivos e negativos. Contém 61.591 *tweets* sendo 32.744 *tweets* classificados como positivo e 28.847 *tweets* classificados como negativo.

4.1.1 Balanceamento de Dados

O balanceamento de dados foi aplicado sobre os 80% dos dados destinados ao desenvolvimento do problema. Deste modo, não houveram descartes na seleção da base de testes ao utilizar o algoritmo de *Random UnderSampler*. A Tabela 6 apresenta a quantidade de *tweets* que serão avaliados para cada *dataset*.

Um número significativo de dados foi descartado pelo *Random UnderSampler* que é uma técnica utilizada para equilibrar conjuntos de dados desbalanceados. Ele faz isso removendo aleatoriamente exemplos da classe que tem mais amostras (a classe majoritária). Ao fazer isso, o algoritmo busca igualar o número de exemplos entre as diferentes classes.

[2] Disponível em: <https://github.com/TatulMesropyan/ntlk-task> Acesso em 24 de agosto de 2024
[3] Disponível em: <https://data.world/crowdflower/deflategate-sentiment> Acesso em 24 de agosto de 2024
[4] Disponível em: <https://www.kaggle.com/datasets/crowdflower/first-gop-debate-twitter-sentiment> Acesso em 24 de agosto de 2024
[5] Disponível em: <https://www.kaggle.com/datasets/augustop/portuguese-tweets-for-sentiment-analysis?select=TweetsWithTheme.csv> Acesso em 24 de agosto de 2024

Para evitar que os mesmos exemplos sejam removidos várias vezes, as amostras são escolhidas sem reposição, ou seja, uma vez removidas, não são consideradas novamente (KABIR, 2024). Especialmente em *US Airlines*, 5393 exemplos foram descartados, indicando que os dados originais eram altamente desequilibrados, com uma alta predominância de uma das classes.

Tabela 6 – Dados coletados

Dataset	Inicial	Classificação (20%)	Treino (80%)	Descartados
Coachella	2997	274	886 (443 Positivo, 443 Negativo)	1837
Deflategate	6269	1255	2014 (1007 Positivo, 1007 Negativo)	3000
GOP Debate	10728	2147	3606 (1803 Positivo, 1803 Negativo)	4975
US Airlines	11533	2308	3832 (1916 Positivo, 1916 Negativo)	5393
Tweets with theme	61591	12318	46018 (23009 Positivo, 23009 Negativo)	3255

4.1.2 Pré-processamento

O pré-processamento consistiu inicialmente na remoção de termos indesejados tais como menções, hashtags, links, caracteres não alfabéticos, emoticons, guids, padrões de risadas e pontuações, que, em geral, não possuem valor semântico para a análise. Embora se reconheça que elementos como emoticons e padrões de risadas possam carregar valor sentimental, optou-se por sua remoção para reduzir a complexidade mantendo o foco na análise das relações textuais entre as palavras.

Em seguida, devido à distinção de linguagem presente nos *datasets*, duas abordagens distintas e com propósito similar são elencadas. Primeiro, para os *datasets* na língua inglesa, a expansão de contrações. Contrações são comuns na linguagem, a redução é muito útil para normalizar o texto antes de gerar vetores de palavras ou caracteres. A redução de contrações é efetuada através de regras de substituição simples das contrações inglesas normalmente utilizadas. A Tabela 7 demonstra alguns exemplos de contrações comuns da lingua inglesa. Segundo, para o *dataset* em português, a normalização das palavras. A normalização consiste em expandir abreviações e corrigir eventuais erros gramaticais. A Tabela 8 demonstra alguns exemplos de normalização.

Na sequência, o texto resultante é tokenizado e submetido a uma filtragem de *stopwords*. O objetivo principal desta etapa é a obtenção de *tokens* válidos e úteis para a análise de sentimentos. Diz-se úteis, aquelas palavras que dão sentido ao texto. Uma vez reduzido o texto a *tokens* significativos e bigramas desses *tokens*, um *corpus* com a frequência de cada palavra no texto é construido. Esse passo é importante pois a partir disso cada bigrama terá um peso relacionado a sua intensidade de uso, o que será útil nas próximas etapas. A Figura 5 elucida todo o processo de mineração descrito acima.

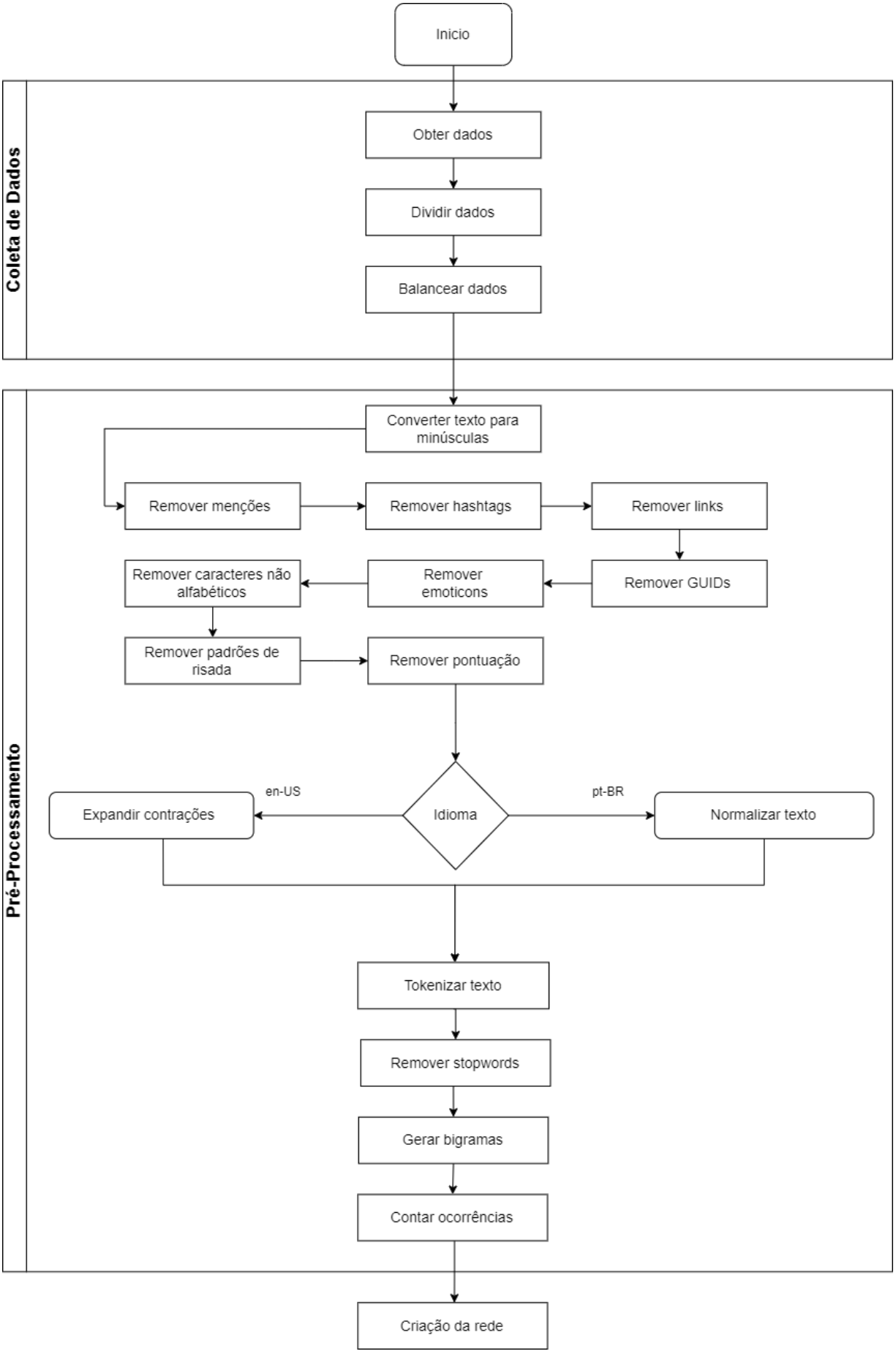

Figura 5 – Fluxograma da mineração de dados.

Tabela 7 – Contrações comuns em inglês e suas formas expandidas

Contração	**Forma expandida**
I'm	I am
You're	You are
He's	He is
She's	She is
It's	It is
We're	We are
They're	They are
Can't	Cannot
Won't	Will not
Didn't	Did not

Tabela 8 – Abreviações e erros gramaticais comuns em português e suas expansões/correções

Abreviação/Erro Gramatical	**Expansão/Correção**
vc	você
pq	porque
n	não
tb	também
q	que
ñ	não
kd	cadê
blz	beleza
mto	muito
ta	está

4.2 Rede complexa e ACO

A partir dos *tokens* gerados na fase de pré-processamento, nesta etapa espera-se a criação de uma rede complexa para cada polaridade desejada. Os relacionamentos serão determinados pelos bigramas obtidos dos dados processados na fase anterior. Cada palavra única se torna um vértice da rede. Uma aresta é criada entre dois nós, u e v, se a sequência de palavras uv constitui um bigrama identificado no texto. Portanto, a rede modela as relações de co-ocorrência adjacente entre as palavras.

Para a análise, são geradas duas redes complexas independentes: uma rede positiva e uma rede negativa. A rede positiva é construída utilizando exclusivamente os bigramas extraídos dos textos classificados como positivos no conjunto de treinamento. De forma análoga, a rede negativa é construída apenas com os bigramas de textos negativos. Essa separação permite que o método calcule a importância de cada bigrama dentro de seu contexto sentimental específico.

As arestas entre os vértices possuirão um peso, determinado pela frequência do bigrama no *dataset*. Os pesos nas arestas representam a atratividade de uma determinada aresta. As formigas, nos ciclos iniciais, ao escolherem a próxima aresta a seguir, são mais

propensas a escolher arestas com pesos maiores.

Deste modo temos um cenário ideal para a execução do ACO, que tem por objetivo não uma solução, como encontrar um determinado vertice, mas o que chamaremos de reconhecimento de área, onde as formigas apenas andarão pelos vértices distribuindo o feromônio de forma semi-determinística, o qual será coletado posteriormente.

O algoritmo implementa uma simulação de colônia de formigas para explorar e otimizar caminhos em uma rede complexa, utilizando o conceito de feromônios para guiar as formigas. Cada formiga começa em um vértice inicial e se move pelo grafo, escolhendo seu próximo passo com base em uma probabilidade que considera a quantidade de feromônio e a atratividade entre os vértices. À medida que as formigas percorrem o grafo, elas depositam feromônios nas arestas que atravessam, reforçando os caminhos mais promissores. No entanto, a energia da formiga é limitada, e ela pode "morrer" se sua energia se esgotar antes de visitar todos os vértices disponíveis. De igual modo, formigas também são recompensadas com energia se escolhem bons caminhos. Após cada iteração, a quantidade de feromônio nas arestas é reduzida por um processo de evaporação, o que evita que soluções antigas se mantenham dominantes por muito tempo e permite que o algoritmo explore novos caminhos ao longo do tempo. O processo é repetido por várias iterações, e o estado final dos feromônios é salvo para análise posterior. O algoritmo utiliza paralelismo através de *threads* para simular o movimento de várias formigas simultaneamente, o que aumenta a eficiência da simulação. A Figura 6 descreve o fluxo mencionado acima e será abordado pelos tópicos seguintes.

