UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

IBILCE / UNESP - CÂMPUS DE SÃO JOSÉ DO RIO PRETO

Luís Marcello Moraes Silva

Abordagem bioinspirada híbrida de seleção de atributos para classificação de sentimentos em mídias sociais

São José do Rio Preto
2022

Luís Marcello Moraes Silva

Abordagem bioinspirada híbrida de seleção de atributos para classificação de sentimentos em mídias sociais

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Financiador: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES-DS. Proc. 88887.512827/2020-00.

Orientador: Prof. Dr. Carlos Roberto Valêncio

São José do Rio Preto
2022

S586a

Silva, Luís Marcello Moraes

Abordagem bioinspirada híbrida de seleção de atributos para classificação de sentimentos em mídias sociais / Luís Marcello Moraes Silva. -- São José do Rio Preto, 2022

92 f. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Carlos Roberto Valêncio

1. Ciência da computação. 2. Inteligência coletiva. 3. Algoritmos genéticos. 4. Redes sociais. 5. Sistemas especialistas. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Luís Marcello Moraes Silva

Abordagem bioinspirada híbrida de seleção de atributos para classificação de sentimentos em mídias sociais

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Financiador: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES-DS. Proc. 88887.512827/2020-00.

Comissão examinadora:

Prof. Dr. Carlos Roberto Valêncio
UNESP – São José do Rio Preto
Orientador

Prof. Dr. Geraldo Francisco Donegá Zafalon
UNESP – São José do Rio Preto

Prof. Dr. Angelo Cesar Colombini
Universidade Federal Fluminense – Niterói - RJ

São José do Rio Preto
29 de abril de 2022

“Gather ye rose-buds while ye may,

Old Time is still a-flying;

And this same flower that smiles today

Tomorrow will be dying.”

(ROBERT, 1963, p.13-14)

Agradecimentos

Agradeço primeiramente aos meus familiares que me apoiaram e me acompanharam no período de estudos. Agradeço aos meus colegas da computação do IBILCE, agradeço especialmente aos membros e amigos do grupo GBD, em especial ao André Morielo, William Tenório, Gustavo Molina e Sofia Pazzoti. Por fim, agradeço ao professor Valêncio pelo auxílio na iniciação científica, nos trabalhos externos e nesta dissertação de mestrado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

A análise de sentimentos em mídias sociais consiste em extrair informações de usuários presentes nos comentários destas redes sociais. Este tema tem sido amplamente estudado nos últimos anos, pois, por exemplo, pode auxiliar o processo de tomada de decisão de empresas e até identificar intenções e opiniões sobre candidatos em eleições. No entanto, devido ao ambiente *Big Data* no qual estes dados estão inseridos, sua análise tradicional pode ficar comprometida por conta do elevado número de atributos somados a outros fatores. Isto acaba por implicar em uma análise com alto custo computacional e com baixa qualidade de resultados, além do fato de que tal análise é inviável manualmente, pois excede a capacidade humana de entendimento. Pesquisas recentes têm focado em como analisar os sentimentos de usuários com técnicas de aprendizado de máquina somadas às técnicas inspiradas pela natureza, e assim, possibilitar o estudo de opiniões de usuários sobre um determinado tópico. Com o intuito de se analisar tais dados de modo mais preciso, uma seleção de atributos por meio destas abordagens, somado a análises léxicas, tornou-se uma alternativa atrativa para contornar este desafio e viabilizar seu processamento. Este trabalho tem como objetivo a apresentação de uma abordagem híbrida bioinspirada, cuja contribuição científica é a melhoria de um modelo preditivo de classificação de sentimentos multi-idiomas que considera diferentes contextos dos dados. Por meio dos resultados, é possível verificar que o modelo obteve melhorias de acurácia entre 10% e 17%, enquanto que o método de seleção utilizou cerca de 45% dos atributos em relação à análise tradicional.

Palavras-chave: Ciência da computação. Inteligência coletiva. Algoritmos genéticos. Redes sociais. Sistemas especialistas (Computação).

Abstract

The social media sentiment analysis consists on extracting information from users in comments made in their social network. Such topic has been the focus of many study works in the last few years. It can assist the decision-making process of companies, aid teaching methods and even identify and boost intentions and opinions about candidates in elections. However, due to the Big Data environment in which these data are inserted, the traditional analysis can be compromised because of the high dimensionality added to other factors. The implication on the analysis is resulted by high computational cost and low quality of results. Besides that such analysis is impracticable manually as it exceeds the human capacity of understanding. Up to date research has given a focus on how to analyze feelings of users with machine learning and techniques inspired by nature, allowing the study of users' opinions. In order to analyze such data effectively, a feature selection through these approaches is proposed. Machine learning added to lexical analysis has become an attractive alternative to overcome this challenge and facilitate its processing. This paper aims to present a hybrid bio-inspired approach to realize feature selection and improve sentiment classification quality. The scientific contribution is the improvement of a classification model considering pre-processing of the data with different languages and contexts. The results prove that the developed method enriches the predictive model by improving the accuracy by 10% to 17%. This method selected 45% of the attributes in average compared to traditional analysis.

Keywords: Computer science. Swarm intelligence. Genetic algorithms. Social networks. Knowledge acquisition (Expert systems).

Lista de Ilustrações

Figura 1 - Tópicos de pesquisas em AS 22
Figura 2 - Processo ETA 25
Figura 3 - Diagrama de algoritmos inspirados pela natureza 31
Figura 4 - Fluxograma do processo de seleção de atributos 32
Figura 5 - Demonstração do VL em duas dimensões 34
Figura 6 - Pseudocódigo do algoritmo BC 35
Figura 7 - Operador genético de reprodução 39
Figura 8 - Operador genético de mutação 40
Figura 9 - Pseudocódigo do AG 40
Figura 10 - Diagrama do trabalho desenvolvido 48
Figura 11 - Exemplo de pré-processamento 50
Figura 12 - Fluxograma do algoritmo BCG 60
Figura 13 - Pseudocódigo do algoritmo BCG 61
Figura 14 - Função de teste Rastrigin 66
Figura 15 – Comparação do desempenho de BCG com variação da ocorrência 67
Figura 16 - Desempenho da BCG com $\boldsymbol{f1}$ 73
Figura 17 - Desempenho da BCG com $\boldsymbol{f2}$ 73
Figura 18 – Comparação do desempenho da BCG com BC e AG 81

Lista de Tabelas

Tabela 1 - Comparativo entre trabalhos correlatos 44
Tabela 2 - Exemplos de sinônimos relacionados a contexto 52
Tabela 3 - Especificações do ambiente de teste 63
Tabela 4 - Especificações das bases de testes 64
Tabela 5 - Parâmetros do algoritmo BCG 68
Tabela 6 - Acurácia base com quatro com classificadores 69
Tabela 7 – Acurácia com $f1$ na base A com quatro classificadores 70
Tabela 8 – Quantidade de atributos com $f1$ na base A com classificador ME em cinco estados 70
Tabela 9 - Acurácia com $f2$ em quatro classificadores 71
Tabela 10 - Quantidade de atributos com $f2$ com classificador ME em cinco estados 72
Tabela 11 – Comparação de tempo de execução das funções $f1$ e $f2$ com BCG 74
Tabela 12 - Acurácia do método BCG com $f1$ nos quatro classificadores 74
Tabela 13 - Quantidade de atributos selecionados com $f1$ nos quatro classificadores 75
Tabela 14 - Acurácia (%) com método filtro *K-Best* 76
Tabela 15 - Acurácia (%) com método *wrapper* RFE 76
Tabela 16 – Valores de aumento na acurácia com métodos tradicionais 77
Tabela 17 – Comparação de acurácia com método ME entre BC e BCG 78
Tabela 18 – Métricas de qualidade e desempenho entre BC e BCG 78
Tabela 19 – Comparação de acurácia com método ME entre AG e BCG 79
Tabela 20 – Métricas de qualidade e desempenho entre AG e BCG 80
Tabela 21 – Testes estatísticos sobre acurácia entre AG e BCG 81
Tabela 22 – Comparação de aspectos de correlatos com o trabalho feito 85

Lista de Abreviaturas e Siglas

ACO – Otimização por Colônia de Formigas

AS – Análise de Sentimentos

AG – Algoritmo Genético

BC – Busca Cuco

BCG – Busca Cuco Genético

CRF – Campos Aleatórios Condicionais

CSV – Valores Separados por Vírgulas

DF – Frequência de Documento

DM – Mineração de Dados

ETA – Extração, Transformação e Armazenamento

GBD - Grupo de Banco de Dados

KDD - Descoberta de Conhecimento em Bases de Dados

ME – Entropia Máxima

ML – Aprendizado de Máquina

NB – Bayesiano Ingênuo

NLP – Processamento de Linguagem Natural

PSO – Otimização por Enxame de Partículas

RF – Floresta Aleatória

RFE – Eliminação de Atributos Recursiva

SFS – Seleção de Atributos Sequencial

SQL – Linguagem de Consulta Estruturada

SVM – Máquina de Vetores de Suporte

TDM – Matriz de Termos de Documentos

TF-IDF – Frequência do Termo-Inverso da Frequência de Documentos

VL – Voos de Levy

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Motivação e escopo 14
1.1.1 Tratamento léxico de dados textuais 14
1.1.2 Abordagem meta-heurística de seleção de atributo 15
1.2 Objetivos e metodologia 15
1.3 Contribuições 16
1.4 Organização da dissertação 16
2 ANÁLISE DE SENTIMENTOS EM *BIG DATA* 17
2.1 Contexto *Big Data* 17
2.2 Mídias Sociais 19
2.2.1 Mídias sociais e *Big Data* 20
2.2.2 Utilização de contexto em mídias sociais 20
2.3 Análise de sentimentos 21
2.3.1 Mineração de opinião em português 23
2.4 Extração e transformação de dados 24
2.5 Seleção de atributos 27
2.5.1 Definição formal 27
2.5.2 Abordagem determinística 28
2.5.3 Abordagem estocástica 30
2.6 Busca cuco 33
2.7 Algoritmo genético 38
2.8 Trabalhos correlatos 41
2.8.1 Um método de duas etapas para análise de sentimentos embasada em aspectos 41
2.8.2 Desafios no desenvolvimento de abordagens híbridas para análise de sentimentos 42
2.8.3 Análise de sentimentos com busca cuco para seleção de atributos em *tweets* 42
2.8.4 Seleção de atributos por meio de inteligência de enxames 43
2.8.5 Filtro híbrido eficiente e abordagem evolutiva para análise de sentimentos 43
2.8.6 Comparação com trabalhos correlatos 44
2.9 Considerações finais 45

3 ABORDAGEM HÍBRIDA BIOINSPIRADA 46
3.1 Escopo do algoritmo 46
3.2 Coleta de dados 48
3.3 Pré-processamento 49
3.3.1 Tradução e limpeza inicial 50
3.3.2 Aplicação de contexto 51
3.3.3 Padronização final 53
3.4 Extração de características 54
3.5 Algoritmo bioinspirado 54
3.5.1 Abstração das soluções 55
3.5.2 Geração de solução por BC 56
3.5.3 Operadores genéticos 57
3.5.4 Visão geral 58
3.6 Considerações sobre o trabalho desenvolvido 61
4 TESTES E RESULTADOS 62
4.1 Metodologia de experimentação 62
4.2 Materiais e métodos 63
4.2.1 Bases de dados 63
4.2.2 Especificações dos testes 65
4.3 Avaliação do método 65
4.4 Testes com BCG 68
4.4.1 Comparações das funções de aptidão 69
4.4.2 Comparação com métodos determinísticos 75
4.4.3 Comparação com métodos meta-heurísticos 78
4.5 Considerações finais sobre os testes 82
5 CONCLUSÃO 83
5.1 Contribuição científica 84
5.2 Trabalhos futuros 85
REFERÊNCIAS 87

Capítulo 1

Introdução

Atualmente, a sociedade passa por um período de transformações tecnológicas, uma parte desta transformação ocorre devido ao fato de que as mídias sociais passaram a fazer parte do cotidiano das pessoas em vários segmentos (KRAIEM et al., 2019; YUE et al., 2019), além de terem se tornado um recurso de comunicação importante (YADAV; VISHWAKARMA, 2020). Tais redes sociais como *Facebook* e *Twitter* acabaram por representar um papel importante para várias áreas que necessitam de suporte no processo de tomada de decisão (ROUT et al., 2018).

Estas áreas podem ser: melhorias em produtos e serviços (KO et al., 2017); auxilio em processos de ensino (OKTAVIA et al., 2017); detecção de notícias falsas (SHU et al., 2017); identificação de perfis de suicídio (VIOULÈS et al., 2018); e até mesmo meios de comunicação e influência política (YADAV; VISHWAKARMA, 2020; YUE et al., 2019; OLIVEIRA; BERMEJO; DOS SANTOS, 2017).

Desta forma, cada vez mais pesquisas sobre análise de mídias sociais são desenvolvidas, em particular, pesquisas que possuem foco na mineração de opinião ou análise de sentimentos (AS) e, consequentemente, na área de processamento de linguagem natural, do inglês *natural language processing* (NLP), têm recebido bastante destaque (LIMA; DE CASTRO; CORCHADO, 2015; TRIPATHY; AGRAWAL; RATH, 2016).

Tal área é importante no processo de tomada de decisão, pois permite aos pesquisadores avaliar quais os pensamentos de determinados grupos de pessoas sobre um dado assunto (ROUT et al., 2018). Análises podem ser realizas em nível de documento, em que se considera que todo o texto possui apenas uma opinião, ou uma análise em nível de sentença, em que cada frase possui uma opinião própria, tais opiniões podem ser positivas, negativas ou neutras (YUE et al., 2019; APPEL et al., 2016).

Um importante aspecto para que a extração de conhecimento seja possível é a seleção de atributos (YUE et al., 2019). Uma vez que os dados da análise de opinião, mesmo com poucas amostras, podem gerar uma grande quantidade de atributos ou características, associados à quantidade de palavras e símbolos diferentes. Isto pode gerar efeitos negativos junto ao processo de extração de conhecimento, tais como: baixa precisão, baixa acurácia e tempo de processamento elevado se não for tratado adequadamente (UYSAL, 2016).

Devido à natureza dos dados, a tarefa de selecionar os melhores atributos pode ser custosa em termos computacionais, pois, tradicionalmente, para um número inteiro n de atributos, existem 2^n subconjuntos possíveis, o que torna este problema intratável (PANDEY; RAJPOOT; SARASWAT, 2020). Devido ao elevado volume de dados, pesquisadores desenvolveram métodos que exploram heurísticas para encontrar boas soluções, neste caso, bons subconjuntos de atributos de tal forma que sejam encontrados rapidamente e com uma acurácia suficientemente boa (ALARIFI et al., 2020; ABD EL AZIZ; HASSANIEN, 2018; PANDEY; RAJPOOT; SARASWAT, 2020).