4.2.1 Seleção do vértice inicial

Todas as arestas são criadas com o valor de atratividade, provenientes da frequência dos bigramas, além disso, um valor inicial de feromônio será depositado em cada uma das arestas. A seleção do vértice inicial de cada formiga é baseado em certos critérios e distribuições de probabilidade. Inicialmente o algoritmo separa um vetor contendo todos os vértices com o valor de feromônio inicial intacto. Obviamente, no primeiro ciclo, esse vetor representa toda a rede, no entanto a medida que a rede é percorrida, o tamanho desse vetor tende a diminuir.

Uma distribuição Bernoulli é utilizada para decidir se a seleção será feita entre os vértices não percorridos. Essa distribuição tem uma probabilidade de 40%, significando que, em 40% das vezes, o método tentará selecionar um vértice do vetor de vértices não percorridos. Esse limiar foi escolhido empiricamente, com base em testes que indicaram sua eficácia em situações onde não há uma escolha dominante. O valor de 40% permite que alternativas menos favorecidas, mas ainda viáveis, sejam consideradas com a devida frequência. Assim, opções com 40% de preferência são levadas em conta, mesmo que não sejam as mais favorecidas. Se a distribuição indicar verdadeiro (dentro dos 40%) e houver vértices no vetor, um vértice é selecionado aleatoriamente desse vetor usando uma

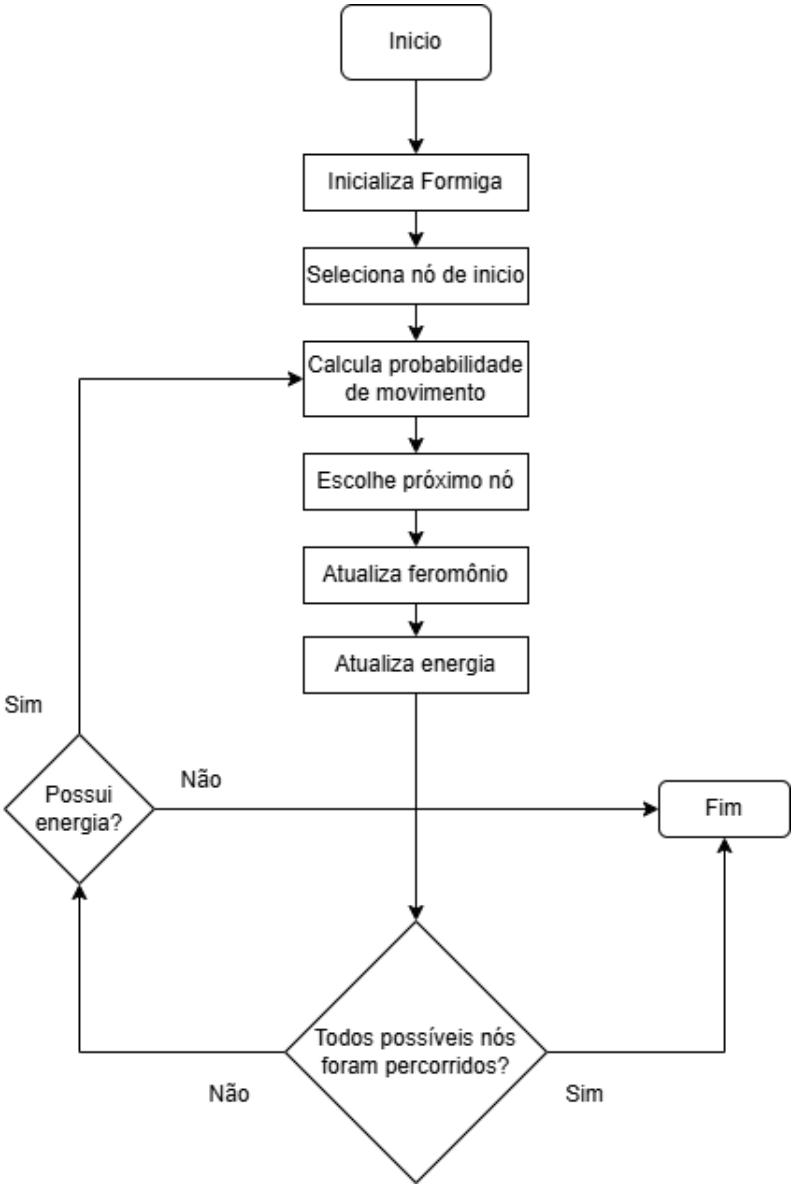

Figura 6 – Fluxograma do ACO proposto.

distribuição uniforme. Caso contrário (ou seja, se a distribuição Bernoulli for falsa ou se o vetor estiver vazio), o método seleciona um vértice aleatório dentre todos os vértices do grafo, também utilizando uma distribuição uniforme.

Seja $G = (V, E)$ um grafo, onde V é o conjunto de vértices e E é o conjunto de arestas. O valor de feromônio em uma aresta $e \in E$ é representado por $\tau(e)$. O conjunto de vértices com feromônio inicial é dado pela Equação (17):

$$V_{\text{não percorridos}} = \{v \in V : \exists e \in E, \tau(e) = \tau_{\text{inicial}}\} \tag{17}$$

A seleção de um vértice v pode ser obtida uniformemente pelas Equações (18) e (19) onde, X é a variável aleatória que representa o resultado do experimento de Bernoulli e pode assumir 0 ou 1:

- $X = 1$ indica que a escolha será feita entre os vértices não percorridos, com probabilidade $P(X = 1) = 0.4$:

$$P(v = v_i \mid v_i \in V_{\text{não percorridos}}) = \frac{1}{|V_{\text{não percorridos}}|} \tag{18}$$

- $X = 0$ indica que a escolha será feita entre todos os vértices do grafo, com probabilidade $P(X = 0) = 0.6$:

$$P(v = v_i \mid v_i \in V) = \frac{1}{|V|} \tag{19}$$

O objetivo é obter uma abordagem probabilística para determinar a escolha de um vértice, balanceando a exploração de vértices ainda não percorridos com a escolha aleatória de qualquer vértice da rede.

4.2.2 Probabilidade de movimento

A probabilidade $p(x, y)$ de uma formiga se mover de um vértice x para um vértice vizinho y é diretamente proporcional tanto à quantidade de feromônio $\tau(x, y)$ quanto à atratividade $d(x, y)$ do vértice vizinho y. Isso significa que caminhos com maior feromônio e maior atratividade são mais prováveis de serem escolhidos. A probabilidade é definida pela Equação (20):

$$p(x, y) = \tau(x, y) \times d(x, y) \tag{20}$$

Onde $\tau(x, y)$ é a quantidade de feromônio na aresta que conecta os vértices x e y e $d(x, y)$ é a atratividade do vértice y em relação ao vértice x.

Diferentemente do caso onde $d(x, y)$ representaria uma distância (o que implicaria em uma relação inversamente proporcional), aqui, quanto maior a atratividade $d(x, y)$, maior a probabilidade de escolha desse caminho.

Além disso, outra abordagem é utilizada, chamaremos de probabilidade acumulada. A probabilidade acumulada é utilizada durante o processo de seleção do próximo vértice. Esta abordagem envolve acumular as probabilidades dos vizinhos até que a soma acumulada seja maior ou igual a um valor aleatório R dentro da somatória total da atratividade dos nós vizinhos, chamado limiar, a qual é expressa pela Equação (21).

$$P_{\text{acum}} = \sum_{y' \in \text{vizinhos}} p(x, y') \tag{21}$$

A formiga começa em um vértice e verifica os vizinhos não visitados. Para cada vizinho, calcula-se a probabilidade de movimentação com base na quantidade de feromônio e atratividade da aresta. Essas probabilidades são acumuladas, e o processo entra em um *loop* onde a soma acumulada é comparada a um limiar aleatório. Quando a soma

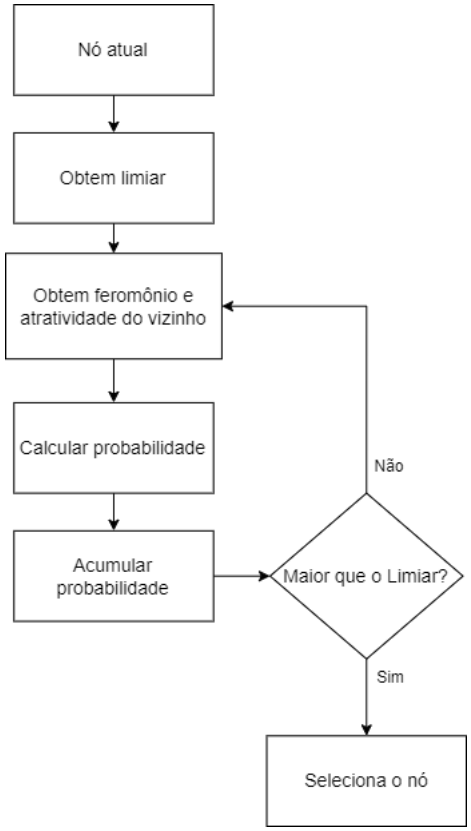

Figura 7 – Probabilidade de movimento.

ultrapassa o limiar, o vizinho correspondente é selecionado como o próximo vértice a ser visitado, conforme ilustrado na Figura 7. Esse *loop* garante uma seleção ponderada, priorizando vértices mais atrativos, mas ainda permitindo diversidade na exploração.

Essa forma de cálculo de probabilidade acumulada permite que a seleção do próximo vértice seja influenciada pelas probabilidades relativas, mas ainda assim mantida como um processo estocástico. Isso significa que vértices com maior feromônio e atratividade têm uma chance maior de serem escolhidos, visto que por possuirem um valor elevado, a sua soma tem maior chance de ultrapassar o limiar, mas todos os vizinhos têm alguma chance a depender do valor aleatório R.

Um aspecto crucial dessa simulação é garantir que as formigas não revisitem vértices já percorridos durante uma única iteração. Isso é gerenciado por meio de um registro dos vértices que a formiga já visitou. Ao considerar os vizinhos do vértice atual, a formiga verifica se um vértice já está presente nesse conjunto antes de decidir o próximo passo. Se o vértice já foi visitado, ele é ignorado, e a formiga avança para o próximo vizinho disponível. Dessa forma, cada formiga explora o grafo sem retornar a um vértice previamente visitado, maximizando a diversidade das soluções encontradas e evitando ciclos desnecessários durante a simulação.

4.2.3 Atualização de feromônio

A taxa de evaporação ρ é definida em um intervalo que varia entre um valor mínimo e máximo, permitindo que a evaporação seja aleatória dentro desses limites. Isso introduz variabilidade no processo de evaporação, simulando um ambiente dinâmico onde a volatilidade do feromônio pode variar.

A Equação (22) descreve o processo de evaporação do feromônio ao longo do tempo. A quantidade de feromônio $\tau(n, v)$ em uma aresta é multiplicada por uma taxa de evaporação ρ, que é um valor aleatório entre um mínimo $\rho_{\min}$ e um máximo $\rho_{\max}$. Este processo evita que as soluções anteriores dominem a busca indefinidamente, permitindo que o algoritmo explore novas rotas e possibilite a adaptação a novas condições durante a execução.

$$\tau'(n, v) = \tau(n, v) \times \rho \quad \text{onde} \quad \rho \sim \text{Uniform}(\rho_{\min}, \rho_{\max}) \tag{22}$$

Onde $\tau'(n, v)$ é o valor do feromônio após a evaporação e o símbolo $\sim$ indica que ρ é uma variável aleatória extraída de uma distribuição uniforme.

4.2.4 Atualização da energia

A energia da formiga é um conceito elaborado para essa proposta e é ajustada com base na quantidade de feromônio presente na aresta que conecta o vértice atual x ao próximo vértice y. Se a quantidade de feromônio $\tau(x, y)$ for maior que um valor inicial τ_{inicial}, a formiga ganha energia ao percorrer essa aresta. Caso contrário, a formiga perde energia ao percorrer uma aresta com pouco ou nenhum feromônio. É descrito pela Equação, (23)

$$E(t + 1) = E(t) + \Delta E_{\text{ganha}} - \Delta E_{\text{perdida}} \tag{23}$$

Onde $E(t + 1)$ é a energia no próximo instante de tempo, $E(t)$ é a energia no instante atual, ΔE_{ganha} é a quantidade de energia ganha e $\Delta E_{\text{perdida}}$ é a quantidade de energia perdida.

O conceito de energia serve como uma métrica para simular a capacidade de exploração das formigas. A energia representa a "vitalidade" da formiga, diminuindo à medida que ela percorre arestas com pouco feromônio (indicando caminhos menos explorados ou menos promissores) e aumentando ao passar por arestas com maior concentração de feromônio (indicando caminhos mais explorados e possivelmente melhores). Isso cria uma dinâmica onde as formigas com mais energia têm maior capacidade de explorar o grafo por mais tempo, enquanto aquelas com menos energia são forçadas a encerrar sua busca mais cedo. Essa abordagem permite um equilíbrio entre exploração e exploração intensiva de áreas promissoras.

A energia atua como uma função de avaliação para medir o desempenho da formiga ao explorar a rede. A energia não apenas limita a duração da exploração, mas também serve como um indicador da qualidade das decisões da formiga, influenciando diretamente sua

capacidade de continuar contribuindo para a construção da solução global. Desta forma, ao "morrer" ou chegar a um vértice "sem saida", ou seja, sem vizinhos disponíveis, a formiga libera o fluxo para a produção de uma nova formiga que repetirá o fluxo proposto pelo algoritmo.

4.3 Pós-processamento

Após o término da execução do ACO, um novo *dataset* é gerado contendo os bigramas e a quantidade de feromônio depositada. Nesta etapa, bigramas que não tiveram suas arestas afetadas por feromônios são excluídos, restando apenas os bigramas que de fato são considerados caminhos válidos, ou seja, que foram, de algum modo, percorridos pelas formigas.

O *dataset* contém uma coleção considerável de palavras, no entanto, dificilmente abrangeria toda a sintaxe de uma linguagem. O objetivo é permitir que o algoritmo ACO opere sobre a rede de palavras em sua forma original, capturando as relações de coocorrência entre os termos exatos utilizados nos textos. Ao fazer isso, a análise de centralidade baseada em feromônios pode identificar a importância de variações morfológicas específicas.

Por causa disso, como demonstrado na Figura 7, serão aplicadas técnicas de lematização e stematização sobre as palavras presentes nos bigramas. A lematização irá gerar equivalência para as palavras, pois, ao reduzir cada palavra ao seu lema, existe a possibilidade de que algumas palavras se tornem iguais, somando assim o peso de suas caracteristicas. A stematização, por sua vez, pode ajudar a identificar as raízes das palavras, unificando ainda mais os termos, garantindo que variações morfológicas não interfiram na análise subsequente dos bigramas. A aplicação posterior de lematização e stematização serve para agregar os valores de feromônio de diferentes formas de uma mesma palavra-raiz, consolidando a importância semântica do conceito após sua relevância ter sido estabelecida pelo ACO.

A aplicação das técnicas de lematização e stematização é fundamental para aprimorar a qualidade e a eficácia do *dataset* gerado após a execução do ACO. Ao unificar palavras que possuem variações morfológicas, essas técnicas reduzem a dimensionalidade do conjunto de dados, eliminando redundâncias que poderiam prejudicar a performance do algoritmos de classificação. Na análise posterior esse ponto terá relevância porque diferentes formas de uma palavra podem ter significados parecidos ou iguais. Além disso, ao consolidar termos relacionados, a lematização e a stematização permitem que o classificador trabalhe com um conjunto mais robusto de bigramas, aumentando a cobertura lexical do *dataset* e, consequentemente, a precisão do modelo na detecção de padrões semânticos relevantes. Essa abordagem garante que o modelo resultante seja mais generalizável e menos suscetível a erros causados por variações linguísticas superficiais.

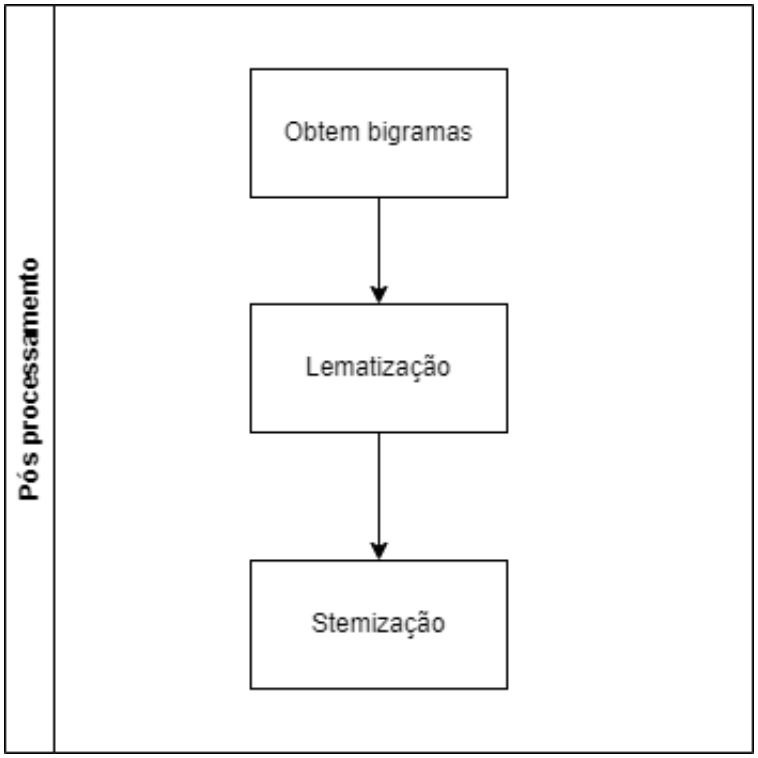

Figura 8 – Pós processamento.

4.4 Classificação e Avaliação

A execução do processo de classificação é exemplificada pela Figura 9 e consiste em aplicar o conjunto de teste extraido pela divisão de dados ao mesmo *pipeline* de pré-processamento (excluindo a contagem de ocorrências) e pós-processamento descrito anteriormente. Ao final, teremos uma lista de bigramas stematizados para cada frase presente no *dataset*.

Propomos a classificação das frases ao analisar os bigramas que aparecem nelas e somar as medidas de feromônio associadas a esses bigramas presentes nos *datasets* positivos e negativos que foram classificados pelo ACO. Para determinar se uma frase tem uma tendência positiva ou negativa, o sistema compara as somas dos feromônios dos bigramas presentes na frase. Esse processo é descrito pela Equação 24:

$$S_{\text{final}} = \sum_{i=1}^{n_{\text{pos}}} P(B_i) - \sum_{j=1}^{n_{\text{neg}}} N(B_j) \tag{24}$$

Onde S_{final} é o score final da frase. $P(B_i)$ é o valor de feromônio associado ao bigrama B_i presente no conjunto de bigramas positivos. $N(B_j)$ é o valor de feromônio associado ao bigrama B_j presente no conjunto de bigramas negativos. n_{pos} é o número de bigramas positivos na frase. n_{neg} é o número de bigramas negativos na frase.

Se $S_{\text{final}} > 0$, a frase é classificada como positiva; caso contrário, é classificada como negativa:

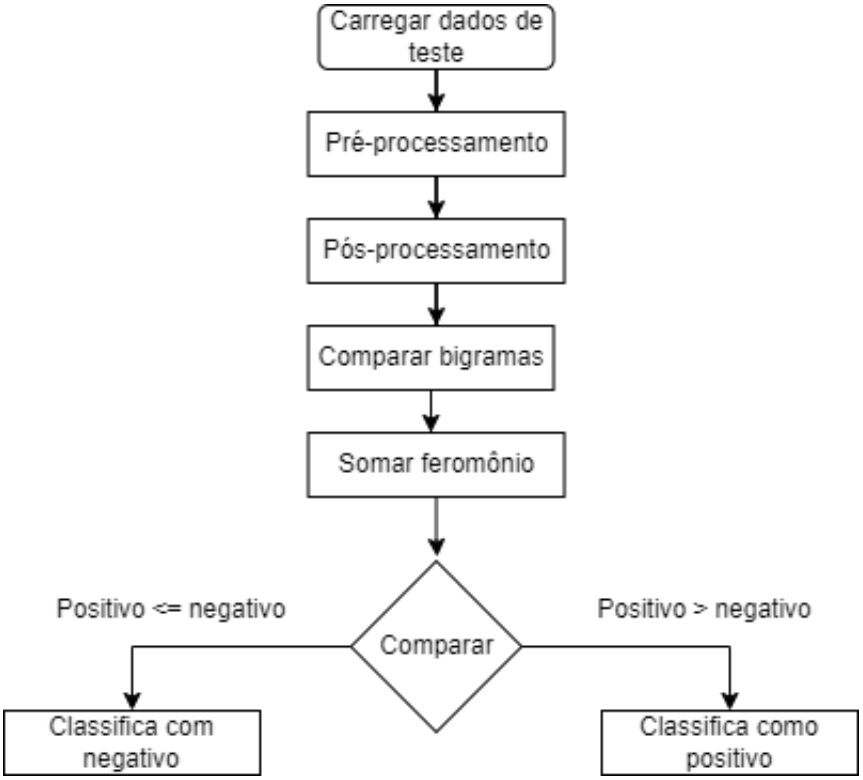

Figura 9 – Fluxograma de classificação.

$$\text{Classificação} = \begin{cases} \text{Positiva} & \text{se } S_{\text{final}} > 0 \\ \text{Negativa} & \text{se } S_{\text{final}} \leq 0 \end{cases}$$

Durante a classificação de um novo texto, é comum que um bigrama esteja presente tanto na rede positiva quanto na negativa. O desempate não se baseia na mera presença, mas no valor da medida de centralidade (feromônio, PageRank, etc.) associado a ele em cada rede. A Equação 24 lida naturalmente com essa situação: o score final é a soma dos valores de centralidade da rede positiva subtraída da soma dos valores da rede negativa. Assim, um bigrama contribui para a polaridade da rede onde ele (e seus vizinhos, no caso de medidas globais) é considerado mais importante.