Além disso, é necessário destacar o desafio de se realizar a análise de opinião em mídias sociais, uma vez que o volume e variedade dos dados excedem a capacidade humana de compreensão (AKHTAR et al., 2017; VALÊNCIO et al., 2020). Uma limitação frequente é o fato de que pesquisadores lidam apenas com uma língua presente nos dados, em geral inglês, e ignoram os dados de outras línguas. Destaca-se que a língua inglesa representa 25% do conteúdo gerado em sites e, por exemplo, o português é uma das cinco línguas mais utilizadas na internet (PEREIRA, 2020).

Processos de tomada de decisão de empresas e instituições podem expandir suas capacidades de análises ao se considerar não apenas um idioma, além de buscar alternativas que os habilitem na execução de suas tarefas em tempo hábil com pouca ou nenhuma perda de qualidade. Para tal, é necessário tratar a análise de sentimentos como um problema do universo *Big Data*, isto é, é preciso considerar alternativas recentes para extrair, transformar e selecionar os melhores atributos a partir dos dados de usuários de mídias sociais.

Este trabalho propõe continuar a pesquisa iniciada por Valêncio et al., (2020), no sentido de buscar um método que tenha como foco realizar a análise de sentimentos em mídias sociais.

Com base nisto, um método fundamentado no algoritmo meta-heurístico Busca Cuco (BC) se mostra interessante para selecionar o melhor conjunto de atributos, e, em conjunto, um método de limpeza e tratamento de dados textuais embasados na literatura tem potencial para lidar com estes desafios (PANDEY; RAJPOOT; SARASWAT, 2020; HASSONAH et al., 2020; ZAINUDDIN; SELAMAT; IBRAHIM, 2018; ABD EL AZIZ; HASSANIEN, 2018; YADAV; VISHWAKARMA, 2020; SHARMA; KAUR, 2020).

1.1 Motivação e escopo

Como mencionado, devido ao fato de que os dados de redes sociais podem ser utilizados em diversas áreas estratégicas (YUE et al., 2019), somado ao fato de que pesquisas em análise de sentimentos tem apresentado crescimento importante (MÄNTYLÄ; GRAZIOTIN; KUUTILA, 2018), este trabalho apresenta-se como uma possibilidade para se detectar padrões e tendências neste contexto.

De acordo com o estado da arte, um método híbrido de seleção de atributos embasado no algoritmo BC, juntamente com um tratamento de dados textuais específico para mídias sociais que lide com o contexto dos dados (KUMAR; GARG, 2019) e com multi-idiomas (PEREIRA, 2020), mostra-se como uma opção pertinente para lidar com os desafios deste cenário (HASSONAH et al., 2020; PANDEY; RAJPOOT; SARASWAT, 2020).

1.1.1 Tratamento léxico de dados textuais

Tradicionalmente, é necessário aplicar um tratamento em dados textuais antes de se aplicar a mineração de dados em si (HASSONAH et al., 2020). Logo, este trabalho visa aplicar os tratamentos léxicos de dados indicados na literatura (HASSONAH et al., 2020; ZAINUDDIN; SELAMAT. IBRAHIM, 2018; CIRQUEIRA et al., 2018), de modo que esta limpeza se adeque aos padrões do respectivo estado-da-arte.

Somado a isto, é importante realizar uma limpeza que consiga tratar de forma adequada não apenas características especiais dos dados oriundos de redes sociais, mas também leve em consideração o contexto dos dados, pois trata-se de recurso que contribui na melhora do processo de extração de conhecimento (KUMAR; GARG, 2019).

Por outro lado, a literatura ressalta a relevância de tratar os dados textuais em língua portuguesa (PEREIRA, 2020), assim, este trabalho leva isto em consideração, de modo que o estudo seja mais amplo e permita esta análise. Então os dados são convertidos para a língua inglesa e, desta forma, é possível aplicar as mesmas regras de limpeza.

1.1.2 Abordagem meta-heurística de seleção de atributo

Além dos desafios relacionados ao tratamento léxico dos dados, existe o problema da seleção de atributos. Mesmo que os dados passem por um processo de normalização, muitas palavras distintas precisam ser analisadas, de modo que cada uma pode ser considerada um atributo e, portanto, quanto maior o número de sentenças dadas como entrada, maior a dimensionalidade do problema.

Com isto, uma meta-heurística que passou a ser utilizada para este caso é a Busca Cuco (PANDEY; RAJPOOT; SARASWAT, 2020; PANDEY; RAJPOOT; SARASWAT, 2017; ABD EL AZIZ; HASSANIEN, 2018; KUMAR et al., 2019), de forma que este método pode ser aplicado para encontrar um subconjunto de atributos que sejam mais relevantes e, assim, garantir que a classificação de polaridade tenha maior qualidade.

Aplicar apenas este método sem alterações não é interessante, pois esta abordagem lida com dados contínuos enquanto que a análise de opinião se utiliza de dados binários, além de que o algoritmo pode convergir prematuramente para uma solução local (YADAV; VISHWAKARMA, 2020). Logo, mesclá-lo com o Algoritmo Genético (AG) pode ser vantajoso, pois pode melhorar a qualidade das soluções e aumentar sua variabilidade de modo a se obter um método balanceado, e o AG é naturalmente ajustável para realizar a classificação de polaridade (LUI; WANG, 2019; YADAV; VISHWAKARMA, 2020; SHARMA; KAUR, 2020).

1.2 Objetivos e metodologia

O objetivo deste trabalho é apresentar um sistema de tratamento de dados de mídias sociais que suporte a análise de sentimentos embasada em uma abordagem híbrida, inicialmente com uma análise léxica, e, posteriormente, por meio de um algoritmo de seleção de atributos desenvolvido que combina as abordagens meta-heurísticas BC e AG.

Para tal, inicialmente os dados textuais são submetidos a um processo de formatação que inclui a aplicação de técnicas de limpeza de dados. Somado a isto, efetua-se uma análise multi-idiomas com dados em português e inglês, assim como considera o contexto dos dados.

Um conjunto de testes foi conduzido com o propósito de estimar os parâmetros iniciais da técnica desenvolvida, e, posteriormente, analisar como a abordagem de seleção de atributos alterou a execução de algoritmos classificadores, de forma a ratificar a qualidade dos resultados em comparação a técnicas tradicionais e estocásticas encontradas na literatura.

1.3 Contribuições

A contribuição científica deste trabalho é a apresentação de um método de seleção de atributos no cenário de mineração de opinião multi-idiomas e que considera o contexto dos dados em sua análise. Tal contribuição permite a simplificação e seleção de um conjunto de atributos mais relevantes, o que garante um processamento mais rápido e com maior qualidade mesmo em casos com diferentes contextos e idiomas. Os resultados empíricos revelam que o algoritmo elaborado conseguiu identificar soluções com valores de acurácia 12% maiores em média, além de apresentar reduções de 50% a 65% no número de atributos a depender da função de aptidão utilizada.

Além disso, este estudo contribui com a divulgação de um algoritmo que combina os benefícios das técnicas BC e AG de modo a se obter um algoritmo balanceado que descobre boas soluções em um tempo de execução reduzido, além selecionar um menor subconjunto de atributos com soluções expressivas.

1.4 Organização da dissertação

Neste capítulo foram introduzidos os desafios relacionados à análise de sentimentos para dados de mídias sociais, bem como os desafios ligados ao processamento léxico neste cenário e os objetivos. O restante da dissertação está organizado como segue: Capítulo 2 apresenta a fundamentação teórica que embasa o desenvolvimento do trabalho e o estado da arte sobre seleção de atributos por meio de métodos meta-heurísticos; Capítulo 3 apresenta a proposta deste trabalho e a sua descrição; Testes e resultados apresenta os experimentos realizados e os resultados obtidos; Capítulo 5 discorre sobre as conclusões.

Capítulo 2

Análise de Sentimentos em *Big Data*

Neste capítulo é apresentada a fundamentação teórica e recursos descritos na literatura que suportam as atividades de análise de sentimentos em mídias sociais. Conceitos sobre *Big Data* e mídias sociais são apresentados e as técnicas de análise e de limpeza mais empregadas. Por fim, é apresentado o estado da arte sobre seleção de atributos, tanto de métodos tradicionais quanto de abordagens bioinspiradas, juntamente com as vantagens e desvantagens de cada abordagem.

2.1 Contexto *Big Data*

O conceito denominado *Big Data* é amplamente difundido e é um grande alvo de pesquisas científicas devido ao seu potencial no auxílio em tomadas de decisão. A análise dos dados deste universo tornou-se uma tendência tanto em empresas quanto em instituições de pesquisa com o propósito de identificar informações úteis nestes dados (YUE et al., 2019).

Este conceito é representado por seus V’s e pode ser entendido como um conjunto de dados, que além dos dados convencionais, é composto por dados textuais, áudios, imagens, vídeos e os dispositivos geradores destes dados, como plataformas de internet das coisas, sensores e afins.

Tais dados estão dispersos de forma que ambos, o seu volume de dados e a velocidade de geração necessitam de um processamento que excede os recursos computacionais convencionais (SIVARAJAH et al., 2017).

Para ilustrar tal abrangência, é estimado que, em 2014, a quantidade de dados gerados por dia ultrapassava 2,5 exabytes (10^{18} *bytes*) e ainda, cerca de 90% destes dados eram não estruturados. Somado a isto, especula-se que em 2022 o volume de dados gerados seja de mais de 97 *zettabytes* (10^{21} *bytes*) segundo o portal *Statista* (2022).

Com isto, é possível perceber os desafios associados ao *Big Data* e que podem ser divididos em três categorias: características dos dados, processos e gerenciamento destes dados. As características dos dados são apresentadas na literatura de acordo seus V's. Na literatura, vários aspectos são considerados, no entanto é possível destacar os seguintes (SIVARAJAH et al., 2017; STOREY; SONG, 2017; GANDOMI; HAIDER, 2015):

a) Volume – Trata-se da magnitude dos conjuntos de dados que podem ter desde *terabytes* (10^{12} *bytes*) até exceder os *zettabytes* (10^{21} *bytes*) e são oriundos de diversas fontes de dados distintas;
b) Velocidade – Trata-se da velocidade de se gerar, capturar, extrair, processar e armazenar os dados, ou a elevada taxa de chegada dos dados heterogêneos que precisam ser tratados de forma suficientemente rápida;
c) Variedade – Trata-se de múltiplos formatos de dados que podem ser estruturados, semiestruturados e não estruturados, eles são: textos; imagens; bases de dados; vídeos; dados de sensores; dentre outros;
d) Variabilidade – Trata-se dos dados cujo significado está em constante mudança, logo os dados podem ter sua semântica alterada com o passar do tempo;
e) Valor – Trata-se da utilidade da extração de conhecimento feita nos dados sem perder a qualidade que os dados podem ter, e assim, ser capaz de identificar as informações relevantes para os analistas;
f) Veracidade – Trata-se da confiabilidade, complexidade, qualidade, imprecisão e inconsistência que são inerentes aos dados deste contexto.

Na literatura recente, é possível identificar outros V's, tais como visualização, validade, volatilidade e vulnerabilidade, em que, para alguns casos, incluem conceitos já descritos nos V´s anteriores (STOREY; SONG, 2017).

2.2 Mídias Sociais

Um subconjunto do universo *Big Data* são as mídias sociais. Mídias ou redes sociais podem ser consideradas como uma série de aplicações que são disponibilizadas na internet, de modo que suas principais características incluem o fato de serem continuamente modificadas pelos seus usuários por meio de participações, colaborações e compartilhamentos de dados (GHANI et al., 2019). Devido ao desenvolvimento e evolução de *hardware* e *software* nos últimos anos, foi possível que cerca de 2,9 bilhões de usuários passassem a utilizar as redes sociais como o *Facebook*, segundo o site *Datareportal* (2022), e, com isto, tem-se disponível para estudo, dados gerados por usuários (GHANI et al., 2019).

De acordo com BATRINCA e TRELEAVEN (2015), as redes sociais são a maior, mais rica e dinâmica representação de comportamento humano feita nos últimos anos, de forma que traz possibilidades de estudos de indivíduos, grupos específicos e sociedade de modo conjunto. Com isto, existem três áreas que podem se utilizar das redes sociais para melhorar o seu processo de tomada de decisão, estas áreas são as de: negócios; biociências e ciências sociais, explicadas a seguir.

Segundo BATRINCA e TRELEAVEN (2015), os estudos nas áreas de negócios possuem foco principalmente em comércio e finanças, tais estudos são centrados em reconhecimento do impacto da marca ou empresa, melhoria e descoberta de produtos e serviços (KO et al., 2017), estratégias de marketing e até detecção de fraudes ou de notícias falsas (SHU et al., 2017). Neste sentido, foi possível identificar e analisar os sentimentos dos usuários e, assim, ter um auxílio em relação à predição de preços de produtos em um caso de estudo (BATRINCA; TRELEAVEN, 2015).

Em relação a biociências, os dados são coletados para analisar as mudanças de comportamentos em humanos e monitorar impactos de eventos, isto pode ser aplicado para analisar os casos de tabagismo, obesidade, ideação de suicídios (VIOULÈS et al., 2018), dentre outros problemas relacionados à saúde (BATRINCA; TRELEAVEN, 2015).

Por fim, ciências sociais computacionais incluem alguns estudos como: monitoramento de respostas dos usuários a discursos, anúncios ou eventos de natureza política; detecção de comportamento de grupos específicos de usuários; e identificação precoce de eventos recentes; por exemplo, foi possível identificar por meio do *Twitter* as preferências de votos em uma eleição, e assim, prever de forma semelhante às pesquisas de intenções de votos quem ganharia a eleição (OLIVEIRA; BERMEJO; DOS SANTOS, 2017).

2.2.1 Mídias sociais e *Big Data*

É importante ressaltar que devido ao fato das redes sociais serem uma aplicação contida no universo *Big Data*, os seus dados podem possuir as características dos V's apresentadas. As áreas de descoberta de conhecimento mais comuns em mídias sociais e *Big Data* são: descobrimento de tendências, análise de mídias sociais e mineração de opinião (GHANI et al., 2019). O descobrimento de tendências se refere à detecção de tópicos ou eventos que ocorrem no cotidiano e que refletem em qual é o assunto mais comentado e como é sua repercussão segundo os usuários (NAZIR et al., 2019). Por outro lado, a análise de sentimentos também conhecida por mineração de opinião, em inglês *opinion mining*, pode se referir a uma área que busca identificar qual o sentimento presente em uma dada frase ou oração, assim como busca identificar quem emitiu a opinião e qual o tópico citado (YUE et al., 2019).