Esse processo permite que o sistema avalie o sentimento geral da frase com base nas combinações de palavras que ela contém. Dessa forma, o sistema se baseia em uma análise estatística dos bigramas, identificando aqueles que têm maior influência em determinar o sentimento de uma frase.

Para validar e comparar a abordagem proposta, o mesmo processo de classificação será aplicado utilizando quatro outras métricas estabelecidas na literatura: Degree Centrality, Eigenvector, PageRank e Katz. A metodologia segue uma estrutura comum: são construídas duas redes de co-ocorrência de palavras, uma a partir dos textos de sentimento positivo (G_{pos}) e outra dos textos de sentimento negativo (G_{neg}). Em seguida, a centralidade de cada palavra (vértice) é calculada em ambas as redes.

A classificação de sentimentos em cada abordagem se baseia em um valor fundamental que quantifica a importância de palavras ou de suas sequências. Embora o objetivo seja sempre avaliar os bigramas de uma frase, os métodos divergem sobre onde esse valor é atribuído: alguns o associam diretamente ao bigrama (a relação entre duas palavras), enquanto outros o calculam para as palavras individuais (as entidades). As abordagens de ACO e de Katz operam no nível do bigrama, ao passo que as métricas de Degree Centrality, Eigenvector e PageRank calculam um score para cada palavra individualmente.

No ACO, o valor fundamental é o nível de feromônio, que é atribuído diretamente à aresta do grafo que representa um bigrama. Esse valor emerge de um processo simulado onde formigas virtuais exploram a rede de palavras, depositando feromônio nos caminhos (bigramas) que percorrem. As rotas semanticamente importantes, mais frequentadas pelas formigas, acumulam uma maior quantidade de feromônio ao longo do tempo, enquanto um mecanismo de evaporação diminui a importância de caminhos menos relevantes. O resultado é um valor que representa a força da conexão contextual entre duas palavras para expressar um determinado sentimento.

Em contraste, as métricas de Degree Centrality, Eigenvector e PageRank calculam um score de centralidade para cada palavra (vértice) da rede. O Degree Centrality mede a conectividade local de uma palavra, baseando-se no número de suas ligações diretas. O Eigenvector, por sua vez, mede a influência global, atribuindo um score maior a palavras conectadas a outras que também são influentes. De forma similar, o PageRank calcula a importância de uma palavra com base na estrutura de links da rede, identificando termos que são destinos frequentes de caminhos semânticos relevantes. Para classificar uma frase, o sistema primeiro identifica seus bigramas e, em seguida, soma os scores de centralidade de cada palavra constituinte, utilizando os valores dos modelos positivo e negativo para gerar um score agregado para a frase inteira.

A abordagem de Katz adota um modelo probabilístico distinto dos demais. Aqui, o valor associado ao bigrama é sua probabilidade condicional, que indica a chance de a segunda palavra aparecer após a primeira. Esse valor é derivado das frequências relativas no corpus de treinamento. A característica central do método é o mecanismo de Katz Backoff, que lida com a escassez de dados. Se um bigrama nunca foi visto, o modelo não lhe atribui probabilidade zero, mas recua para uma versão descontada da probabilidade do unigrama, garantindo que qualquer sequência de palavras possa ser avaliada. A classificação de uma frase é feita calculando-se a probabilidade total da sequência (o produto das probabilidades de seus bigramas) em relação aos modelos positivo e negativo, e atribuindo o sentimento do modelo que gerar a maior probabilidade.

Após a classificação das frases como positivas ou negativas, essas predições são comparadas com as classificações verdadeiras associadas a cada frase no conjunto de dados de teste. Essa comparação permite identificar onde o modelo acertou e onde errou, proporcionando uma base para calcular diversas métricas de desempenho. A matriz de confusão

oferece uma representação visual e numérica das predições feitas pelo modelo em comparação com os valores reais, organizando os resultados em quatro categorias: verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN).

As principais métricas utilizadas para avaliar o desempenho do modelo incluem acurácia, precisão, recall e F1-score que são extraidas da matriz de confusão gerada pela comparação da classificação e os dados reais. A precisão e revocação são aplicadas utilizando uma formula ponderada (Equação (16)). Essa abordagem é utilizada pois o conjunto de teste, ao refletir a distribuição real dos dados, é desbalanceado. O cálculo da média ponderada para a precisão e a revocação garante que o desempenho em ambas as classes (majoritária e minoritária) seja representado de forma proporcional na métrica final, fornecendo uma avaliação mais justa e robusta do desempenho geral do modelo.

A fim de comparar e analisar o desempenho da metodologia proposta de forma robusta, a avaliação foi dividida em duas frentes. Inicialmente, cada medida de centralidade foi aplicada individualmente. Posteriormente, foram compostos dez diferentes conjuntos de classificação, cada um formado por uma combinação única de três das métricas avaliadas.

A classificação final para estes conjuntos foi definida através de um mecanismo de votação majoritária. Para uma determinada frase, cada uma das três medidas no conjunto realiza sua própria classificação, resultando em três votos independentes para positivo ou negativo. A classificação final da frase é então atribuída à polaridade que obtiver a maioria dos votos. Por exemplo, se em uma combinação o PageRank e o ACO classificarem uma frase como positiva e o Katz a classificar como negativa, o resultado final da combinação para aquela frase será positivo, pois esta polaridade obteve dois dos três votos disponíveis.

Capítulo 5

Experimentos e Análise dos Resultados

Neste capítulo são apresentados os experimentos desenvolvidos e suas análises nos quais o método proposto foi aplicado. Os experimentos foram executados em um computador pessoal com processador Intel® Core™ i9-13900H de 2,60 GHz, com 32 GB de RAM e sistema operacional Windows 11. Também foi utilizado o Python 3.12.2 com as bibliotecas (Nltk 3.8.1, Pandas 2.2.2, Enelvo 0.15, Imblearn 0.0, Spacy 3.7.2, Sklearn 1.5.0 e Matplotlib 3.8.3), C++ versão 14, Visual Studio 2022 versão 17.8.3 e Visual Studio Code 1.92.2 como ambiente de desenvolvimento e testes.

Os experimentos foram conduzidos em cinco *datasets*, conforme detalhado na Seção 4.1, aplicando-se a metodologia proposta no Capítulo 4. Após o processamento, a classificação de sentimentos resultante foi avaliada utilizando uma matriz de confusão.

5.1 Experimentos

Cada etapa de pré-processamento, desde a conversão para minúsculas até a remoção de pontuação, contribuiu para a normalização dos textos. Ainda assim, existem *tokens* que não possuem um significado real para a análise. Esses *tokens* surgem decorrente de erros gramaticais, abreviações ou apelidos desconhecidos que não foram completamente eliminados durante o processo e esses elementos residuais introduzem ruído nos dados. A utilização de bibliotecas em *Python* como Enelvo[1], Nltk[2] e Contractions[3], minimizaram esse ruído a um patamar aceitável para a análise posterior. O idioma não representou um empecilho significativo para o processo. Tanto os *datasets* em inglês quanto em português foram tratados de maneira eficiente, graças às ferramentas específicas, presentes nas bibliotecas Nltk e Enelvo, utilizadas para cada idioma. A expansão de contrações em inglês e a normalização de texto em português garantiram que as particularidades de cada língua fossem respeitadas, resultando em uma boa preparação dos dados. O pré e pós-

[1] Disponível em: <https://thalesbertaglia.com/enelvo/>Acesso em 10 de setembro de 2024
[2] Disponível em: <https://www.nltk.org/>Acesso em 10 de setembro de 2024
[3] Disponível em: <https://pypi.org/project/pycontractions/>Acesso em 10 de setembro de 2024

processamento dos dados textuais demonstrou ser eficaz na preparação dos datasets para análise. A tokenização, seguida pela remoção de *stopwords*, refinou ainda mais os dados, tornando-os prontos para a análise de sentimentos. Embora alguns *tokens* ruidosos ainda possam estar presentes, o processo geral conseguiu lidar bem com as complexidades dos textos em ambos os idiomas, proporcionando uma base sólida para a análise subsequente.

Os dados resultantes da etapa de pós-processamento são demonstrados na Tabela 9 e 10 resumindo a quantidade de bigramas negativos e positivos processados. A Tabela 9 exclui a execução do ACO, e os dados são pós-processados logo ao término do pré-processamento. A este primeiro conjunto de bigramas será aplicado o cálculo de centralidade Katz, EigenVector, DegreeCentrality e PageRank. Na Tabela 10 , após a execução do ACO, os bigramas que não receberam feromônio são removidos, resultando em uma quantidade menor de bigramas processados.

Tabela 9 – Número de bigramas processados por Katz, EigenVector, DegreeCentrality e PageRank

Dataset	**Negativos**	**Positivos**
Coachella	1837	1703
Deflategate	4653	5999
GOPDebate	8730	7235
Tweets With Theme	151508	188398
US Airlines	13648	9379

Tabela 10 – Número de bigramas visitados e bigramas não visitados pelo ACO

	Visitados		**Não Visitados**	
Dataset	**Negativos**	**Positivos**	**Negativos**	**Positivos**
Coachella	1823	1680	14	23
Deflategate	4073	5202	580	797
GOPDebate	7502	6058	1228	1177
Tweets With Theme	106777	133027	44731	55371
US Airlines	11441	7850	2207	1529

Pode-se observar uma redução significativa no número de bigramas processados pelo ACO em comparação com os métodos tradicionais, especialmente nos *datasets Tweets With Theme* e *US Airlines*, onde milhares de bigramas não foram visitados. Essa diferença é particularmente importante, pois indica que o ACO é mais seletivo, concentrando-se em bigramas com maior potencial de relevância para a classificação de sentimentos.

5.1.1 Classificação por ACO

No algoritmo, vários parâmetros desempenham papéis essenciais na simulação do movimento das formigas. A taxa de evaporação é definida como 0,25 determinando a quantidade de feromônio que se dissipa em cada iteração, influenciando a persistência das trilhas.

O feromônio inicial é 0,1 atribuído a todas as arestas ao serem criadas, representando a quantidade mínima de feromônio. O valor do feromônio depositado é 10, indicando a quantidade de feromônio que uma formiga adiciona à trilha ao percorrer uma aresta. A energia inicial das formigas é 100, controlando a distância que uma formiga pode percorrer antes de morrer. A formiga ganha 0,3 de energia ao percorrer uma aresta com feromônio e perde 10 ao percorrer uma sem feromônio. No final, o algoritmo simula 1500 formigas em 50 iterações para explorar a rede, ajustando dinamicamente esses parâmetros para otimizar a busca de soluções. A definição de cada valor é justificada a seguir:

- Taxa de Evaporação (0,25): Uma taxa de evaporação de 25% significa que o feromônio diminui relativamente rápido a cada iteração. Isso sugere que as trilhas antigas perdem sua influência rapidamente, incentivando as formigas a explorar novas trilhas. Esse valor pode promover uma exploração mais diversificada do espaço de soluções, evitando que o algoritmo convirja prematuramente para um caminho subótimo.

- Feromônio Inicial (0,1): O valor baixo do feromônio inicial garante que todas as arestas comecem com pouca atratividade, o que faz com que a influência inicial seja mínima. Isso permite que as trilhas mais promissoras sejam construídas com base no comportamento das formigas durante a simulação, e não com base em valores predefinidos.

- Feromônio Depositado (10): A quantidade significativa de feromônio depositada em cada iteração (10 vezes maior que o feromônio inicial) sugere que, uma vez que uma formiga encontra uma trilha promissora, essa trilha rapidamente se torna mais atrativa para outras formigas. Isso acelera a escolha dos melhores caminhos. Um valor maior poderia levar a uma exploração insuficiente de outras soluções potenciais.

- Energia da Formiga (100): A energia inicial das formigas permite que elas explorem uma parte significativa do grafo antes de morrerem. Isso é importante para assegurar que as formigas possam percorrer várias arestas. No entanto, a energia ganha (0,3) e a energia perdida (10) ao percorrer arestas com e sem feromônio, respectivamente, indicam que percorrer trilhas sem feromônio é altamente penalizador. Esse balanço entre energia inicial e energia perdida favorece a exploração de trilhas ricas em feromônio. A probabilidade acumulada é um fator decisivo desta etapa, pois frequentemente obriga as formigas a explorarem caminhos subótimos.

- Número de Formigas e Iterações (1500 formigas, 50 iterações): O uso de 1500 formigas ao longo de 50 iterações reflete um esforço substancial de exploração e exploração múltipla das possíveis soluções. Esses valores indicam que o algoritmo foi projetado para ter uma robustez significativa, permitindo a cada formiga contribuir para a

busca de soluções, o que pode garantir que as formigas conseguirão percorrer a maior parte, senão toda a rede.

O resultado da classificação de cada *dataset* é apresentado na Tabela 11. Entre os *datasets* analisados, o *US Airlines* obteve a melhor acurácia, com um valor de 0,77. Essa alta acurácia sugere que o modelo foi mais eficiente em prever corretamente as instâncias para este conjunto de dados específico. Em contraste, o *Coachella* teve a menor acurácia, com 0,54, indicando uma performance relativamente inferior na classificação das instâncias desse *dataset*. A análise das outras métricas como precisão, revocação e F1-score também revela variações significativas entre os *datasets*, refletindo diferentes desafios e características de cada um.

Além da acurácia, as métricas de precisão e revocação também mostram variações consideráveis. Por exemplo, o *Tweets With Theme* apresentou um F1-score de 0,70, o que é relativamente alto e demonstra um bom equilíbrio entre precisão e revocação. O desempenho foi consistente entre as bases, com o F1-score variando de 0,59 no dataset Coachella até 0,79 no US Airlines. O Deflategate também apresentou um bom equilíbrio entre precisão e revocação, alcançando um F1-score de 0,72. Isso demonstra que a abordagem manteve uma performance robusta em diferentes contextos.

Tabela 11 – Desempenho do ACO em diferentes bases de dados (F1-score corrigido)

Base	**Acurácia**	**Precisão**	**Revocação**	**F1-score**
US Airlines	0,77	0,81	0,77	0,79
Coachella	0,54	0,64	0,54	0,59
Deflategate	0,69	0,75	0,69	0,72
Tweets With Theme	0,70	0,71	0,70	0,70
GOPDebate	0,66	0,74	0,66	0,70

5.1.2 Classificação usando outras medidas de centralidade

Para obter parâmetros de comparação e mensurar o desempenho da classificação utilizando os feromônios, os *datasets* foram submetidos à classificação após os bigramas serem processados pelos algoritmos Katz, EigenVector, Degree Centrality e PageRank. Esses métodos foram aplicados para avaliar como diferentes medidas de centralidade influenciam os resultados de classificação em termos de acurácia, precisão, revocação e F1-score, além de examinar as respectivas matrizes de confusão. Os resultados obtidos para cada *dataset* é ilustrado pela Figura 10 e descrito em detalhes na Tabela 12.

Analisando os resultados da Tabela 12, observa-se que o método PageRank obteve a melhor performance geral no *dataset* US Airlines, alcançando um F1-score de 0,83. Para a base Deflategate, o Katz se destacou com um F1-score de 0,76, enquanto para a base GOPDebate, o Katz também liderou com 0,77. Isso sugere que, embora o PageRank seja

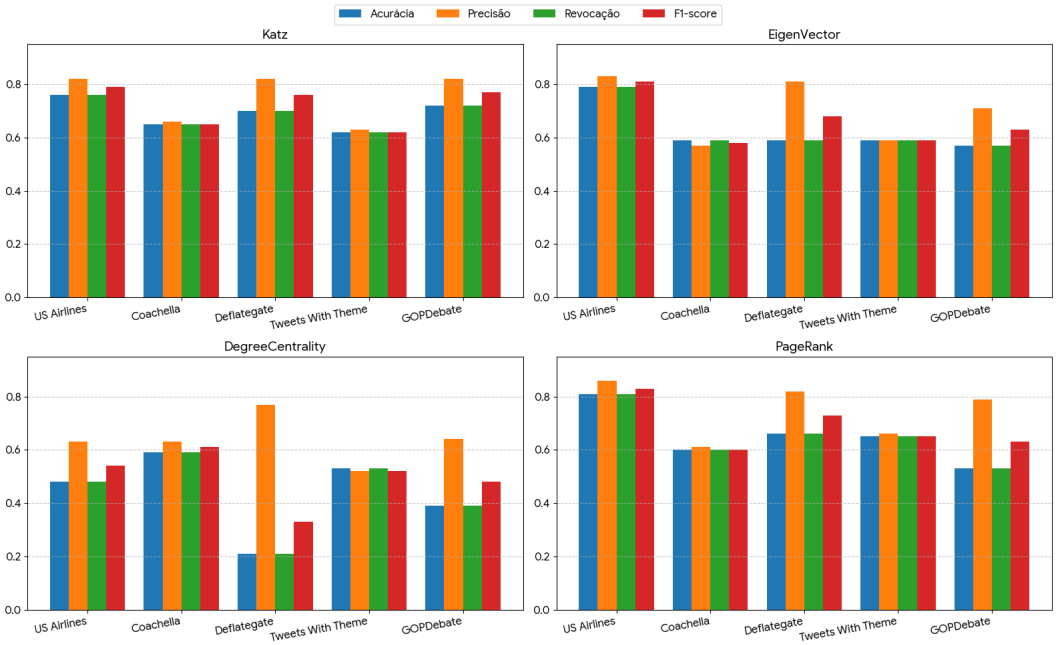

Figura 10 – Gráfico do desempenho da classificação com diferentes medidas de centralidade.

Tabela 12 – Desempenho da classificação com diferentes medidas de centralidade (F1-score corrigido e melhores valores em negrito)

Base	Método	Acurácia	Precisão	Revocação	F1-score
US Airlines	Katz	0,76	0,82	0,76	0,79
	EigenVector	0,79	0,83	0,79	0,81
	DegreeCentrality	0,48	0,63	0,48	0,54
	PageRank	**0,81**	**0,86**	**0,81**	**0,83**
Coachella	Katz	**0,65**	**0,66**	**0,65**	**0,65**
	EigenVector	0,59	0,57	0,59	0,58
	DegreeCentrality	0,59	0,63	0,59	0,61
	PageRank	0,60	0,61	0,60	0,60
Deflategate	Katz	**0,70**	**0,82**	**0,70**	**0,76**
	EigenVector	0,59	0,81	0,59	0,68
	DegreeCentrality	0,21	0,77	0,21	0,33
	PageRank	0,66	**0,82**	0,66	0,73
Tweets With Theme	Katz	0,62	0,63	0,62	0,62
	EigenVector	0,59	0,59	0,59	0,59
	DegreeCentrality	0,53	0,52	0,53	0,52
	PageRank	**0,65**	**0,66**	**0,65**	**0,65**
GOPDebate	Katz	**0,72**	**0,82**	**0,72**	**0,77**
	EigenVector	0,57	0,71	0,57	0,63
	DegreeCentrality	0,39	0,64	0,39	0,48
	PageRank	0,53	0,79	0,53	0,63

robusto, a medida de Katz demonstrou um excelente equilíbrio entre precisão e revocação em cenários de debate e controvérsia.

Outro ponto interessante é a variação de precisão e revocação entre os métodos. O

algoritmo EigenVector, por exemplo, alcançou uma alta precisão em vários *datasets*, como 0,83 em *US Airlines*, mas não necessariamente se traduziu em uma alta revocação ou F1-score em todos os casos. Essa discrepância indica que, embora um método possa ser bom em identificar corretamente as classes positivas, ele pode falhar em capturar todos os verdadeiros positivos, o que impacta diretamente o equilíbrio entre as métricas e a eficácia geral do modelo.

Ao comparar o ACO com outras técnicas de centralidade, observa-se que o feromônio se mostra uma medida competitiva. No dataset US Airlines, por exemplo, o ACO alcançou um F1-score de 0,79, muito próximo do melhor resultado obtido pelo PageRank (0,83). Para a base Tweets With Theme, o desempenho foi quase idêntico, com o ACO atingindo 0,70 e o PageRank 0,65. Isso indica que o feromônio, como medida de centralidade, é tão eficaz quanto as abordagens tradicionais para balancear precisão e revocação, validando sua robustez. Esse resultado indica que o ACO é capaz de sustentar um bom desempenho, mesmo sem sofrer impacto de vieses em favor de classes majoritárias ou minoritárias.

Entretanto, algumas técnicas de centralidade, como PageRank, ainda se destacam ao oferecer um balanceamento ligeiramente melhor entre precisão e revocação, como observado no *dataset US Airlines*, onde PageRank alcança um F1-score de 0,62 em comparação com 0,52 do ACO. Embora o ACO tenha mostrado resultados consistentes em termos de acurácia, o ajuste fino entre precisão e revocação em algumas técnicas de centralidade podem possuir um desempenho ligeiramente superior em cenários onde as classes estão bem balanceadas. Essa vantagem sutil pode estar relacionada à forma como as técnicas de centralidade tratam a estrutura de conexões e interações nos dados, o que facilita a classificação correta de instâncias de ambas as classes. Mesmo em bases equilibradas, essas nuances entre as técnicas de classificação tornam o PageRank e o Katz opções viáveis em termos de F1-score.

5.1.3 Classificação por Votação Majoritária

Por fim, o último experimento consiste em testar a eficácia de um modelo de votação majoritária, combinando as abordagens anteriores para avaliar o desempenho de classificação em um *dataset*. Para cada frase do *dataset*, será atribuído um resultado com base em uma combinação das métricas utilizadas, buscando determinar a polaridade final de acordo com a maioria das classificações obtidas.