Em geral, os tomadores de decisão de empresas e instituições possuem interesse em entender quais são os tópicos que podem beneficiar ou prejudicar a imagem da organização, além de ser importante analisar os dados e prever possíveis cenários e comportamento dos usuários, isto é viável por meio da análise de mídias sociais (NAZIR et al., 2019). Contudo, é preciso desenvolver técnicas computacionais que propiciem tal extração de conhecimento uma vez que a análise manual é inviável (AKHTAR et al 2017).

2.2.2 Utilização de contexto em mídias sociais

Um aspecto importante para considerar ao se aplicar KDD em mídias sociais, sobretudo quando o foco é analisar os sentimentos e opiniões de usuários, é o contexto dos dados (KUMAR; GARG, 2019). Ao extrair informações, neste caso, a partir de textos quase sempre informais, é importante saber que, somente a análise do texto pode ser um fator limitante em termos da qualidade da informação obtida. Tais limitações ocorrem porque a análise de sentimentos deve levar em consideração o contexto social do usuário ou outros fatores que possam interferir (KUMAR; GARG, 2019).

Uma determinada palavra pode apresentar sentimentos diferentes dependendo do contexto em que está inserida. Por exemplo, a palavra "imprevisível" se utilizada em uma crítica de filmes porque possui muitas reviravoltas, normalmente está associada a um sentimento positivo, enquanto que a palavra "imprevisível", se associada como uma descrição de uma pessoa, geralmente implica em uma opinião negativa sobre o indivíduo.

Algumas variações na escrita das palavras decorrem do fato dos textos de mídias sociais serem essencialmente informais, assim abreviações não convencionais, gírias, emojis e erros ortográficos são comuns. Isto gera dificuldades no processamento, o que reforça a necessidade de se analisar o contexto e, com isto, é possível melhorar a qualidade da análise de sentimentos, de modo que outras limitações tenham um impacto menor (KUMAR; GARG, 2019; HASSONAH et al., 2020).

2.3 Análise de sentimentos

Formalmente, a análise de sentimento é definida como um conjunto de cinco atributos da forma $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, em que: e_i representa a entidade alvo; a_{ij} representa uma característica da entidade e_i; s_{ijkl} representa o valor da opinião ou sentimento do detentor h_k desta opinião; e, por fim, t_l representam o momento temporal em que a opinião foi feita (APPEL et al., 2016). O atributo s_{ijkl} associa-se a apenas os valores a seguir: positivo, negativo ou neutro, isto é, pode ser interpretado como $s \in S$; $S = \{positivo, negativo, neutro\}$ (YUE et al., 2019; KUMAR; GARG, 2019). Algumas variações podem ocorrer na literatura, de modo que pode haver um sexto atributo p_l que se refere ao local onde o sentimento foi expresso (YUE et al., 2019).

Deste modo, a análise de sentimentos pode ser entendida como um conjunto métodos de análise de textos para extrair informações sobre os sentimentos presentes, assim, trata-se de uma técnica de NLP que tem o foco na mineração de texto ou *text mining* em inglês (KUMAR; GARG, 2019). Existem outras abordagens que levam em consideração imagens e vídeos presentes nas redes sociais (LI et al., 2019), mas não são o foco.

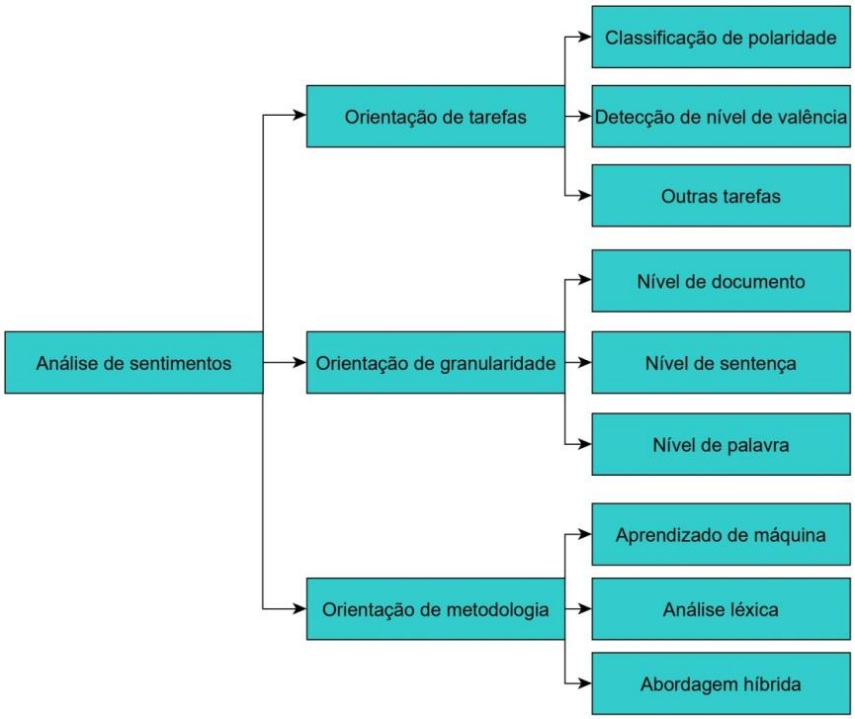
Assim, a análise de sentimentos em textos de mídias sociais está principalmente voltada para a forma como os usuários respondem ou reagem a uma determinada publicação ou acontecimento, isto se dá por meio dos comentários. É possível analisá-los de diversas formas, isto é, o comentário pode ser analisado sob a ótica de cada um dos seus atributos, além de que o texto original pode ser segmentado e reorganizado. Junto a isto, existem métodos que realizam a seleção de atributos que podem ser mais relevantes (ABD EL AZIZ; HASSANIEN, 2018), e métodos que buscam realizar manipulações nos comentários para identificar qual o sentimento presente no texto (APPEL et al., 2018).

Isto implica que diversos estudos podem ser realizados sob a ótica de vários métodos, e devido às propriedades altamente variáveis dos dados, não há um modelo que seja generalista o bastante para suprir a demanda de forma razoável para qualquer caso neste

cenário. No entanto, dentre eles há uma estratégia interessante embasada na metodologia híbrida, tal estratégia consegue levar em consideração os aspectos positivos de cada técnica, e, ao mesmo tempo, consegue lidar com os desafios impostos pelas mídias sociais no que diz respeito às características *Big Data* (LIMA; DE CASTRO; CORCHADO, 2015; ZAINUDDIN; SELAMAT; IBRAHIM, 2018; APPEL et al., 2018).

Segundo Yue et al., (2019), existem três grandes áreas para as quais os pesquisadores direcionam suas pesquisas. Assim, a análise de sentimentos pode ser dividida da seguinte forma: a tarefa da análise; a dimensão do texto; e a técnica de extração de conhecimento, como representado na Figura 1. Como apresentado na Figura 1, a análise de sentimentos pode ter ênfase em diversos tópicos, de modo que é possível considerá-los parcialmente ou em sua totalidade. Cabe ao pesquisador optar por quais subáreas deseja lidar no trabalho.

Figura 1 - Tópicos de pesquisas em AS



Fonte: Adaptado de Yue et al., (2019) e Pereira (2020)

Em relação à orientação de tarefas, o objetivo é analisar a finalidade do estudo. De acordo com Yue et al., 2019, isto pode ser dividido como classificação de polaridade, níveis de valência, dentre outros. A classificação consiste em associar um sentimento ou rótulo a uma amostra ou documento textual, enquanto que os níveis associam um valor de intensidade ao sentimento presente no texto.

Sobre a granularidade, está relacionada à quantidade de texto analisada por amostra, isto é, o NLP pode ser feito em um texto inteiro, em parágrafos, em períodos ou mesmo em

palavras (YUE et al., 2019). A mineração de opinião pode ser feita em diferentes níveis: documento, sentença, palavra e atributo.

O nível de documento considera uma única opinião em um texto como um todo; o nível de sentença determina a opinião em cada frase, o nível de palavra extrai uma opinião de cada palavra; e no nível de atributo, são extraídas entidades do texto como empresas ou pessoas e então se determina a opinião para cada entidade extraída (LIMA; DE CASTRO; CORCHADO, 2015).

A orientação quanto à metodologia é dividida em três módulos: aprendizado de máquina, ou *machine learning* (ML) do inglês, abordagem léxica e abordagem híbrida (YUE et al., 2019). O método de ML supervisionado também é conhecido como classificação em DM. Tal método consiste em utilizar um conjunto de dados de treinamento cujos rótulos ou classes são conhecidos previamente, assim, o método busca identificar o padrão de cada classe e aplicar tal informação para classificar objetos cuja classe é desconhecida (YUE et al., 2019; BALAZS; VELÁSQUEZ, 2016). O método não supervisionado considera que os dados de treinamento não possuem rótulos e o semi-supervisionado que possui dados rotulados e não rotulados.

Segundo Yue et al., (2019), existem diversos tipos de algoritmos de classificação, de modo que os principais paradigmas são: algoritmos baseados em regras; árvores de decisão; máquinas de vetores de suporte ou *support vector machine* (SVM) do inglês; redes neurais artificiais; aprendizado profundo; regressão; entre outros métodos. Apesar do método supervisionado não considerar o custo de construção da base de dados de treinamento rotulados, ele é um método bastante utilizado por pesquisadores em análise de sentimentos (YUE et al., 2019).

Por fim, é possível mesclar ambas as metodologias de análise léxica e ML de forma a criar um método híbrido (LI et al., 2017). A partir disto, os dados textuais podem ser inicialmente submetidos a um processamento léxico que, além de realizar uma limpeza inicial, faça também uma seleção das palavras com maior chance de serem relevantes.

2.3.1 Mineração de opinião em português

Além de apresentar as diferentes perspectivas relacionadas ao modo de como realizar a AS, seja em relação à tarefa, à granularidade ou ao método, é importante destacar como isto se aplica na língua portuguesa. Segundo Cirqueira et al., (2018), o Brasil é a capital universal das mídias sociais, cuja população é a segunda mais ativa em redes sociais e cujo português é

a língua padrão. Devido ao fato de que, aproximadamente, apenas 25% dos usuários da internet usavam o inglês como língua principal em abril de 2019, existe uma lacuna em relação a análise de sentimentos em outras línguas, de modo que estudar outras línguas pode ser vantajoso (PEREIRA, 2020).

Adicionalmente, existem poucos trabalhos e ferramentas que lidam diretamente com a mineração de opinião na língua portuguesa, de modo que a estratégia que gerou melhores resultados foi utilizar traduções de diferentes línguas para o inglês, contudo, esta estratégia está limitada a linguagens com recursos linguísticos limitados (PEREIRA, 2020).

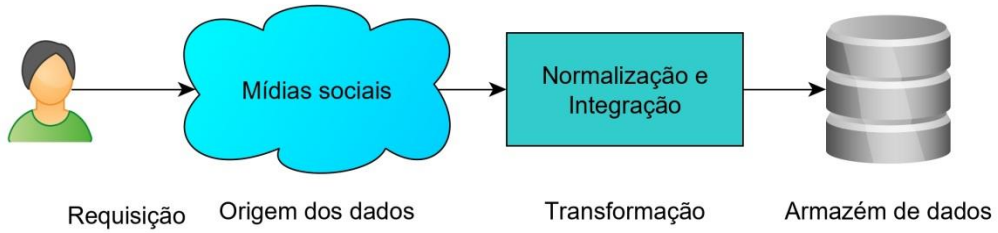
Apesar de o idioma português ser um dos cinco idiomas mais falados na internet, é tratado como uma língua com poucos recursos linguísticos desenvolvidos. Logo, é afirmado por Pereira (2020) que é interessante realizar um pré-processamento nos dados textuais em português antes de convertê-los para o inglês, pois línguas específicas geralmente contêm conhecimento específico que se apresenta na forma de jargões, e expressões idiomáticas locais. Em contrapartida, é importante que a tradução automática dos dados em questão mantenha a mesma polaridade, uma vez que as ferramentas de tradução são bastante competitivas em relação a outras técnicas devido a sua aplicabilidade (ARAÚJO; PEREIRA; BENEVENUTO, 2020).

Neste cenário, a despeito de que a tradução de dados textuais gera uma perda de expressões, este é o método que consegue extrair os melhores resultados, o que reforça a necessidade de desenvolver modelos que consigam realizar um tratamento adequado nos dados. É demonstrado que apenas cerca de 7% das traduções automáticas, como as realizadas por ferramentas como Google Tradutor, acabam por inverter a polaridade do documento, o que implica que o processo de tradução não é ideal, mas é razoável afirmar que este método não gera uma análise de sentimentos contraditória (ARAÚJO; PEREIRA; BENEVENUTO, 2020).

2.4 Extração e transformação de dados

O processo como um todo é conhecido na literatura como processo ETA, que significa extração, transformação e armazenamento (KRAIEM et al., 2019). Tal processo está representado na Figura 2, em que o programador inicialmente realiza requisições específicas de dados, assim, são extraídos das redes sociais e então são submetidos a um processo de transformação. Ao final, os dados devidamente limpos são carregados no armazém de dados, de forma a possibilitar estudos posteriores com estes dados.

Figura 2 - Processo ETA



Fonte: Elaborado pelo autor

Em relação à transformação de dados textuais, segundo Kumar e Garg (2019), é necessário primeiramente integrar os dados e depois aplicar técnicas de limpeza. Tal integração diz respeito ao modo de unificar os dados de diferentes origens com o intuito de se apresentá-los sob uma única ótica, assim os atributos semelhantes podem ser tratados da mesma forma na análise posterior, enquanto os distintos são realocados ou desconsiderados (MOALLA; NABLI; HAMMAMI, 2017).

Neste contexto, as técnicas de limpeza de dados são um conjunto de regras cujo objetivo é refinar os dados, de forma a tornar a sua análise mais eficiente ou mesmo possível. Vários trabalhos utilizam diferentes técnicas em análise de sentimentos para diferentes cenários, assim, as principais técnicas foram apresentadas juntamente aos trabalhos que as empregaram:

a) Remoção de palavras de parada (*stopwords*) – Trata-se de palavras com pouca ou nenhuma relevância que aparecem com frequência nos textos. Palavras como preposições e artigos, por exemplo, “de”, “o”, “as”, são frequentes e não são úteis durante o DM e devem ser removidas. (NAZIR et al., 2019; ROUT et al., 2018; TRIPATHY; AGRAWAL; RATH, 2016);
b) Remoção de “caracteres especiais – É o processo de se remover caracteres como: “?”, “!”, “#”, “%”, entre outros caracteres. Tal processo também retira caracteres de pontuação e acentos (NAZIR et al., 2019; ROUT et al., 2018);
c) Tokenização – Este é o processo de dividir o texto em palavras chamadas “tokens”, isto é feito de modo que cada conjunto de caracteres justapostos é considerado um “token”. Tal método é importante para identificar cada palavra e analisá-la. (CIRQUEIRA et al., 2018). Um método bastante utilizado para realizar tal tarefa é o n-gramas, este método consiste em verificar as n seguintes palavras de um dado texto, de forma que este modelo auxilia a predição do próximo termo esperado após uma determinada palavra aparecer (TRIPATHY; AGRAWAL; RATH, 2016);

d) Remoção de URL – Trata-se do processo de detecção e remoção de links ou URLs (*Uniform Resource Locator*) que significa localizador uniforme de recursos do inglês (NAZIR et al., 2019; ROUT et al., 2018; PANDEY; RAJPOOT; SARASWAT, 2017);
e) Remoção de usuários (menções) – É o processo de remoção de menções a outros usuários no texto (CIRQUEIRA et al., 2018; PANDEY; RAJPOOT; SARASWAT, 2017);
f) Radicalização – É o processo que busca alterar o tamanho das palavras de modo a reduzi-las para o seu radical (NAZIR et al., 2019; CIRQUEIRA et al., 2018);
g) Tratamento de emojis – É o processo de detectar determinados emojis previstos e remover o restante, de modo que eles podem auxiliar na classificação de polaridade (CIRQUEIRA et al., 2018).