No total, 10 combinações distintas, descritas na Tabela 13, serão testadas e avaliadas. Dentre essas combinações, 6 incluem a classificação baseada no ACO, enquanto 4 são compostas apenas pelas outras métricas de centralidade (Katz, Eigen Vector, Degree Centrality, e PageRank). Esta avaliação é crucial para entender como o ACO se integra e se comporta em comparação com outras abordagens de centralidade, possibilitando uma análise mais abrangente do impacto do ACO no processo de classificação.

Tabela 13 – Combinações de algoritmos e suas siglas

Sigla	Katz	PageRank	Eigen Vector	Degree Centrality	ACO
EDA			x	x	x
KDA	x			x	x
KEA	x		x		x
KED	x		x	x	
KPA	x	x			x
KPD	x	x		x	
KPE	x	x	x		
PDA		x		x	x
PEA		x	x		x
PED		x	x	x	

Os resultados obtidos para cada *dataset* é ilustrado pela Figura 12 e descrito em detalhes na Tabela 14. O desempenho das combinações de métricas variou significativamente entre os diferentes conjuntos de dados, refletindo a complexidade e a natureza específica de cada base.

A análise dos resultados demonstra que a formação de comitês por votação majoritária é uma estratégia eficaz, contanto que as métricas constituintes sejam robustas. Em geral, as combinações que incluíam as métricas mais fortes (Katz, PageRank, ACO) apresentaram um desempenho consistentemente elevado. A combinação KPA (Katz+PageRank+ACO), por exemplo, alcançou o melhor F1-score para o dataset *Deflategate* (0,76), enquanto a PEA (PageRank+EigenVector+ACO) obteve a maior acurácia (0,84) em *US Airlines*. Em contrapartida, a inclusão da Degree Centrality em combinações como EDA e KDA geralmente resultou em um desempenho inferior, confirmando que a escolha dos membros do comitê é crucial para o sucesso da votação.

A contribuição específica do ACO para o sucesso dos comitês torna-se clara ao comparar diretamente as combinações que o incluem com aquelas que não o fazem. Modelos como KPA, KEA e PEA frequentemente apresentam um desempenho superior em acurácia e F1-score em relação a conjuntos como PED e KED. Essa superioridade sugere que o ACO introduz uma perspectiva de análise que captura padrões não totalmente cobertos pelas outras métricas. Na base *US Airlines*, por exemplo, todas as cinco melhores combinações em F1-score incluem o ACO, ressaltando sua importância na construção de um classificador mais preciso e equilibrado.

A Figura 11 ilustra visualmente essas conclusões. Ao observar o gráfico de Acurácia, a eficácia das combinações (barras azuis) é aparente, pois frequentemente superam os melhores métodos individuais (barras laranjas). O dataset *US Airlines* serve como um excelente exemplo, onde a combinação PEA atinge o pico de 0,84, um valor superior ao melhor resultado individual de 0,81 (PageRank), evidenciando o ganho obtido pela sinergia dos métodos no processo de votação.

A análise dos gráficos de Precisão e Revocação aprofunda essa percepção. Embora

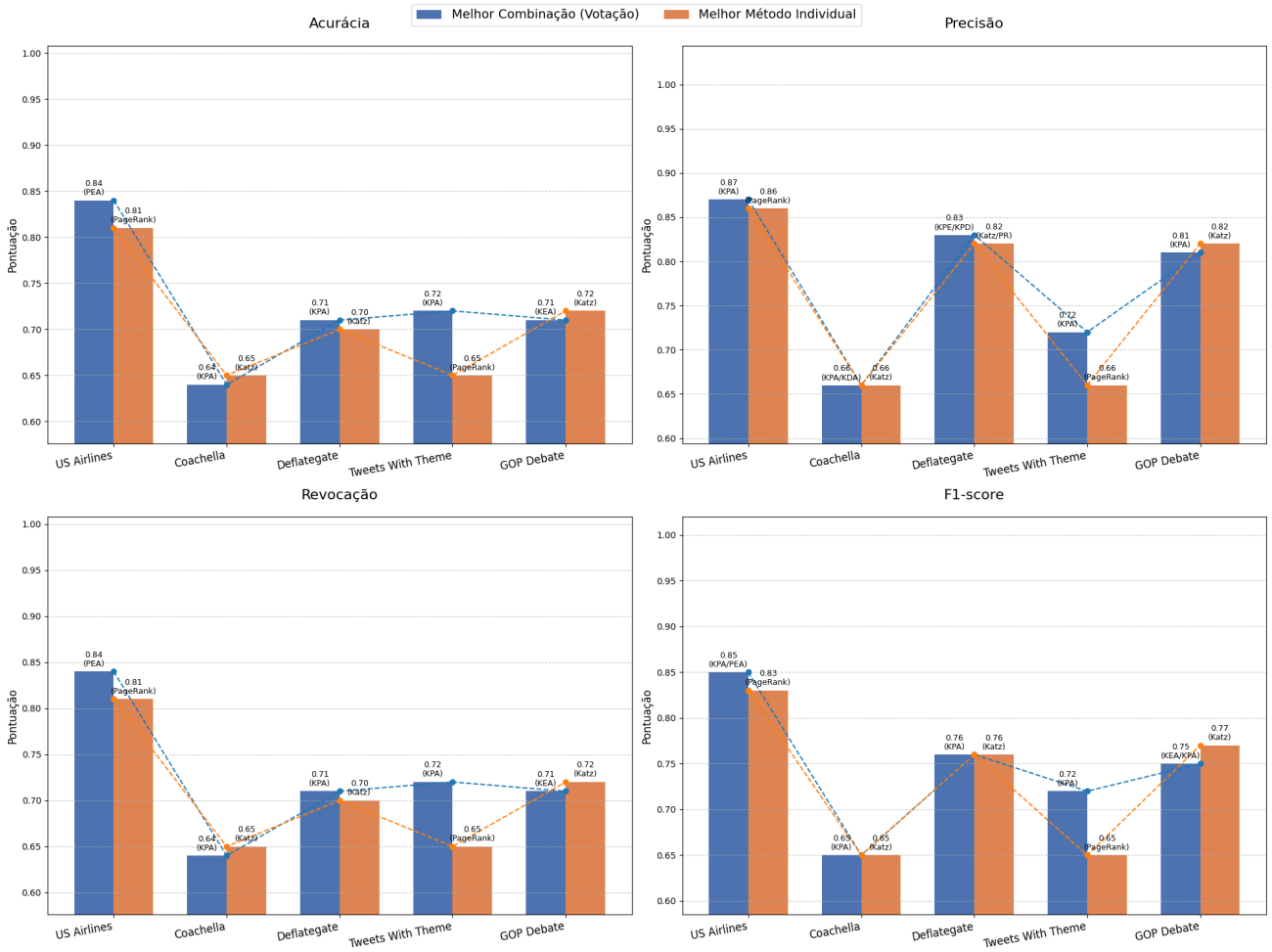

Figura 11 – Comparativo de desempenho máximo para métricas de avaliação

um método individual possa se destacar em uma dessas métricas isoladamente, as combinações que integram o ACO demonstram uma notável estabilidade, mantendo um alto desempenho em ambas. No quesito de Precisão, a combinação KPA alcança o pico de 0,87 no dataset *US Airlines*. De forma similar, na Revocação, a combinação PEA atinge o valor máximo de 0,84 para o mesmo *dataset*. Essa performance consistente demonstra que os comitês com ACO são robustos, sendo precisos em suas classificações e eficazes em identificar o conjunto total de instâncias relevantes.

O gráfico de F1-score, por ser a métrica que harmoniza precisão e revocação, é onde a vantagem da abordagem combinada com o ACO se torna mais proeminente. Nota-se que as combinações superam os métodos individuais na maioria dos cenários. Em *US Airlines*, as combinações KPA e PEA atingem um F1-score de 0,85, enquanto o melhor método individual alcança 0,83. A mesma tendência se repete em *Deflategate* e *GOP Debate*, onde as combinações que incluem o ACO (KPA e KEA) registram os maiores F1-scores (0,76 e 0,75, respectivamente), confirmando que a perspectiva única do algoritmo ajuda

a votação a alcançar um balanço superior.

A razão para essa melhoria reside na diversidade metodológica que o ACO introduz. Enquanto as outras métricas avaliam a importância das palavras com base na estrutura estática da rede, o ACO a explora de maneira dinâmica, através de agentes que depositam feromônios. Isso resulta em uma opnião sobre a centralidade das palavras que é baseada em caminhos e não apenas em conectividade. Essa diversidade no processo de votação cria um modelo final mais sinérgico e resiliente, capaz de mitigar os pontos fracos de cada método isoladamente.

Em suma, os resultados indicam que, embora métodos individuais como Katz e PageRank sejam fortes, a formação de comitês por votação majoritária é uma estratégia superior, e a participação do ACO nesse comitê atua como um fator crucial para o sucesso. O algoritmo não apenas contribui com uma análise competitiva, mas funciona como um potencializador, elevando a robustez e a eficácia da classificação final em cenários complexos ao agregar uma perspectiva de análise que as outras métricas não oferecem.

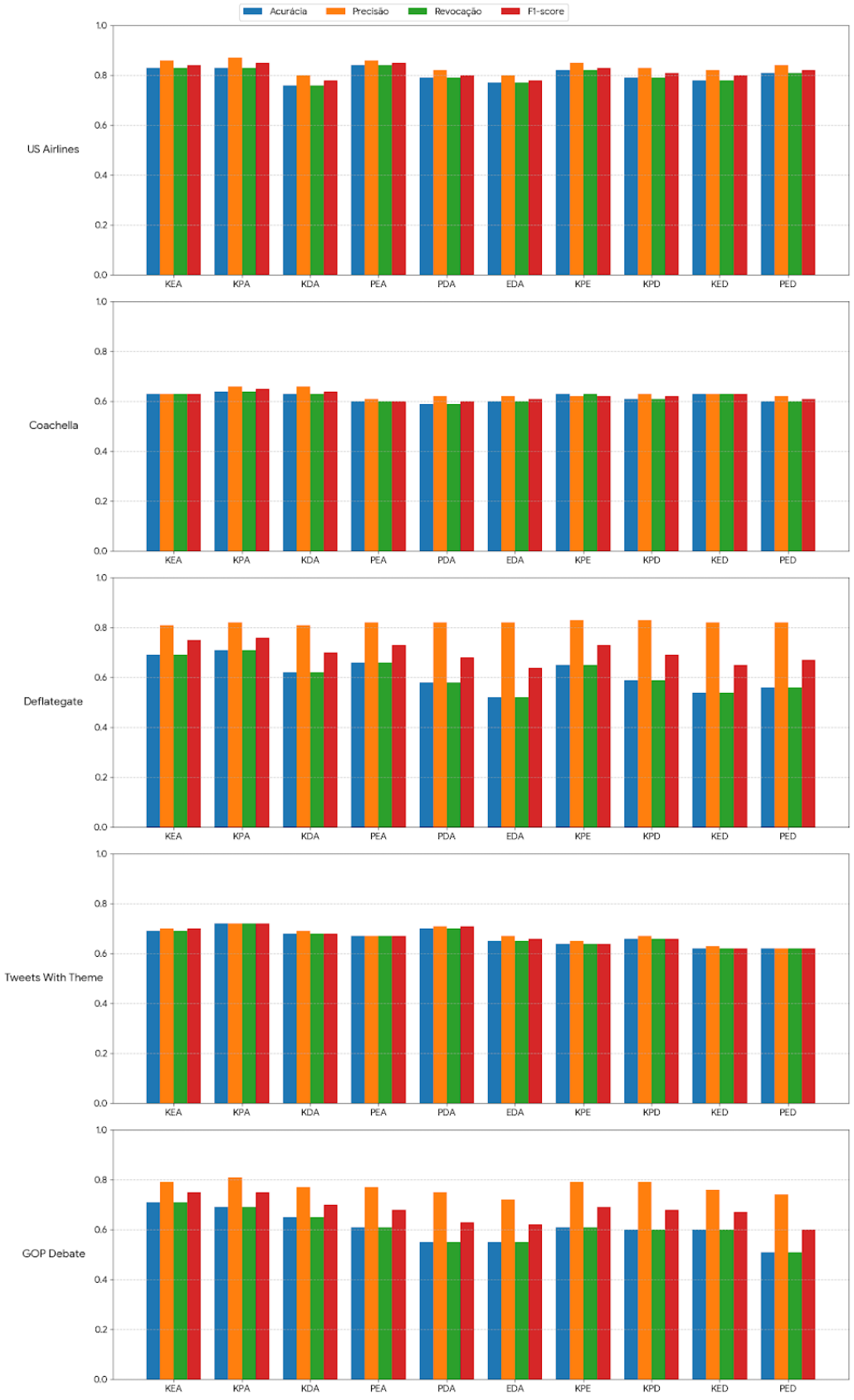

Figura 12 – Gráfico do desempenho das combinações de algoritmos de centralidade para os *datasets* avaliados.

Tabela 14 – Desempenho das combinações de algoritmos de centralidade para os *datasets* avaliados (F1-score corrigido e melhores valores em negrito)

Datasets	**Combinações**	**Acurácia**	**Precisão**	**Revocação**	**F1-score**
US Airlines	KEA	0.83	0.86	0.83	0,84
	KPA	0.83	**0.87**	0.83	**0,85**
	KDA	0.76	0.80	0.76	0,78
	PEA	**0.84**	0.86	**0.84**	**0,85**
	PDA	0.79	0.82	0.79	0,80
	EDA	0.77	0.80	0.77	0,78
	KPE	0.82	0.85	0.82	0,83
	KPD	0.79	0.83	0.79	0,81
	KED	0.78	0.82	0.78	0,80
	PED	0.81	0.84	0.81	0,82
Coachella	KEA	0.63	0.63	0.63	0,63
	KPA	**0.64**	**0.66**	**0.64**	**0,65**
	KDA	0.63	**0.66**	0.63	0,64
	PEA	0.60	0.61	0.60	0,60
	PDA	0.59	0.62	0.59	0,60
	EDA	0.60	0.62	0.60	0,61
	KPE	0.63	0.62	0.63	0,62
	KPD	0.61	0.63	0.61	0,62
	KED	0.63	0.63	0.63	0,63
	PED	0.60	0.62	0.60	0,61
Deflategate	KEA	0.69	0.81	0.69	0,75
	KPA	**0.71**	0.82	**0.71**	**0,76**
	KDA	0.62	0.81	0.62	0,70
	PEA	0.66	0.82	0.66	0,73
	PDA	0.58	0.82	0.58	0,68
	EDA	0.52	0.82	0.52	0,64
	KPE	0.65	**0.83**	0.65	0,73
	KPD	0.59	**0.83**	0.59	0,69
	KED	0.54	0.82	0.54	0,65
	PED	0.56	0.82	0.56	0,67
Tweets With Theme	KEA	0.69	0.70	0.69	0,70
	KPA	**0.72**	**0.72**	**0.72**	**0,72**
	KDA	0.68	0.69	0.68	0,68
	PEA	0.67	0.67	0.67	0,67
	PDA	0.70	0.71	0.70	0,71
	EDA	0.65	0.67	0.65	0,66
	KPE	0.64	0.65	0.64	0,64
	KPD	0.66	0.67	0.66	0,66
	KED	0.62	0.63	0.62	0,62
	PED	0.62	0.62	0.62	0,62
GOP Debate	KEA	**0.71**	0.79	**0.71**	**0,75**
	KPA	0.69	**0.81**	0.69	**0,75**
	KDA	0.65	0.77	0.65	0,70
	PEA	0.61	0.77	0.61	0,68
	PDA	0.55	0.75	0.55	0,63
	EDA	0.55	0.72	0.55	0,62
	KPE	0.61	0.79	0.61	0,69
	KPD	0.60	0.79	0.60	0,68
	KED	0.60	0.76	0.60	0,67
	PED	0.51	0.74	0.51	0,60

Capítulo 6

Conclusão

O objetivo principal deste projeto consistiu na analise da eficiência de métodos de algoritmos inteligentes, comparando o feromônio como medida de centralidade com métricas tradicionais de centralidade como Katz, PageRank, EigenVector e Degree Centrality. Os resultados demonstraram que o algoritmo de colônia de formigas (ACO), utilizando o feromônio, pode oferecer desempenho equivalente ou superior, tornando-o uma alternativa robusta para técnicas de classificação.

Ao aplicar o ACO a redes complexas, foi possível coletar uma tabela de feromônios que reflete a centralidade das palavras para a classificação de textos. O ACO mostrou-se eficaz na identificação de caminhos de palavras relevantes e na geração de bigramas, maximizando o uso da rede de palavras e capturando a importância contextual.

Ao comparar os resultados com o estado da arte, as combinações que utilizam o ACO apresentaram melhorias significativas nas métricas de acurácia e F1-score. Isso evidencia que o ACO oferece vantagens notáveis em relação às abordagens tradicionais. O feromônio revelou-se particularmente eficaz na análise de sentimentos. Sua capacidade de ajustar e otimizar os pesos das características permitiu capturar nuances e padrões complexos, tornando a análise mais precisa e detalhada.

Embora o uso do feromônio exija maior poder computacional, o investimento se justifica pelos benefícios que oferece. A capacidade de captar nuances detalhadas e otimizar a seleção de características para a análise de sentimentos contribui para uma análise mais precisa e detalhada. A maior exigência computacional do ACO deriva de sua natureza iterativa e baseada em agentes. A complexidade pode ser aproximada por $O(I \times F \times C)$, onde I é o número de iterações, F é o número de formigas e C é o custo para uma formiga construir uma solução (que depende do tamanho da rede). Em contraste, medidas como a Degree Centrality são computacionalmente mais simples (próximas de $O(V + E)$, onde V são vértices e E são arestas). Essa sobrecarga computacional, no entanto, é o que permite a construção de uma medida de centralidade dinâmica e contextual, que captura os diferentes aspectos dos dados.

A presente proposta, ao utilizar o feromônio como uma medida de centralidade, oferece

uma vantagem significativa em termos de explicabilidade e interpretabilidade. As trilhas de feromônio podem ser visualizadas, revelando os caminhos semânticos (bigramas) mais influentes na determinação de um sentimento.

6.1 Principais Contribuições

1. Introduzir o uso do algoritmo de otimização por colônia de formigas (ACO) como uma nova abordagem para análise de sentimentos, propondo o feromônio como uma métrica de centralidade alternativa às tradicionais, como Katz, PageRank, EigenVector e Degree Centrality.

2. Demonstrar a eficácia do ACO na seleção de características e na otimização de bigramas, demonstrando de forma prática um desempenho satisfatório nas métricas avaliadas.

3. Comprovar que o ACO é uma alternativa viável e robusta em relação às técnicas tradicionais de classificação de textos.

4. Aplicar o ACO em redes complexas para explorar caminhos relevantes entre palavras, maximizando a representatividade dos textos e mostrando seu potencial como uma ferramenta versátil para análise de grandes volumes de dados textuais.

5. Utilizar algoritmos de inteligência coletiva, como o ACO, em conjunto com técnicas de Processamento de Linguagem Natural (PLN) para melhorar a análise semântica e estrutural de textos, permitindo uma compreensão mais profunda das interações entre palavras e temas.

6.2 Trabalhos Futuros

Os resultados desse projeto demonstram um bom desempenho obtido pelo método proposto. Novas pesquisas podem ser iniciadas a partir, por exemplo, da análise de unigramas somados ao uso de bigramas e até mesmo trigramas afim de se obter novas nuances presentes no texto. A análise de sentimento é apenas uma das várias formas possiveis de se utilizar essa abordagem, sumarização de textos, detecção de tópicos são outras abordagens que se aproveitariam da capacidade do ACO de identificar padrões e relações semânticas. Este método seria particularmente útil em cenários onde o porque de uma classificação é tão importante quanto a própria classificação, como na análise de discurso, identificação de vieses ou na compreensão detalhada da percepção do consumidor. Outra linha de investigação é a utilização do ACO em tempo real em uma rede alimentada dinâmicamente, análogamente ao trabalho de Goel e Prakash (2016), no entanto sob a abordagem proposta neste projeto.

6.3 Contribuições em Produção Bibliográfica

O seguinte artigo é fortemente relacionado a pesquisa elaborada nesta dissertação e atualmente encontra-se em fase de análise:

> TAVEIRA, R. de O.; BACKES, A. R. Sentiment Analysis using Word Centrality based on Ant Colony Pheromones. Submitted to Applied Soft Computing Journal and Neurocomputing.

Referências

ADEDOYIN-OLOWE, M.; GABER, M. M.; STAHL, F. A survey of data mining techniques for social media analysis. **Journal of Data Mining & Digital Humanities**, Episciences. org, v. 2014, 2014.

AGARWAL, A.; XU, J.; LIU, B. Sentiment analysis of twitter data. ACM, p. 1281–1289, 2011.

AGGARWAL, S.; CHHABRA, B. Sentimental analysis of tweets using ant colony optimizations. In: **2017 International Conference on Intelligent Sustainable Systems (ICISS)**. [S.l.: s.n.], 2017. p. 1219–1223.

AHMAD, J. et al. Ant colony optimization for text feature selection in sentiment analysis. **Computers, Materials & Continua**, Tech Science Press, v. 59, n. 1, p. 105–122, 2019.

AIELLO, L. M. et al. Deteccao de eventos no twitter atraves de grafos de visibilidade natural. In: SBC. **Simposio Brasileiro de Computacao**. [S.l.], 2013.

ALOISE D., N. T. F. D. M. R. d. S. B. V. A. D. J. Heuristicas de colonia de formigas com path-ranking para o problema de otimização da alocação de sondas de produção terrestre-spt. **XXXIV Simpósio Brasileiro de Pesquisa Operacional**, 2002.