Somado a isto, devido ao fato de que os dados são textuais, ainda é preciso fazer com que os dados sejam computacionalmente processáveis, o que é conhecido como extração de características ou seleção de termos (CHANG et al., 2020).

Para tal, existem alguns processos que transformam os dados em vetores de características, isto é, é possível transformar um documento ou uma sentença em um conjunto da forma $v(d) = \{v_1, v_2 \dots v_n\}$ em que $v_i = \{1|0\}$, em que n é o número de palavras ou atributos totais, e o valor 1 representa a ocorrência de uma determinada palavra no documento d na posição i, e o valor 0 representa a ausência (LIU; WANG, 2019).

Segundo Chang et al., (2020), um valor limite pode ser utilizado para remover palavras pouco ou muito frequentes, pois se admite que estas palavras não contribuem para o DM, isto é usado em técnicas como a frequência de documentos, *document frequency* (DF) em inglês, e a frequência do termo-inverso da frequência de documentos, *term frequency–inverse document frequency* (TF-IDF) do inglês. De acordo com estes autores, existem outras técnicas embasadas em estatística como: qui-quadrado, ganho de informações, diferença proporcional categórica, dentre outras.

Apesar da existência de vários métodos, a técnica TF-IDF costuma apresentar os melhores resultados (AHUJA et al., 2019) e é indicada para casos de bases de dados reais (CHANG et al., 2020). Este método transforma o vetor binário de características em um vetor numérico de modo a colocar os pesos nas posições, e cada frase ou documento pode ser

representado desta forma. Com isto, um conjunto de documentos de entrada pode ser convertido na matriz de termos de documentos ou *term-document matrix* (TDM) em inglês.

A frequência de um termo ou TF, termo em inglês, é um cálculo que leva em consideração quantas vezes um termo t ou atributo apareceu no documento d em relação à quantidade de termos totais do documento d, como representado na Equação (1).

A partir disto, é necessário calcular a frequência inversa para medir a importância do termo em questão, conhecido como IDF, isto é calculado por meio da Equação (2) (CHANG et al., 2020). Finalmente, a métrica TF-IDF pode ser calculada como apresentado na Equação (3 , de modo que isto mede a importância geral do termo em todos os documentos levados em consideração (CHANG et al., 2020).

$$TF(t,d) = \frac{número\ de\ vezes\ que\ t\ ocorre\ no\ documento\ d}{número\ total\ de\ termos\ em\ um\ documento\ d} \quad (1)$$

$$IDF(t) = \log_e \frac{número\ total\ de\ documentos}{número\ de\ documentos\ com\ o\ termo\ t} \quad (2)$$

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (3)$$

2.5 Seleção de atributos

Em linhas gerais, a seleção de atributos é um problema que busca selecionar, entre os n atributos existentes em um determinado conjunto, um subconjunto com m atributos de acordo com algum critério, tal que $m < n$.

Isto é feito com o intuito de selecionar os atributos mais relevantes do conjunto original, de forma que se busca remover atributos desnecessários, o que pode fazer com que algoritmos de aprendizado aumentem sua acurácia e diminuam os custos computacionais como tempo de execução (CAI et al., 2018).

Uma estratégia que identifica todos os possíveis subconjuntos, o que significa encontrar a solução ótima eventualmente, requer um algoritmo cuja complexidade está na ordem de $O(2^n)$, o que torna este problema intratável porque se trata de um problema NP-completo (CHANDRASHEKAR; SAHIN, 2014).

2.5.1 Definição formal

A seleção de atributos pode ser entendida como um problema de otimização, cujo objetivo é encontrar o menor, ou melhor, subconjunto de atributos que atinja os menores valores de erro e os maiores valores de acurácia em um processo de classificação.

Sejam os atributos do conjunto de dados original definidos como um vetor $A = \{a_1, a_2, \dots, a_n\}$, e $|A| = n$, em que n é o número de atributos, neste caso, o número de palavras distintas. Após a seleção de atributos, o intuito é encontrar o subconjunto $B = \{b_1, b_2, \dots, b_m\}$, e $|B| = m$, tal que $m < n;\ m, n \in \mathbb{N}$ e $m, n > 0$ (SHARMA; KAUR, 2020).

Assim, a função objetivo, também chamada de função de aptidão ou *fitness* do inglês, pode ser uma combinação linear da forma $Z = F(X) = \sum_{i=0}^{n} x_i\, c_i$, em que $c_i \in \mathbb{Z}$. Esta função pode representar tanto a acurácia quanto o erro de um classificador, assim como utilizado em trabalhos na literatura (KUMAR et al., 2019; RODRIGUES et al., 2013). Formalmente, um problema de otimização pode ser representado como:

$Minimizar$

$$Z = F(X), X = [x_1, x_2, \dots, x_n]^T, X \in \mathbb{R}^n$$

$sujeito\ a$

$$g_j \geq c, \qquad j = 1, 2, \dots, k$$

$$X_i^{(L)} \leq X \leq X_i^{(U)}, \quad i = 1,2, \dots, n$$

O item $F(X)$ representa a função objetivo a ser otimizada e g_j, $X_i^{(L)}$ e $X_i^{(U)}$ são restrições relacionadas ao problema, de tal forma que $c \in \mathbb{R}$ é uma constante e $X_i^{(L)}$ e $X_i^{(U)}$ são valores de limite inferior e superior, respectivamente.

2.5.2 Abordagem determinística

O método de seleção de atributos filtro é aplicado em uma etapa anterior à mineração de dados e independe do algoritmo de DM, neste caso, a abordagem realiza um ranqueamento dos atributos e usa um critério para selecionar os atributos com melhor ranque (CAI et al., 2018). Assim, tal técnica funciona, tipicamente, em duas etapas, a primeira consiste em atribuir uma pontuação a cada atributo, e a segunda consiste em filtrar os atributos com menor pontuação de acordo com um valor limite (LI et al., 2017). Uma limitação é que o subconjunto escolhido por este método pode não ser o ótimo, e um subconjunto redundante pode ser escolhido (CHANDRASHEKAR; SAHIN, 2014).

Em contrapartida, o método embutido é amplamente empregado neste contexto, de modo que a ideia da técnica é incorporar o processo de seleção de atributos como parte da etapa de treinamento de algoritmos de ML, assim os atributos escolhidos são automaticamente extraídos ao fim do processo de treino (CHANDRASHEKAR; SAHIN, 2014; CAI et al., 2018).

Este método pode ser visto como uma tentativa de mesclar as características dos métodos *wrapper* e filtro, de modo que incluem algoritmos de ML e são mais eficientes que os outros métodos pelo motivo de não precisar de uma reavaliação de seu subconjunto a cada iteração (LI et al., 2017).

Adiante, o método *wrapper* ou "empacotado" realiza a seleção por meio do algoritmo de ML, isto é, um dado conjunto de atributos é escolhido e o algoritmo faz a sua avaliação de modo a usar seu erro ou sua acurácia como uma medida de qualidade do conjunto escolhido (CAI et al., 2018). Devido ao fato de existirem 2^n subconjuntos possíveis, torna-se inviável testar todos os subconjuntos. Logo, os melhores subconjuntos são encontrados por meio de estratégias de busca, que geralmente encontram um subconjunto de forma heurística. Neste cenário, estratégias de busca como o algoritmo genético e otimização por enxame de partículas, conhecido pelo nome em inglês *particle swarn optimization* (PSO), se mostram alternativas mais eficientes para este problema (CHANDRASHEKAR; SAHIN, 2014).

Adicionalmente, o método *wrapper*, quando comparado ao método filtro, tende a obter subconjuntos menores e maiores valores de acurácia, porém ele não possui uma boa capacidade de generalização e seu tempo de execução é maior (CAI et al., 2018). Existem, ainda, os métodos híbridos, tais métodos podem ser combinações de duas ou mais técnicas vistas anteriormente, cujo principal objetivo é contornar as desvantagens de cada método e aproveitar as vantagens de cada um, assim, os resultados tendem a ser mais robustos pelo fato de que os atributos selecionados com este método tendem a ter sua relevância garantida por mais de um método isolado (LI et al., 2017).

Segundo Chandrashekar e Sahin, (2014), uma abordagem determinística do método *wrapper* se baseia no algoritmo de seleção sequencial, neste caso, o algoritmo inicia um conjunto de atributos vazio e o preenche com atributos tidos como relevantes, até que uma função de maximização atinge seu valor máximo, e, para manter o tempo de execução aceitável, um critério é utilizado para incrementar a função de maximização com o menor número possível de atributos.

Tal método de seleção de atributos encontra o melhor subconjunto de atributos eventualmente, mas não é viável verificar cada possível combinação. Para resolver este problema, técnicas estocásticas como PSO (AKHTAR et al., 2017), AG (UYSAL, 2018), BC (PANDEY; RAJPOOT; SARASWAT, 2020) e variações que mesclam técnicas (KANAGARAJ; PONNAMBALAM; JAWAHAR, 2013) são utilizadas para se chegar a uma boa solução em um tempo de execução hábil.

Assim, os critérios de avaliação dependem da técnica empregada e devem ser escolhidas para o cenário específico em que está inserida. Quanto ao tipo de busca, é possível escolher entre um tipo que começa com o subconjunto vazio e, à medida que o algoritmo verifica os atributos, eles são adicionados no subconjunto; ou então é possível iniciar o subconjunto todos os atributos e removê-los conforme o algoritmo executa. Tais buscas são conhecidas, em inglês, como *increase foward* e *backward deletion*, respectivamente (CAI et al., 2018).

2.5.3 Abordagem estocástica

Com o intuito de encontrar boas soluções, neste caso bons subconjuntos de atributos, várias abordagens meta-heurísticas foram desenvolvidas nos últimos anos, de modo que cada abordagem possui vantagens e desvantagens (SHARMA; KAUR, 2020). Tais estratégias buscam identificar uma boa solução, de forma a percorrer eficientemente o espaço de busca e convergir para uma boa, eventualmente, a melhor solução de modo suficientemente rápido (YADAV; VISHWAKARMA, 2020).

Em particular, a análise de sentimentos propiciou um cenário interessante para tais abordagens, segundo Yadav e Vishwakarma, (2020), o número de publicações de trabalhos voltados para an com métodos heurísticos está em crescimento, o que indica que é uma área vasta que demanda vários estudos para lidar com os diferentes desafios que surgem atualmente.

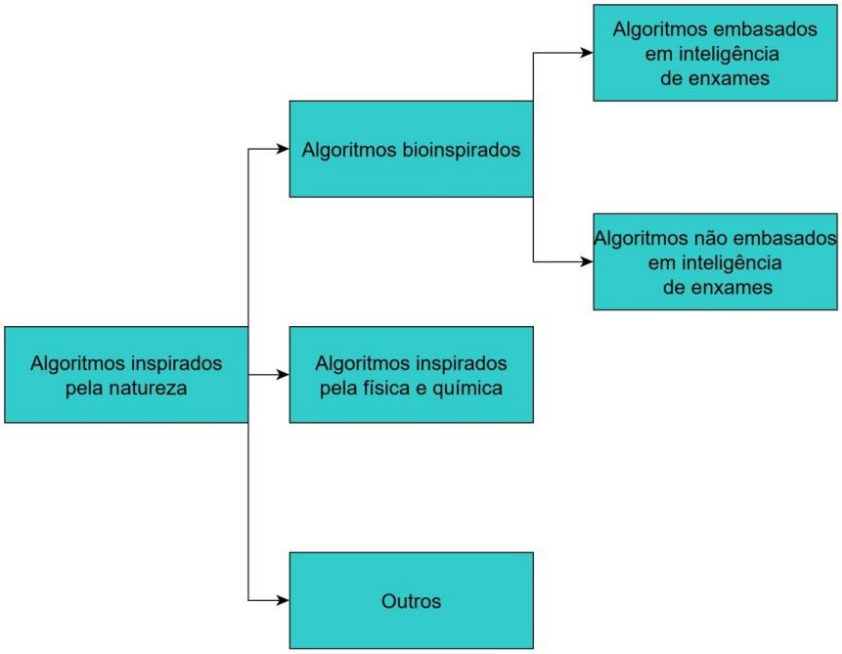
Com isto, os maiores desafios nesta área consistem em encontrar a melhor forma de lidar com a alta dimensionalidade dos dados, com a relevância dos atributos e com a sua redundância (SHARMA; KUMAR, 2020), assim como ainda são necessários esforços para tratar os casos pouco explorados como análise embasada em contexto dos dados e outras línguas além da inglesa (YADAV; VISHWAKARMA, 2020).

Os tipos de algoritmos inspirados pela natureza podem ser visualizados na Figura 3. A computação inspirada pela natureza é composta por várias outras subáreas mais específicas, de modo que a variação da origem da estratégia determina sua subárea.

Neste aspecto, existem os algoritmos heurísticos que usam a estratégia de tentativa e erro e tais métodos buscam encontrar boas soluções em um curto período de tempo, ao passo que os algoritmos meta-heurísticos são um conjunto de estratégias de alto nível genéricas, elas buscam a eficiência de algoritmos heurísticos ao guiar e modificar suas operações para que para que atinjam os melhores resultados, além de que podem ser aplicados em vários

contextos sem um conhecimento prévio do problema, pois se tratam de métodos que resolvem problemas de otimização (YADAV; VISHWAKARMA, 2020).

Figura 3 - Diagrama de algoritmos inspirados pela natureza



Fonte: Adaptado de Yadav e Vishwakarma, (2020)

Tais algoritmos podem ser divididos de acordo com suas premissas que o inspiraram, como mostrado na Figura 3, de tal forma que existem os algoritmos: bioinspirados que são embasados em processos biológicos geralmente associados a seres vivos; estes algoritmos podem ainda ser subdivididos em algoritmos que utilizam a inteligência de exames como o PSO, BC, otimização por colônia de formigas (ACO), BC, colônia de abelhas, libélulas, dentre outros (YADAV; VISHWAKARMA, 2020).

Por outro lado, existem os algoritmos que não se baseiam em inteligência de enxames e algoritmos genéticos que são inspirados nos genes e como eles se reproduzem. Ainda, existem algoritmos embasados em processos físicos e químicos, como a têmpera simulada, busca harmônica e busca gravitacional. Por fim, existem técnicas que são inspiradas em outros processos da natureza como redes neurais, sistemas imunes e evolução diferencial (YADAV; VISHWAKARMA, 2020).

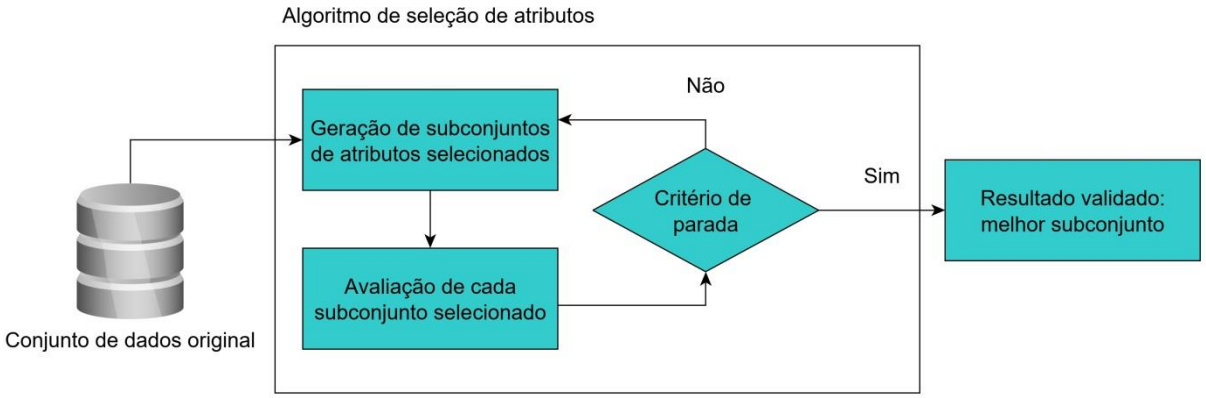
É válido ressaltar que os algoritmos bioinspirados fazem parte de um conjunto que é conhecido como computação evolutiva, tal nome é dado porque há um paradigma nos algoritmos de que, a cada iteração ou época, os chamados indivíduos ou possíveis soluções são avaliados e então os indivíduos evoluem, de modo que eles sofrem alterações de uma iteração para a outra (LIU; WANG, 2019).

A partir disto, é possível explicitar como tais algoritmos podem funcionar e como podem ser aplicados para resolver os desafios da seleção de atributos. Mesmo quando se trata dos métodos bioinspirados, ainda podem ser aplicados nas categorias de métodos de seleção por filtro, *wrapper*, e embutidos (LIU; WANG, 2019).

A ideia de funcionamento básica destes métodos pode ser visualizada por meio da Figura 4 e opera da seguinte forma: dado um conjunto de dados completos ou conjunto original, são gerados subconjuntos de soluções, por vezes de formas aleatórias ou com base em sua heurística, e assim, cada subconjunto é avaliado e os melhores são escolhidos.

Tal processo se repete até que um critério seja satisfeito, então o melhor subconjunto encontrado é apresentado como saída, não é necessariamente o melhor subconjunto e tampouco o algoritmo heurístico obrigatoriamente retorna o mesmo subconjunto a cada execução, porém é garantido que tal subconjunto seja encontrado em tempo hábil e que seja pelo menos uma boa solução desde que o algoritmo esteja corretamente configurado.

Figura 4 - Fluxograma do processo de seleção de atributos



Fonte: Adaptado de Sharma; Kaur, (2020)

No caso dos métodos *wrapper*, subconjuntos são gerados e então cada um é avaliado, de forma que os melhores são mantidos e os piores são substituídos. Para garantir uma boa exploração do espaço de busca, cada algoritmo heurístico se utiliza de algum paradigma para explorá-lo de forma satisfatória e obter melhores subconjuntos com base nos subconjuntos anteriores (LIU; WANG, 2019). Neste contexto, a função de avaliação pode levar em conta tanto a acurácia obtida quanto o tamanho do subconjunto.

Por outro lado, métodos filtros utilizam processos que independem do algoritmo de aprendizado de máquina, mas dependem apenas do conjunto original em si, de modo que relevância e redundância são identificadas por funções específicas. Com isto, a ideia principal para uma boa seleção está no critério de seleção adequado para um tipo de dado específico

(LIU; WANG, 2019). Em contrapartida, os métodos embutidos não são muito explorados em conjunto com algoritmos meta-heurísticos, enquanto que métodos híbridos são mais comuns e buscam mesclar os métodos de seleção citados (LIU; WANG, 2019).

Após escolher um método de seleção e um algoritmo meta-heurístico, é necessário adaptá-lo para lidar com o problema de seleção de atributos, isto ocorre porque são técnicas desenvolvidas para resolver problemas contínuos e, portanto, precisam ser convertidos para problemas em que as soluções, em geral, são vetores da forma $v \in \{0,1\}^n$, em que n é o tamanho da dimensão e cada posição do vetor é uma representação de cada atributo (LIU; WANG, 2019).

Apesar das vantagens relacionadas ao desempenho de algoritmos meta-heurísticos, é preciso destacar que tais algoritmos não contornaram o problema de forma totalmente satisfatória. No caso dos métodos *wrapper*, isto ocorre porque a função de aptidão é chamada intensamente para avaliar cada indivíduo a cada iteração, logo isto é outro ponto desafiador neste contexto (LIU; WANG, 2019). Existem vários desafios em aberto no cenário da AS, dentre eles é interessante destacar o fato de que a maioria dos trabalhos aborda o problema da classificação com apenas duas classes, trabalha apenas com a língua inglesa e não leva o contexto dos dados em consideração (YADAV; VISHWAKARMA, 2020).

2.6 Busca cuco

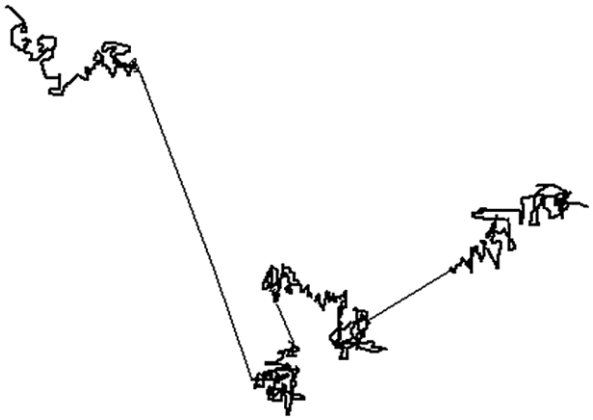
Segundo Sharma e Kaur (2020), o número de artigos relacionados a métodos meta-heurísticos para seleção de atributos está em ascensão desde 2010 e a busca cuco é o segundo método mais empregado em seleção de atributos. De acordo com Yadav e Vishwakarma (2020), a BC é um dos métodos mais usados no cenário de AS, além de ter sido aplicado em análise de sentimentos com dados em português e inglês, somado ao fato de que foi capaz de realizar uma classificação de sentimentos efetiva em diferentes bases de dados.

Tal algoritmo foi proposto por Yang e Deb (2009) e se trata de um método meta-heurístico inspirado no comportamento de reprodução das aves cuco. A metáfora deste método consiste no seguinte comportamento: certas espécies da ave cuco possuem um comportamento peculiar em relação ao modo como se reproduzem, praticam o parasitismo de ninho. Tal estratégia consiste em colocar os ovos do cuco em ninhos de outras espécies na expectativa de que o filhote de cuco seja alimentado por outras aves. (DE MOURA MENESES et al., 2020).

Adicionalmente, os autores Yang e Deb (2009) também utilizaram um conceito importante chamado voos de Lévy (VL), tal técnica faz parte do núcleo do algoritmo BC e se trata um método de busca aleatória de soluções eficiente, de forma que consegue percorrer boa parte do espaço de busca em poucas iterações.

Tal estratégia de busca em duas dimensões pode ser visualizada por meio da Figura 5, em que a exploração é feita em um espaço de duas dimensões de forma que, com poucas iterações, é possível percorrer uma grande região e explorar soluções de modo mais eficiente do que outros métodos (YANG; DEB, 2009). Na Figura 5, uma caminhada aleatória é exemplificada por meio dos VL, que consiste em, dado um ponto inicial, identificar boas soluções na região próxima; posteriormente, há uma probabilidade de se dar um salto para uma região distante e reiniciar o processo. O VL não considera a qualidade dos pontos. Além disso, o tamanho de uma caminhada é atingido por meio da distribuição Lévy descrita na Equação (4, que possui variância e média infinita (PANDEY; RAJPOOT, 2019).

Figura 5 - Demonstração do VL em duas dimensões



Fonte: Pandey e Rajpoot (2019)

$$\text{Lévy} \sim u = t^{-\lambda},\ (1 < \lambda \leq 3) \tag{4}$$

O algoritmo faz uso de uma função de avaliação ou função de aptidão, cujo objetivo pode ser minimizá-la ou maximizá-la; cada iteração no laço de repetição principal do algoritmo é chamada de geração; cada solução ou indivíduo da população é considerado um ninho e cada ninho contém uma solução. Logo, cada ninho contém um subconjunto de atributos, e os ovos do cuco representam uma nova solução, que pode ser aceita como uma nova solução ou pode ser descartada.

Neste contexto, as regras são abstrações do modelo real e são utilizadas como estratégia de busca e de melhoria das soluções. De acordo com Yang e Deb (2009), as três regras básicas são:

a) Uma ave cuco pode colocar seus ovos em qualquer ninho e que escolhe apenas um ninho de forma aleatória em cada geração;
b) Os melhores ninhos com os melhores ovos, isto é, as melhores soluções são selecionadas para serem mantidas para a próxima geração;
c) A quantidade de ninhos hospedeiros é fixa (tamanho da população) e o ovo colocado em um ninho por um cuco pode ser descoberto pela ave hospedeira com probabilidade $P_a \in [0,1]$.

Quando um ovo de cuco é descoberto, a ave hospedeira pode descartar o ovo do cuco, isto é, descartar a nova solução ou simplesmente abandonar o ninho e construir outro. Com isto, a fração P_a dos piores ninhos é substituída por novas soluções aleatórias e o restante da população é mantido. O pseudocódigo do algoritmo é representado por meio da Figura 6.

Figura 6 - Pseudocódigo do algoritmo BC

Início
Função de aptidão $f(X), X = (x_1, \dots, x_n)^T$
Gere a população inicial de n ninhos hospedeiros x_i $(i = 1,2, \dots, n)$
Enquanto (t<MaxGerações) ou (Critério)
Selecione um cuco aleatoriamente por VL e avalie sua aptidão F_i
Escolha um ninho j aleatório entre os n possíveis
Se ($F_i > F_j$)
Substitua a solução do ninho j por i
Fim_se
Uma fração (P_a) dos piores ninhos é abandonada e novas soluções são geradas
Manter as melhores soluções ou os melhores ninhos
Ordene as soluções e salve a melhor solução
Fim_enquanto
Exibir a melhor solução
Fim

Fonte: Adaptado de Yang e Deb (2009)

O método é iniciado com os seguintes parâmetros: uma função de aptidão a ser maximizada; o tamanho da população de ninhos; o número de gerações máximo; a proporção de ninhos a serem descartados P_a; e as constantes que influenciam no tamanho do salto α e β.

Após isto, uma população de ninhos é iniciada aleatoriamente e então o laço principal é iniciado. Até que um critério de parada seja atingido, uma nova solução é descoberta por VL e então um ninho aleatório é selecionado, se a solução descoberta for melhor, então a troca é feita, caso contrário, a população se mantém. Ao final, as soluções são ordenadas de modo

que uma parte das melhores soluções sejam mantidas, enquanto que as piores soluções restantes são descartadas e novas soluções aleatórias são colocadas em seus lugares.

Desta forma, quando uma nova solução for gerada para a iteração seguinte, isto é, $t + 1$, basta calcular qual é o tamanho do passo a ser somado como presente na Equação (5), em que $\oplus$ representa operações de multiplicação termo a termo. Além disso, α é um número real tal que $\alpha > 0$ e geralmente ele é configurado para $\alpha = 1$ (YANG; DEB, 2009; DE MOURA MENESES et al., 2020).

É interessante destacar que a técnica é bastante simples se comparada a outros métodos, pois precisa apenas dos seguintes parâmetros: tamanho da população; número de gerações; número de dimensões; probabilidade (P_a) e um fator de escala α. De acordo com De Moura Meneses et al., (2020), é possível especificar as operações envolvidas no cálculo do passo da busca, de modo que utilizaremos β como o índice de Lévy, em que o valor de β é obtido com a Equação (6).

Yang (2014) fornece os detalhes sobre as operações do algoritmo BC, de forma que o algoritmo utiliza uma combinação balanceada de buscas locais e globais, em que a busca local é ajustada pelo parâmetro P_a e está representada na Equação (7). Nesta equação, $x_i^{(t)}$ é o i-ésimo candidato a ser substituído na iteração t, α_1 é o fator de escala geralmente ajustado para 1, s é o tamanho do passo e é um valor real entre 0 e 1, $H(.)$ é a função degrau, ϵ é um valor real obtido a partir da distribuição uniforme entre 0 e 1, x_j^t e x_k^t são duas soluções selecionadas de forma aleatória, e o símbolo $\otimes$ representa multiplicações termo a termo entre os vetores com as mesmas dimensões.

Adiante, Yang (2014) apresenta a busca global que está representada pela Equação (8), em que α_2 é um valor real geralmente atribuído ao valor 0,01, de modo que a busca não dê saltos muito distantes. O número $L(\beta)$ é obtido por meio da Equação (4), e representa o valor de perturbação para a nova solução, em que β é configurado para 1. O fator $L(\beta)$ pode ser calculado de diversas formas em termos computacionais, porém o modo mais simples de calculá-lo é por meio do algoritmo de Mantegna (YANG, 2014), assim sua estabilidade e simetria são mantidas e ele pode ser calculado por meio da Equação (9).

Assim, para calcular tal perturbação são utilizados dois vetores u e v, e a diferença entre a solução atual e a melhor encontrada até a iteração t. Os vetores u e v são obtidos por meio da distribuição normal apresentada nas Equações (10) e (11), respectivamente. Por fim, σ_v é geralmente atribuído ao valor 1 e σ_u é calculado por meio da Equação (12), em que $\Gamma(.)$ é a função Gamma padrão (DE MOURA MENESES et al., 2020).