ALVARES, R. V. **Investigação do processo de Stemming na língua portuguesa**. Tese (Tese (Doutorado)) — Universidade Federal Fluminense, 2005.

ALVAREZ, L.; VILARES, J. A survey of sentiment analysis methods. **Natural Language Engineering**, v. 17, n. 1, p. 1–28, 2011.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, v. 5, n. 1, p. 1–10, 2006.

BARABÁSI, A.-L. **Network Science**. [S.l.]: Cambridge University Press, 2016.

BARION, E. C. N.; LAGO, D. Mineração de textos. **Revista de Ciências Exatas e Tecnologia**, Kroton Educacional, III, n. 3, p. 123–140, 2008. Disponível em: <https://revista.pgsskroton.com/index.php/rcext/article/view/2372/2276>.

BATISTA, G. et al. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, 2004.

BEDI, M. K.; SINGH, S. Comparative study of two natural phenomena based optimization techniques. **International Journal of Scientific Engineering Research**, v. 4, p. 1–4, 2013.

BONACICH, P. Factoring and weighting approaches to status scores and clique identification. **Journal of Mathematical Sociology**, Routledge, v. 2, n. 1, p. 113–120, 1972.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. In: BRISBANE, AUSTRALIA. **Seventh International World-Wide Web Conference (WWW 1998)**. Brisbane, Australia, 1998. p. 107–117.

CHAKRABORTY, K.; BHATTACHARYYA, S.; BAG, R. A survey of sentiment analysis from social media data. **IEEE Transactions on Computational Social Systems**, v. 7, n. 2, p. 450–464, 2020.

CHIRE, J.; MAHMOOD, K.; LIANG, Z. **Complex Networks for Pattern-Based Data Classification**. 2025. Disponível em: <https://arxiv.org/abs/2503.05772>.

DORIGO, M. **Optimization, Learning and Natural Algorithms**. Tese (Doutorado) — Politecnico di Milano, Milan, Italy, 1992.

DORIGO, M.; BLUM, C. Ant colony optimization theory: A survey. **Theoretical Computer Science**, Elsevier, v. 344, n. 2-3, p. 243–278, 2005.

DORIGO, M.; STÜTZLE, T. (Ed.). **Handbook of Metaheuristics**. [S.l.]: Springer, 2019. v. 272. (International Series in Operations Research & Management Science, v. 272).

ERDőS, P.; RéNYI, A. On random graphs. **Publicationes Mathematicae (Debrecen)**, v. 6, p. 290–297, 1959.

ESTRADA, E. Graph and network theory in physics. **Physics Reports**, Elsevier, v. 714, p. 1–119, 2016.

F. de Arruda, H. et al. Paragraph-based representation of texts: A complex networks approach. **Information Processing Management**, v. 56, n. 3, p. 479–494, 2019. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457318305211>.

FERREIRA, J. Z. **Analise de sentimento em documentos financeiros com múltiplas entidades**. Dissertação (Mestrado) — Programa de Pós-graduação em Informática, 2014. Instituto de Computação. Disponível em: <http://tede.ufam.edu.br/handle/tede/4122>.

FREEMAN, L. C. Centrality in social networks: Conceptual clarification. **Social Networks**, Elsevier, v. 1, n. 3, p. 215–239, 1979.

GIACHANOU, A.; CRESTANI, F. A survey of twitter sentiment analysis methods. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 49, n. 2, p. 1–41, 2016.

GOEL, L.; PRAKASH, A. Sentiment analysis of online communities using swarm intelligence algorithms. In: **2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)**. [S.l.: s.n.], 2016. p. 330–335.

GUL, H. et al. A systematic analysis of community detection in complex networks. **Procedia Computer Science**, v. 201, p. 343–350, 2022. ISSN 1877-0509. The 13th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 5th International Conference on Emerging Data and Industry 4.0 (EDI40). Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050922004598>.

HAMDI, M. Affirmative ant colony optimization based support vector machine for sentiment classification. **Electronics**, v. 11, n. 7, 2022. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/11/7/1051>.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. **Intelligent Data Analysis**, IOS Press, v. 6, n. 5, p. 429–449, 2002.

JIM, J. R. et al. Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. **Natural Language Processing Journal**, v. 6, p. 100059, 2024. ISSN 2949-7191. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2949719124000074>.

JOHNSON, J. M.; KHOSHGOFTAAR, T. M. Survey on deep learning with class imbalance. **Journal of Big Data**, Springer, v. 6, n. 1, p. 1–54, 2019.

JURAFSKY, D.; MARTIN, J. **Speech and Language Processing**. 3.ª. ed. Pearson, 2024. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

KABIR, S. Smoking related lung cancer prediction: A comprehensive review and prediction mechanisms. **Department of CSE, East West University, Dhaka Bangladesh**, 2024.

KAPLAN, A. M.; HAENLEIN, M. Using social media analytics to monitor public opinion. **Public Relations Review**, Elsevier, v. 36, n. 1, p. 116–125, 2010.

KATZ, L. A new status index derived from sociometric analysis. **Psychometrika**, Springer, v. 18, n. 1, p. 39–43, 1953.

KAUR, H.; MAHAJAN, P.; KAUR, D. Sentiment analysis of twitter data using hybrid method of support vector machine and ant colony optimization. **International Journal of Computer Applications**, Foundation of Computer Science (FCS), NY, USA, v. 146, n. 13, p. 24–29, 2016.

KIM, Y.; HOVY, E. A deep learning approach to sentiment analysis. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.]: Association for Computational Linguistics, 2014. p. 1366–1375.

KNöBEL, C. From random graphs to complex networks. **arXiv preprint arXiv:cond-mat/0406457**, 2004.

KOTSIANTIS, S. B.; KANELLOPOULOS, D.; TAMPAKAS, V. Supervised machine learning: a review of classification techniques. **International Journal of Information Technology Decision Making**, v. 6, n. 4, p. 491–510, 2007.

LEOPARDI, F.; BISIO, A. A survey on term frequency-based text analysis. **ACM Computing Surveys (CSUR)**, v. 47, n. 4, p. 67, 2015.

LI, G. et al. Identification of disease propagation paths in two-layer networks. **Scientific Reports**, Springer Nature, v. 13, p. 6357, 2023.

LING, C.; LI, C. Data mining for direct marketing: Problems and solutions. In: **Proceedings of The Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: AAAI Press, 1998. p. 73–79.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: **Mining text data**. [S.l.]: Springer, 2012. p. 415–463.

LIU, M. et al. Study on mechanism of jiawei chaiqin wendan decoction in treatment of vestibular migraine based on network pharmacology and molecular docking technology. **Evidence-Based Complementary and Alternative Medicine**, v. 2021, p. 1–12, 10 2021.

MULLEN, R. J. et al. A review of ant algorithms. **Expert Systems with Applications**, Elsevier, v. 36, n. 6, p. 9608–9617, 2009.

OLIVA, S. Z. et al. Text structuring methods based on complex network: a systematic review. **Scientometrics**, Springer, v. 126, n. 2, p. 1471–1493, 2021.

PALMER, D. D. Text preprocessing. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of Natural Language Processing, Second Edition**. Chapman and Hall/CRC, 2010. p. 9–30. Disponível em: <http://www.crcnetbase.com/doi/abs/10.1201/9781420085938-c2>.

PALTOGLOU, G.; THELWALL, M. Sentiment analysis of online opinion using machine learning techniques. **Journal of the American Society for Information Science and Technology**, v. 61, n. 12, p. 2190–2201, 2010.

PALUKURI, M. V.; PATIL, R. S.; MARCOTTE, E. M. Molecular complex detection in protein interaction networks through reinforcement learning. **BMC Bioinformatics**, Springer, v. 24, p. 306, 2023.

PAN, S. J. et al. Sentiment analysis using a boltzmann machine. In: **Proceedings of the 20th ACM Conference on Information and Knowledge Management**. [S.l.]: ACM, 2004. p. 721–728.

PEAKE, J. et al. Paco-vmp: Parallel ant colony optimization for virtual machine placement. **Future Generation Computer Systems**, Elsevier, v. 129, p. 174–186, 2022.

PLISSON, J.; LAVRAC, N.; MLADENIC, D. A rule based approach to word lemmatization. In: JOžEF STEFAN INSTITUTE. **Proceedings of the Department of Knowledge Technologies**. Ljubljana, Slovenia, 2004. p. 1–10.

PONTES, A. **Portuguese Tweets for Sentiment Analysis**. Kaggle, 2023. Disponível em: <https://www.kaggle.com/datasets/augustop/portuguese-tweets-for-sentiment-analysis>.

PRUSA, J. et al. Using random undersampling to alleviate class imbalance on tweet sentiment data. In: **2015 IEEE International Conference on Information Reuse and Integration**. [S.l.: s.n.], 2015. p. 197–202.

RIBAS, L. C. **Aprendizado de representações e caracterização de redes complexas com aplicações em visão computacional**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brasil, 2021. [acesso 2024-09-10].

ROTA, A. R.; NICODEMO, T. L. Arquivos pessoais e redes sociais: o twitter construído como documento histórico. **Estudos Históricos (Rio de Janeiro)**, FGV CPDOC, v. 36, n. 79, p. 362–382, may-aug 2023.

SALAVATI, C.; ABDOLLAHPOURI, A. Identifying influential nodes based on ant colony optimization to maximize profit in social networks. **Swarm and Evolutionary Computation**, v. 51, p. 100614, 2019. ISSN 2210-6502. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2210650218304577>.

SAXENA, A.; IYENGAR, S. **Centrality Measures in Complex Networks: A Survey**. 2020. Disponível em: <https://arxiv.org/abs/2011.07190>.

SILVA, T. C.; ZHAO, L. Machine learning in complex networks. Springer, 2016.

SILVA, T. O. B.; SANTOS, E. et al. Redes complexas: aspectos básicos, importância e aplicações. **Revista Educação e Ciências Sociais**, v. 6, n. 11, p. 28–49, 2024.

SOTO, P. de. Los sistemas de transporte romanos y la configuración territorial en el noroeste peninsular. In: _____. [S.l.: s.n.], 2013.

THAKKAR, M. H. V. Twitter sentiment analysis using hybrid naive bayes. 2013.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. **Data Min Knowl Disc**, v. 24, p. 478–514, 2012.

VEMULA, S. L.; RATHEE, N. Bio-inspired feature selection techniques for sentiment analysis – review. In: **2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)**. [S.l.: s.n.], 2023. p. 808–816.

VILARINHO, G. N.; RUIZ, E. E. S. Global centrality measures in word graphs for twitter sentiment analysis. In: **2018 7th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.: s.n.], 2018. p. 55–60.

WATTS, D. J. Using complex networks to understand the spread of information. **Annual Review of Sociology**, Annual Reviews, v. 33, p. 211–236, 2007.

YADOLLAHI, A.; SHAHRAKI, A. G.; ZAIANE, O. R. Current state of text sentiment analysis from opinion to emotion mining. **ACM Computing Surveys (CSUR)**, ACM, v. 50, n. 2, p. 1–33, 2017.