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Lévy(\lambda) \tag{5}$$

$$\beta = \lambda - 1 \tag{6}$$

$$x_i^{(t+1)} = x_i^{(t)} + \alpha_1 . s \otimes H(P_a - \epsilon) \otimes (x_j^t - x_k^t) \tag{7}$$

$$x_i^{(t+1)} = x_i^{(t)} + \alpha_2 \otimes L(\beta) \tag{8}$$

$$L(\beta) \sim \frac{u}{|v|^{\frac{1}{\beta}}} (x_{melhor}^t - x_i^t) \tag{9}$$

$$u \sim N(0, \sigma_u^2) \tag{10}$$

$$v \sim N(0, \sigma_v^2) \tag{11}$$

$$\sigma_u = \left\{ \frac{\Gamma(1+\beta) . sen(\frac{\beta\pi}{2})}{\Gamma\left(\frac{1+\beta}{2}\right) . \beta . 2^{\frac{1-\beta}{2}}} \right\}^{\frac{1}{\beta}} \tag{12}$$

O algoritmo BC usa elitismo e se baseia na ideia de que a nova solução trazida pelo cuco seja melhor. Eventualmente, todos os ninhos são visitados pelo cuco e então, eventualmente todos os ninhos conseguem evoluir para que tenham ovos melhores. Somado a isto, o seu processo de randomização é mais eficiente e explora o espaço de busca de forma equilibrada, em que é explorado em pontos próximos e em pontos distantes, além de ter menos parâmetros de entrada (YANG; DEB, 2009).

Adicionalmente, tais vantagens explicadas anteriormente ainda são combinadas com o fato de o algoritmo ser potencialmente mais genérico, estável e adaptável a problemas do que algoritmos como PSO e AG, por exemplo. Foi verificado que o algoritmo BC não precisa de um ajuste fino nos parâmetros, assim, o algoritmo é bastante estável dados valores suficientes para os parâmetros de entrada (YANG; DEB, 2009). Com isto, tal algoritmo pode atingir um patamar interessante quanto à exploração e à explotação, isto é, consegue explorar bem o espaço de busca e consegue encontrar uma boa solução eventualmente (YANG, 2014).

No entanto, algumas limitações deste método apontadas na literatura estão fundamentadas no fato de que, apesar de ser um método que converge rapidamente e utiliza poucos recursos computacionais, o algoritmos BC pode ficar preso em pontos ótimos locais em problemas de otimização (PANDEY; RAJPOOT, 2019; YADAV; VISHWAKARMA, 2020), além de haver problemas de conversão prematura (PANDEY; RAJPOOT; SARASWAT, 2019) ou tardia (ABD EL AZIZ; HASSANIEN, 2018) devido ao fato de apenas uma parte da população é atualizada a cada geração.

2.7 Algoritmo genético

O algoritmo genético se trata de uma heurística bastante utilizada em ciência da computação para resolver diversos problemas, desde aprendizado de máquina até problemas de otimização. A sua principal ideia é gerar soluções e fazer com que melhorem de forma a se diversificar. Similarmente ao algoritmo BC visto anteriormente, um de seus fundamentos é a evolução de soluções, assim busca-se mimetizar o comportamento da natureza de modo que as melhores soluções ou mais bem adaptadas sejam selecionados para se reproduzir, e com isto, as soluções boas surgem até se estabilizarem (KRAMER, 2017).

O AG básico consiste em gerar uma população de indivíduos ou soluções e aplicar três operadores genéticos: seleção; reprodução e mutação, e utiliza conceitos de genética para representar suas operações (YANG, 2017). Neste cenário, o indivíduo pode ser representado como um vetor binário, assim, é chamado de genótipo ou cromossomo, em que cada caractere representa um gene ou característica. O conjunto destes genes forma um fenótipo, ou uma característica física que pode ser interpretada como a aptidão (KRAMER, 2017).

Deste modo, os operadores genéticos são responsáveis por produzir novas soluções no espaço de busca, logo, a caminhada neste espaço é feita por meio deles. De acordo com Kramer (2017), o primeiro operador aplicado é a seleção, ela consiste em selecionar alguns indivíduos da população de forma que alguns podem ser selecionados mais de uma vez, enquanto que outros não são escolhidos. Uma das formas de se fazer a seleção é por meio de torneio, neste caso, um número x de indivíduos escolhidos aleatoriamente é colocado em um torneio e apenas a solução com melhor aptidão é selecionada.

Outra forma de seleção é feita por meio de roleta, em que a aptidão de cada indivíduo é colocada em uma “roleta”, de modo que quanto maior a aptidão, maior a área ou intervalo do indivíduo na roleta, assim, a roleta é rodada n vezes até gerar toda uma nova população (KRAMER, 2017). Ainda, pode haver elitismo na seleção, de modo que o melhor indivíduo é obrigatoriamente selecionado para permanecer na próxima geração.

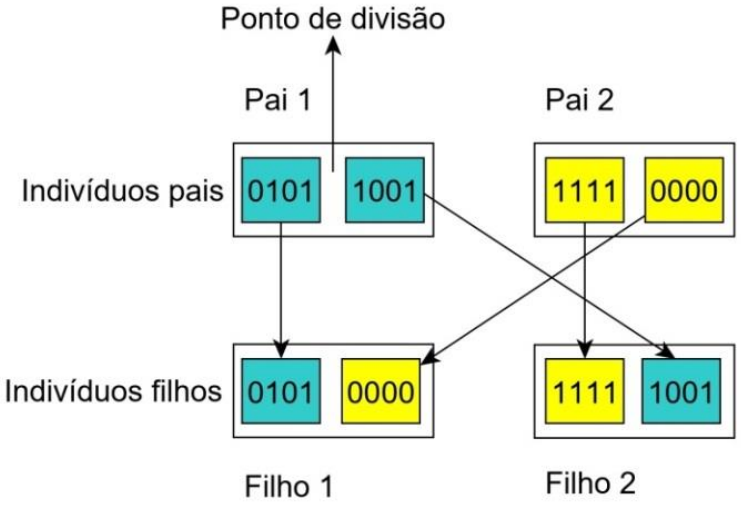
Após isto, a próxima etapa é a reprodução. Tal operador permite a combinação de materiais genéticos de dois ou mais indivíduos. Dois indivíduos podem compor outros dois indivíduos, de modo que os filhos possuem parte do material genético de ambos em igual proporção. Esta operação ocorre de acordo com um parâmetro inicial chamado taxa de cruzamento. Por exemplo, se a reprodução ocorrer com apenas um ponto de divisão, então cada indivíduo pai é cortado em uma determinada localização, tal como, entre a posição 4 e 5, somado a isto, cada parte compõe os dois filhos, tal situação está ilustrada na Figura 7, em

que parcelas que compõem os indivíduos pais são movidas para os indivíduos filhos. O tamanho da parcela é definido pelo ponto de divisão. A ideia desta operação é que partes de duas soluções razoavelmente boas se recombinem e criem soluções melhores (KRAMER, 2017).

Posteriormente, é possível aplicar outro operador denominado mutação, ele atua na solução de modo a perturbá-la, isto é, faz pequenas alterações para explorar o espaço de busca nas regiões próximas de cada indivíduo, assim se espera que descobrir uma eventual melhor solução na região. Com o intuito de ser uma operação pouco custosa, é feita de forma aleatória e a intensidade da perturbação é chamada taxa de mutação (KRAMER, 2017).

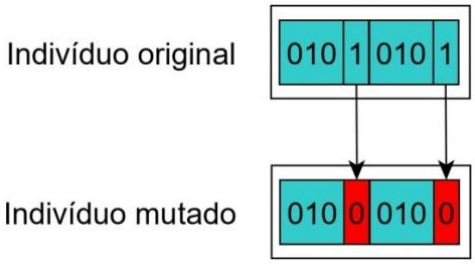
Para que a mutação ocorra, é preciso que cada ponto do espaço de busca seja alcançável. Além disso, o operador deve ser imparcial, logo, não deve fazer com que as soluções convirjam para apenas uma direção ou subespaço no espaço de busca original, a não ser que se tenha um conhecimento prévio sobre uma região. Por fim, a mutação deve ser escalável, e deve permitir que graus de liberdade adaptáveis, de forma que as operações de mutação geralmente ocorrem com base em uma distribuição de probabilidade, assim, a mutação pode seguir uma distribuição normal ou, neste caso, cada um dos n cromossomos pode ser ter seu valor alterado com uma probabilidade de $\frac{1}{n}$. Por exemplo, uma mutação em gene com oito cromossomos pode ocorrer nos genes 4 e 8, como exibido na Figura 8, em que o indivíduo original sofre o processo de mutação ao trocar o valor de dois genes pelo oposto.

Figura 7 - Operador genético de reprodução



Fonte: Adaptado de Kramer (2017)

Figura 8 - Operador genético de mutação



Fonte: Elaborado pelo autor

Um pseudocódigo do AG está representado na Figura 9 e nele as probabilidades de cruzamento e mutação são representadas por P_c e P_m, respectivamente. Inicialmente, a população é criada aleatoriamente e cada indivíduo é avaliado, então uma técnica de seleção é aplicada na população. Após isto, os indivíduos selecionados são cruzados e sofrem mutação de acordo com suas probabilidades, com isto, a população é reavaliada e o melhor indivíduo é salvo e inserido na população selecionada, e isto ocorre até que o critério de parada seja alcançado.

Tal algoritmo é amplamente utilizado no cenário de AS, em especial, ele é utilizado em conjunto com outros métodos (YADAV; VISHWAKARMA, 2020). Além disso, é sabido que tal algoritmo é relativamente mais custoso em termos computacionais se comparado aos outros métodos (SHARMA; KAUR, 2020). Por conta disto, autores geralmente buscam mesclar técnicas com o AG ou outros métodos de forma a unir o melhor das técnicas. Neste caso, a metodologia padronizada de operadores genéticos somados ao método BC pode ser interessante (KANAGARAJ; PONNAMBALAM; JAWAHAR, 2013).

Figura 9 - Pseudocódigo do AG

Início
Iniciar a população P
Avaliar a população e calcular a aptidão de cada indivíduo
Aplicar SELEÇÃO em P
Enquanto (t<MaxGerações) ou (Critério)
Aplicar CRUZAMENTO em P com P_c
Aplicar MUTAÇÃO em P com P_m
Avaliar a população e calcular a aptidão de cada indivíduo
Salvar o melhor indivíduo em I_{melhor}
Aplicar SELEÇÃO em P
Inserir I_{melhor} em P
Fim_enquanto
Exibir a melhor solução
Fim

Fonte: Adaptado de De Castro (2006)

Aplicar uma técnica que combine a BC e o AG pode gerar resultados interessantes para o contexto de análise de sentimentos em mídias sociais, pois, de acordo com Kanagaraj, Ponnambalam e Jawahar (2013), mesclar ambos os algoritmos podem gerar as seguintes características: método simples de se desenvolver e alterar; equilíbrio entre exploração e explotação, uma vez que parte das soluções busca o melhor global com base em boas soluções atuais, enquanto outra parte faz buscas com soluções ruins que podem se tornar boas futuramente; e deve ser apresentar um custo computacional competitivo se comparado a métodos similares.

2.8 Trabalhos correlatos

Os principais trabalhos relacionados encontrados na literatura são apresentados, de modo a destacar a seleção de atributos por meio de abordagens heurísticas inspiradas pela natureza na área de análise de sentimentos em dados de mídias sociais.

2.8.1 Um método de duas etapas para análise de sentimentos embasada em aspectos

Inicialmente, Akhtar et al., (2017) propõem um trabalho que se divide em dois pontos: o primeiro faz uma seleção de atributos enquanto o segundo realiza a construção de um conjunto que combina as saídas de três classificadores. Ambas as etapas utilizaram o algoritmo PSO tanto para selecionar os atributos quanto para construir o conjunto final. Ainda, os autores fizeram esta AS em nível de aspecto, isto é, é necessário extrair os aspectos ou palavras de cada amostra e analisá-los previamente para então submeter os dados ao processo de classificação.

Desta forma, o trabalho utiliza tanto o conjunto de dados de treinamento quanto os de teste e gera os melhores subconjuntos de atributos com a técnica PSO somada a um classificador para avaliar o conjunto. Estes classificadores são: campos aleatórios condicionais, cuja sigla mais conhecida é referenciada em inglês – CRF; máquinas de vetores de suporte e entropia máxima, cuja sigla é referenciada como ME. Para cada classificador, um grupo dos melhores subconjuntos de atributos encontrados é então dado como entrada para outro algoritmo PSO, de modo que é responsável por selecionar as melhores saídas de cada classificador. A partir disto, os autores utilizaram a base SemEval-2014, de modo que duas bases menores fossem extraídas. Tais bases possuem 3044 e 3045 amostras de treinamento, respectivamente, e 800 amostras de teste cada.

Neste estudo, foram considerados três rótulos para a classificação. Posteriormente, o trabalho exibe os resultados em relação a outras técnicas tradicionais de seleção de atributos, em que os valores de acurácia foram todos maiores para as técnicas desenvolvidas. Houve um aumento de aproximadamente 16% e 24% nos valores de acurácia devido ao método proposto em relação às bases de dados "restaurante" e "*laptop*", ambas obtidas a partir da base SemEval-2014.

2.8.2 Desafios no desenvolvimento de abordagens híbridas para análise de sentimentos

Em contrapartida, Appel et al., (2018) realiza um estudo sobre os sucessos e desafios na elaboração de técnicas em AS. Em particular, os autores trabalham principalmente com o desenvolvimento de uma abordagem híbrida que mescle análise léxica, regras semânticas, tratamento de negação e ambiguidade e variáveis linguísticas. Tal trabalho foi aplicado em duas bases de dados, uma de críticas de filmes e outras com *tweets* com sentimentos.

Neste caso, o estudo teve um foco em analisar os dados com uma maior granularidade, pois a análise foi feita em nível de sentença, e os testes foram feitos com a consideração de apenas duas classes, a positiva e a negativa. Os resultados atingidos por meio da abordagem híbrida foram comparados aos classificadores: Bayesiano ingênuo, conhecido pela sigla NB, e ME. Com isto, foi possível verificar que a abordagem proposta atingiu valores de acurácia e precisão maiores do que os classificadores.

Resultados mostram que uma abordagem híbrida que junte as regras semânticas e técnicas de processamento de linguagem natural pode ser mais vantajosa do que utilizar técnicas isoladamente, além de que o trabalho apresentou um aumento de aproximadamente 7% e 21% nos valores de acurácia em relação às bases de filmes e de *tweets*, respectivamente.

2.8.3 Análise de sentimentos com busca cuco para seleção de atributos em *tweets*

Da mesma maneira, o trabalho de Kumar et al., (2019) atua diretamente no problema de selecionar atributos de mídias sociais, em especial, o foco do trabalho foram os dados do *Twitter*. Neste contexto, a busca cuco binária foi aplicada em conjunto com a técnica TF-IDF e os classificadores utilizados para identificar os rótulos positivos e negativos foram: Bayesiano ingênuo; árvore de decisão; SVM; K-NN e perceptron multicamadas (MLP).

Além disso, o trabalho apresenta o pseudocódigo utilizado para o algoritmo proposto e realiza uma limpeza prévia nos dados. A limpeza aplicada inclui: conversão para letras minúsculas; remoção de caracteres repetidos e desnecessários; remoção de *hashtags* e usuários; remoção de pontuação e radicalização. Os dados utilizados são públicos e estão na língua inglesa. Os experimentos conduzidos revelaram um ganho de acurácia de 7,45% em média, enquanto que a quantidade de atributos selecionados foi de 53,17%. O melhor resultado apresentou um aumento de até 9,15% no valor de acurácia.

2.8.4 Seleção de atributos por meio de inteligência de enxames

O trabalho de Kumar e Jaiswal (2019) tem o objetivo de melhorar a acurácia na classificação de sentimentos por meio da seleção de atributos. Este trabalho se utilizou dos algoritmos lobo cinzento binário e mariposa binária para selecionar um subconjunto de atributos com o intuito de melhorar a acurácia da classificação de sentimentos em dados do *Twitter*, em particular, foram usadas duas bases de dados, a SemEval 2016[1] e SemEval 2017[2], de forma que três classes foram utilizadas na classificação, elas são: positivas, negativas e neutras. Para realizar a classificação foram utilizados os algoritmos: NB; SVM; K vizinhos próximos; perceptron multicamadas e árvore de decisão. Após realizar um pré-processamento para retirar os ruídos dos dados, os autores extraíram os atributos por meio da técnica de frequência do termo-inverso da frequência nos documentos, cuja sigla em inglês é TF-IDF.

Os resultados mostram que os algoritmos de otimização simplificam o processo de seleção de atributos de forma efetiva, além de propiciarem maiores valores de acurácia, de modo que o número de atributos selecionados foi, em média, 30% menor, enquanto que houve um aumento de 10% no valor da acurácia em ambas as bases de dados. Ao final, o algoritmo da mariposa binária se mostrou mais eficaz em encontrar subconjuntos menores ao mesmo tempo em que o valor da acurácia aumentava em relação ao algoritmo do lobo cinzento binário.

2.8.5 Filtro híbrido eficiente e abordagem evolutiva para análise de sentimentos

Por fim, Hassonah et al., 2020 elaboraram um artigo sobre uma técnica híbrida de aprendizado de máquina, de forma que os métodos filtro e *wrapper* são combinados.

[1] http://alt.qcri.org/semeval2016/
[2] http://alt.qcri.org/semeval2017/

Tal técnica realiza uma classificação ternária de sentimentos positivos, negativos e neutros por meio do algoritmo SVM, NB, árvore de decisão e floresta aleatória (RF), além de combinar duas abordagens de seleção de atributos, elas são a *ReliefF*, que computa a importância dos atributos em relação à distância que possuem de suas instâncias, e o algoritmo inspirado pela natureza otimização por multiverso, ou, do inglês, *Multi-Verse Optimizer* (MVO).

Adicionalmente, o trabalho também extraiu amostras da mídia social *Twitter* e dividiu os dados em relação ao contexto em que estavam inseridos. Com isto, os resultados obtidos revelaram melhores valores de acurácia em comparação a outras técnicas, tais valores variam de 1 a 15% de aumento no valor de acurácia, assim como houve uma redução no número de atributos de até 96,85%.

2.8.6 Comparação com trabalhos correlatos

Nesta seção, são destacadas as características de cada trabalho correlato e ilustradas cada utilização ou ausência de fatores, tais destaques podem ser visualizados por meio da Tabela 1. É possível perceber que todos os trabalhos realizam um pré-processamento e utilizam aprendizado de máquina em um cenário de mineração de opinião.

Tabela 1 - Comparativo entre trabalhos correlatos

	(AKHTAR et al., 2017)	(APPEL et al., 2018)	(KUMAR et al., 2019)	(KUMAR; JAISWAL et al., 2019)	(HASSONAH et al., 2020)
Uso de ML	Sim	Sim	Sim	Sim	Sim
Uso de meta-heurística	Sim	Não	Sim	Sim	Sim
Pré-processamento	Sim	Sim	Sim	Sim	Sim
Cenário de mídias sociais	Não	Sim	Sim	Sim	Sim
Classificação ternária	Sim	Não	Não	Sim	Sim
Tratamento de contexto	Não	Não	Não	Não	Sim
Tratamento de diferentes idiomas	Não	Não	Não	Não	Não

Fonte: Elaborado pelo autor

Contudo, é visível que existem limitações quanto à hibridização de meta-heurísticas e quanto a realizar classificação com três classes, e, sobretudo, os artigos não levam em consideração diferentes idiomas na análise de sentimentos e somente um deles usa o contexto dos dados como uma forma de melhorar a análise.

2.9 Considerações finais

Este capítulo apresentou os principais conceitos relacionados ao *Big Data*, análise de sentimentos, seleção de atributos e características de dois algoritmos meta-heurísticos, além de uma análise do atual estado da arte quanto a estes fundamentos. Nota-se, então, uma tendência de trabalhos que buscam realizar análise de sentimentos específicas com técnicas bioinspiradas.

Capítulo 3

Abordagem híbrida bioinspirada

Neste capítulo são apresentados os aspectos utilizados para compor a abordagem híbrida proposta, assim como a metodologia e os detalhes de implementação utilizados. A primeira parte da abordagem consiste em selecionar os dados e aplicar o pré-processamento adequado, somado a isto, é necessário levar em consideração os dois idiomas, neste caso português e inglês, e os diferentes contextos aos quais os dados estão associados.

Em seguida, os dados devidamente formatados podem passar pelo processo de seleção ou redução de atributos. Esta etapa foi realizada por meio de um algoritmo inspirado pela natureza que mescla as técnicas de AG e BC, ambas bastante utilizadas neste cenário (SHARMA; KAUR, 2020), de modo que as desvantagens individuais de cada método são atenuadas pelas características de cada abordagem.

3.1 Escopo do algoritmo

É válido destacar que o objetivo deste trabalho é viabilizar e auxiliar a análise de sentimentos no sentido de reduzir a quantidade de atributos elevada que o ambiente de mídias sociais apresenta, e com isto, melhorar a qualidade da classificação de polaridade.

Isto significa que quando os dados são submetidos a algoritmos que não estão devidamente preparados para esta alta dimensionalidade, os modelos tendem a ficar em superposição, o que ocasiona uma perda de desempenho em novos dados (LI et al., 2017).

Desta forma, este trabalho tem como foco ampliar o projeto feito em Valêncio et al., (2020), de forma que o principal intuito deste trabalho era a construção de um ambiente de integração de dados relacional, com o intuito de que a organização dos dados fosse adequada para a aplicação de técnicas de DM efetuadas a posteriori.

No presente trabalho, o foco está na qualidade, confiabilidade e valor associado aos dados armazenados. A abordagem desenvolvida pode ser dividida em duas etapas principais, em que a primeira consiste no pré-processamento dos dados previamente coletados e a segunda consiste na seleção dos atributos. Um diagrama que representa o projeto feito está representado por meio da Figura 10.

Como ilustrado na Figura 10, bases de dados públicas oriundas de mídias sociais são obtidas, de forma que tais dados possuem ruídos, duplicações, um contexto associado ao qual foram gravados e dois idiomas originais, português e inglês. Estes dados estão em sua forma bruta, logo, são inseridos no módulo de pré-processamento. Nesta etapa, são traduzidos do português para o inglês inicialmente, e então se inicia o processo de limpeza.

Este processo visa remover os ruídos e reduzir a variedade dos dados por meio de uma formatação mais simples. Primeiro, os dados são separados palavra por palavra e os links e menções a usuários são removidos por meio de expressões regulares assim como feito em Rassol et al., (2019).

A seguir, os dados são submetidos a um processo de filtragem por meio do contexto ao qual pertencem, isto é, um usuário constrói manualmente conjuntos de palavras de parada, sinônimos e símbolos que estão associados a contextos específicos com o intuito de remover ou substituir palavras similares. O usuário deve estar ciente sobre os dados que deseja analisar e estar familiarizado com os contextos a serem analisados.

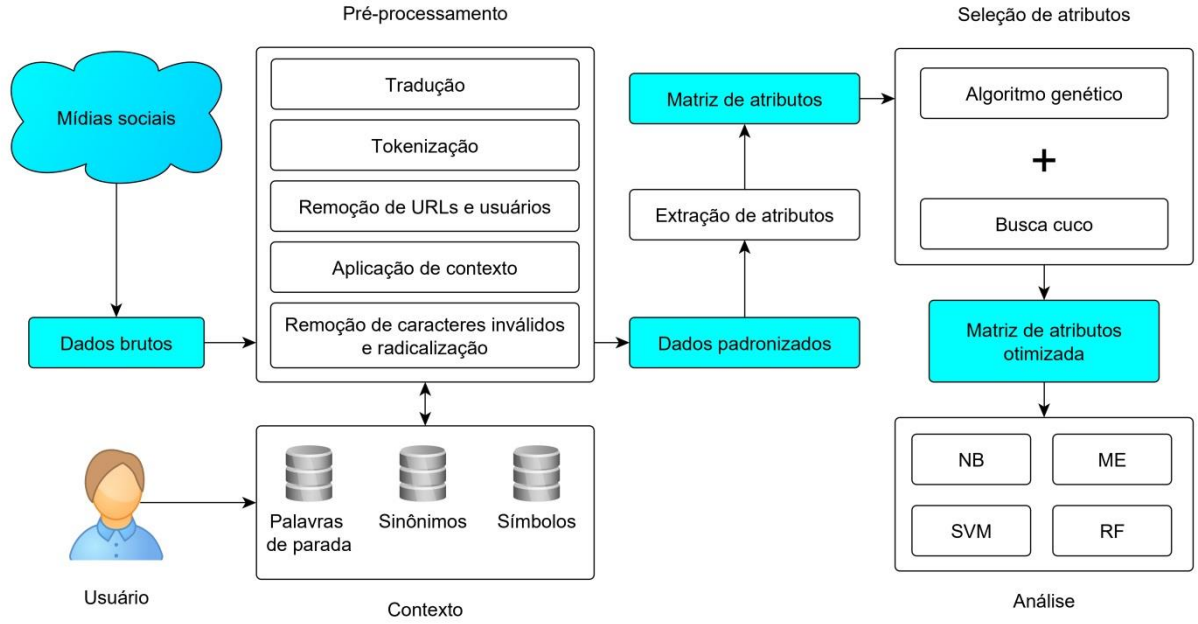
Ao final, os dados passam por uma última remoção de caracteres inválidos, como pontuação e símbolos não previstos, e são radicalizados. Todo este processo é feito com a intenção de se reduzir o número de atributos e facilitar o processamento de dados posterior.

Depois desta etapa, os dados devidamente padronizados são submetidos a um processo de extração de termos e de transformação, no qual os documentos textuais são convertidos em uma matriz de atributos por meio da técnica TF-IDF ou frequência do termo-inverso da frequência de documentos. Com isto, a matriz completa contém cada atributo considerado em

suas colunas e, na próxima etapa, o algoritmo inspirado pela natureza, que une o algoritmo genético (AG) e a busca cuco (BC), realiza uma seleção final dos melhores atributos e gera uma matriz otimizada.

A partir desta matriz são aplicados quatro algoritmos classificadores, a saber: NB, ME, SVM e RF. Então os resultados quanto à qualidade obtida com o método bioinspirado de seleção de atributos são apresentados.

Figura 10 - Diagrama do trabalho desenvolvido



Fonte: Elaborado pelo autor

3.2 Coleta de dados

A obtenção dos dados pode ser feita de diversas formas, em geral, é possível extrair dados de mídias sociais a partir da interface de programação das próprias redes sociais. Este trabalho utilizou dados oriundos de quatro bases diferentes, de forma que foram considerados somente a polaridade do texto, o contexto e o documento em si.

Estas bases de dados foram obtidas em formato CSV, ou valores separados por vírgula em português, e amostras tidas como relevantes foram selecionadas. Informações sobre o contexto e sobre a polaridade de cada documento ou amostra foram obtidas. As bases de dados possuem as seguintes características:

a) Base *Sanders*[3] – Esta base possui três classes e sentenças oriundas do *Twitter* em inglês sobre empresas de tecnologia, entre elas estão companhias como Apple, Microsoft e Google;
b) Base *Crowdflower*[4] - Esta base possui três classes e sentenças em inglês oriundas do *Twitter* sobre empresas de linhas aéreas;
c) Base *Kaggle*[5] - Esta base possui três classes e sentenças em português oriundas do *Twitter* sobre política e noticiários;
d) Base CLASME – Esta base foi criada por Valêncio et al., (2020), possui três classes e sentenças em português e inglês sobre política e empresas de tecnologia.

3.3 Pré-processamento

Desta forma, a função do pré-processamento é reduzir o número de atributos ou, neste caso, de palavras únicas, e garantir que os dados estejam padronizados. As etapas deste processo podem ser divididas em cinco tarefas principais, a saber: a tradução, tokenização, remoção de links e menções de usuários, aplicação de contexto e, por fim, remoção de caracteres inválidos. Na Figura 11 estão ilustradas as ações destas etapas por meio de um exemplo, onde um documento é considerado como entrada inicial e os demais documentos resultantes após passar por cada uma destas etapas. Os quadros em branco representam o documento textual, desde sua forma bruta até a o documento final pré-processado, enquanto que os quadros coloridos representam os processos de transformação aplicados no texto. As etapas estão apresentadas na ordem que são executadas.

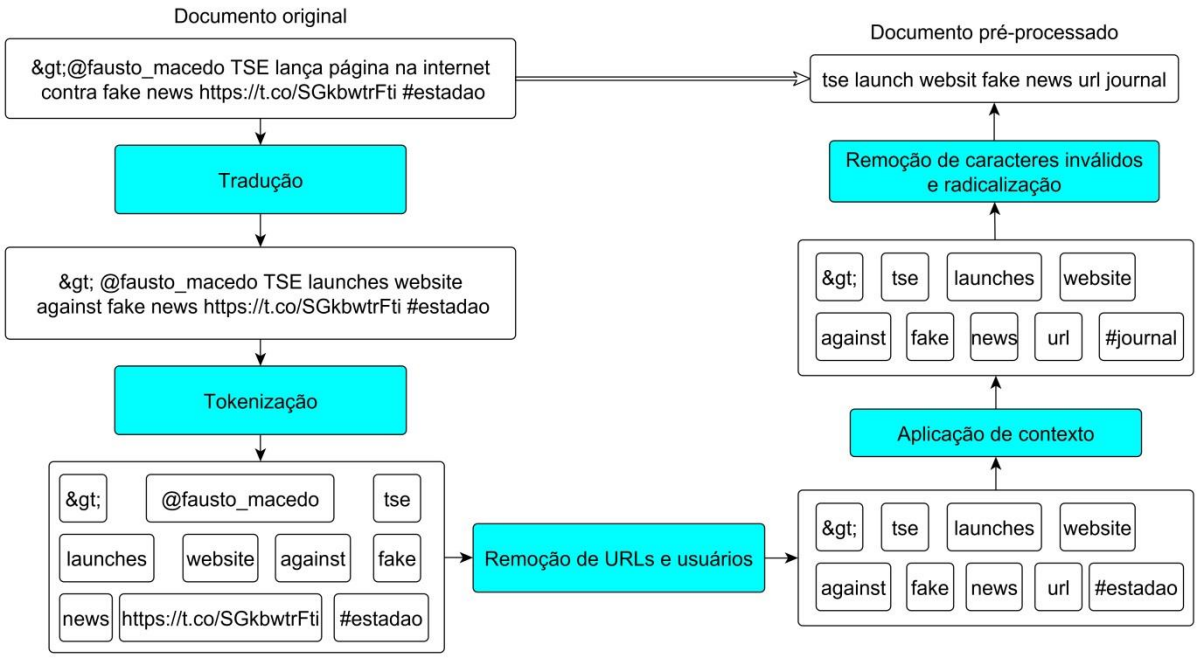
Cada documento, ou amostra, obtida na etapa anterior foi dado como entrada neste estágio com o intuito de se obter o documento padronizado. O processo com um documento de exemplo está exemplificado por meio da Figura 11, em que o dado bruto possui algumas características que podem ocorrer no ambiente de mídias sociais.

Após a tradução para o inglês, as palavras são separadas em objetos individuais, assim, cada palavra é analisada e caso seja um link ou usuário, ela é removida ou substituída. Como mostrado na Figura 11, é possível observar que o usuário "@fausto_macedo" foi removido e que a URL foi substituída por outro termo.

[3] https://github.com/zfz/twitter_corpus/blob/master/full-corpus.csv
[4] https://www.kaggle.com/crowdflower/twitter-airline-sentiment
[5] https://www.kaggle.com/augustop/portuguese-tweets-for-sentiment-analysis

Os dados, então, são processados em relação ao contexto, em que palavras reservadas são trocadas e palavras de parada são removidas, neste caso, a palavra “estadao” foi substituída por “journal”. Ao final, os caracteres inválidos são removidos e os tokens radicalizados, o que resulta na sentença final padronizada. Todo este estágio foi desenvolvido na linguagem de programação Python[6]com as bibliotecas TextBlob[7] e NLTK[8].

Figura 11 - Exemplo de pré-processamento



Fonte: Elaborado pelo autor

3.3.1 Tradução e limpeza inicial

É importante se atentar que os dados estão tanto em português quanto em inglês. Logo, uma boa forma de se lidar com um ambiente com mais de uma língua é por meio da tradução automática, isto porque a análise de sentimentos em português não possui ferramentas bem ajustadas quanto às ferramentas disponíveis para dados em inglês (PEREIRA, 2020).

Para iniciar o pré-processamento, técnicas de tradução automática estão disponíveis por meio da linguagem Python. De acordo com Pereira (2020), a tradução com ferramentas automáticas é uma estratégia competitiva em comparação com técnicas voltadas apenas para português, por mais que a tradução não seja perfeita, a qualidade dos resultados com dados traduzidos é próxima a dos dados originais em inglês.

[6] https://www.python.org/
[7] https://textblob.readthedocs.io/en/dev/
[8] https://www.nltk.org/

Segundo Pereira (2020), além de vários estudos comprovarem que a tradução de dados para o inglês é o que gera os melhores resultados, é necessário desenvolver recursos de linguagem mais efetivos, além de estabelecer conceitos de conhecimento geral e recursos para domínios específicos da língua portuguesa. A justificativa para tal se baseia no fato de palavras mudarão entre sentenças de linguagens diferentes, porém, os métodos de tradução automática, se eficazes, não devem mudar a polaridade do sentimento ou a opinião expressa na sentença, desde que isto se mantenha, este método é razoável para a análise de sentimentos (ARAÚJO; PEREIRA; BENEVENUTO, 2020).

Para realizar a tradução dos dados, a interface de programação do *Google* foi utilizada por meio da linguagem Python. Após toda a base de dados ser traduzida, uma análise manual foi feita nos termos mais frequentes, de forma que se adequassem os termos mais frequentes que não foram traduzidos. Por exemplo, algumas gírias e abreviações como “hj” e “mt” foram traduzidas, nestes casos, respectivamente como “*today*” e “*much*”.

Deste modo, os dados agora podem ser “tokenizados”, ou divididos em partes individuais, em que cada palavra ou aspecto foi isolado. A seguir, é possível identificar uma URLs e menções de usuários por meio de uma expressão regular, de modo a remover ou substituir tais palavras de acordo com a necessidade. Isto é importante não apenas para reduzir o número de atributos, mas para evitar que os nomes destes usuários interfiram na classificação. Nesta etapa, os dados são formatados para sua forma minúscula ao final.

3.3.2 Aplicação de contexto

Conforme mencionado anteriormente, cada amostra possui um contexto associado, isto significa, por exemplo, que o documento pode estar associado a uma opinião sobre um produto. É importante considerar que cada contexto possui algumas palavras especiais que possuem um significado bem definido (EL ANSARI; ZAHIR; MOUSANNIF, 2018).

Um exemplo prático ocorre na base de dados utilizada neste trabalho, a palavra “apple” que significa “maçã” possui uma forte inclinação a estar relacionada como uma fruta em um cenário amplo, porém no contexto de empresas de tecnologia, esta palavra está associada à empresa de nome *Apple*.

Assim, para lidar com este problema, um conjunto de símbolos, palavras de parada e palavras sinônimas, ou tesauros, foi construído manualmente para cada contexto identificado na base de dados. Tal base de dados possui três contextos associados, que são: empresas de tecnologia; empresas aéreas e conteúdo relacionado a notícias e política.

Para se criar o conjunto de sinônimos e de símbolos, os termos mais frequentes de cada contexto foram analisados manualmente e os termos que mais geravam ambiguidade ou redundância foram selecionados.

Com isto, criou-se uma tabela de comparação que foi utilizada para a limpeza e substituição dos dados. Em particular, as palavras seguidas das *hashtags* foram selecionadas e verificadas quanto a sua utilidade, assim como emojis e símbolos que podem auxiliar a identificação do sentimento dos documentos.

A partir disto, o conjunto de palavras foi elaborado e aplicado na base de dados. Alguns casos foram escolhidos e estão representados na Tabela 2, onde é ilustrado alguns dos principais aspectos e dos sinônimos utilizados.

Com isto, é possível identificar que os usuários utilizaram jargões comuns do *Twitter*, mas que podem gerar ambiguidade e interferir negativamente na classificação, por exemplo, a palavra "siri", um conhecido crustáceo, está associada ao nome da assistente digital de uma empresa, e "*ice cream*", que significa "sorvete" em inglês, na verdade está relacionado ao nome do sistema operacional *Android*.

Tabela 2 - Exemplos de sinônimos relacionados a contexto

Contexto	**Palavras originais**	**Sinônimos**
Empresas de tecnologia	*apple*	*company*
	siri	*system*
	ice cream	*system*
	windows	*company*
Empresas de transporte aéreo	*southwestair*	*company*
	jfk	*airport*
	rebook	*book*
	virginamerica	*company*
Notícias e conteúdo político	*folha*	*jornal*
	estadao	*jornal*
	são paulo	*sp*
	g1sp	*journal*

Fonte: Elaborado pelo autor

Da mesma maneira, a palavra "*windows*" cuja tradução literal é janela, também foi substituída. No contexto de transporte aéreo, "jfk" é o nome de um aeroporto e "*rebook*" está associado ao verbo "reservar", uma palavra que também foi substituída. As palavras "folha" e

“estadao” não são traduzidas, pois são nomes próprios, e, neste contexto, se tratam de veículos de notícias. Eles são trocados por seus respectivos sinônimos.

De forma similar, as palavras de parada também foram identificadas manualmente, porém foram unidas com as palavras de parada padrão da língua inglesa da biblioteca sklearn[9]. Após testes empíricos serem realizados, o conjunto de palavras de parada padrão se mostrou o mais estável para a base de dados utilizada, de modo que diferentes arranjos dos dados resultavam em valores de qualidade muito distintos.

Existem processos automatizados como utilizado por Muhammad; Wiratunga e Lothian (2016), porém não foi o foco deste trabalho desenvolver uma estratégia que lidasse diretamente apenas com o contexto dos dados e outras variações de análises léxicas como negação, intensificação e sarcasmo.

3.3.3 Padronização final

A última etapa do processo de limpeza consiste na remoção de caracteres especiais, que se tratam de pontuação, acentos, caracteres não alfanuméricos e quaisquer outros caracteres que não estejam presentes na tabela ASCII. Este processo varre todos os tokens e remove os caracteres considerados inválidos se estiverem presentes.

Além disso, algumas palavras reservadas como “rt” e “gt” também são removidas, pois são termos frequentes e estão associados a respostas ou símbolos que ocorrem em redes sociais.

Nesta etapa, algumas palavras são normalizadas quanto a sua gramática, em especial palavras que atuam como conectivos ou palavras como “sim” e “não”, em que os usuários costumam digitar com erros ou variações, tais como “simmm”, “s”, “n”, “nope” e “naum”. Estas palavras foram algumas das que não foram traduzidas pela tradução automática.

Por fim, o último método empregado foi a radicalização, em que os termos foram reduzidos ao seu radical, isto é, variações por conta de plural, gênero e outras instâncias passaram por um processo que visa reduzir a quantidade geral de atributos e ajuda a identificar palavras semelhantes, o que tende a auxiliar o DM (HASSONAH et al., 2020).

O radicalizador empregado neste caso foi o *Snowball*, disponível na biblioteca NLTK do Python. Tal método apresentou os melhores resultados em relação à diminuição do número de atributos enquanto manteve uma qualidade similar aos outros métodos testados.

[9] https://scikit-learn.org/

3.4 Extração de características

Nesta etapa, os dados padronizados precisam ser convertidos para um formato em que seja possível analisá-los computacionalmente. Para tal, é necessário gerar a matriz de termos de documentos, ou a TDM. Tal matriz foi criada a partir da técnica TF-IDF que consiste em converter as sentenças em um vetor de características, de forma que cada ocorrência única de termos é colocada como uma coluna e cada documento é uma linha.

Este método permite que cada célula da matriz tenha um valor de peso entre 0 e 1, de modo que é possível verificar a frequência de determinados atributos. Somado a isto, foi configurado que atributos que ocorressem em menos de dois documentos, e atributos que ocorressem em mais de 70% dos documentos seriam desconsiderados. Assim, apenas os termos intermediários são utilizados, pois acabam por ter o peso maior, e consequentemente, são os termos mais relevantes, assim como realizado por correlatos (HASSONAH et al., 2020; KUMAR; JAISWAL, 2019).

Além disso, é possível realizar normalizações para evitar divisão por zero e aplicar a técnica n-gramas com diferentes valores, de modo a se verificar qual a melhor combinação. No presente caso, a técnica 1-grama foi utilizada como base, pois se verificou que esta configuração atingiu os melhores valores de qualidade. O método TF-IDF foi implementado com a biblioteca sklearn previamente citada.

Após a extração de características, a matriz de atributos com a base de dados total ainda precisa receber os valores da classe, isto é, cada linha precisa estar associada a uma classe, neste caso, às classes positiva, negativa e neutra, que foram representadas com os valores inteiros 1, -1 e 0 respectivamente. Com isto, os dados podem ser submetidos ao processo de seleção de características propriamente dito.

3.5 Algoritmo bioinspirado

Inicialmente, é importante destacar que o algoritmo de seleção desenvolvido busca tanto obter um subconjunto de características que atinja melhores valores de qualidade, quanto reduzir o máximo possível o número de atributos sem que isto prejudique a classificação. Para realizar tal tarefa, existem dezenas de algoritmos meta-heurísticos propostos na literatura (SHARMA; KAUR, 2020), porém alguns deles se destacam como métodos interessantes para a área de AS.

De acordo com Sharma e Kaur (2020), a BC se destaca por ser um dos algoritmos mais utilizados nos últimos anos, além de ser bastante empregado em vários trabalhos como uma variante binária para lidar com diferentes problemas. Sua principal vantagem é o fato de que se trata de um método simples, eficiente e ajustável a diferentes cenários, contudo este método também possui uma desvantagem que consiste no fato de que a BC pode ficar presa em ótimos locais (YADAV; VISHWAKARMA, 2020).

Com isto, uma das formas de se criar uma maior diversificação de soluções e evitar tal problema consiste em utilizar o AG, pois possui diversas implementações que permitem variar o conjunto de possíveis soluções, além de ser indicado para análise de sentimentos porque sua construção básica usa vetores binários.

Deste modo, o algoritmo criado é uma mescla de BC e AG, denominado algoritmo Busca Cuco Genético ou BCG. O objetivo deste método é ser capaz de identificar um bom subconjunto de atributos de forma a utilizar a estratégia de exploração de soluções da BC de forma rápida, e, com uma dada frequência, faça as operações do AG na população. Assim, o algoritmo BCG almeja encontrar possíveis boas soluções sem comprometer a estabilidade, escalabilidade e custo computacional que a BC possui, uma vez que se sabe que o AG é consideravelmente mais custoso.

A ideia deste método é fazer com que as soluções que eventualmente ficam estáticas devido à estratégia da BC sejam perturbadas, de forma a evitar uma convergência prematura de soluções. Além disso, devido ao fato de que a BC gera apenas uma única solução com potencial para ser a melhor a cada geração, aplicar o AG pode fazer com que outras boas soluções sejam geradas a cada geração, além de ampliar ainda mais a exploração de soluções no espaço de busca com uma estratégia que não seja aleatória.

3.5.1 Abstração das soluções

Portanto, para que seja possível identificar um bom subconjunto de atributos, é necessário construir um conjunto de vetores binários, conhecidos como população, da forma $V_i = \{v_1, v_2, \dots v_d\}$ em que V_i é um vetor de soluções ou atributos binários, $v_j \in \{1|0\}$ e d é a quantidade de atributos totais ou dimensão. O melhor ninho, ou indivíduo, é aquele cujos atributos, ou ovos, com o valor 1 conseguem atingir o maior valor de acurácia, ou o ninho com maior acurácia e o menor número de atributos selecionados, a depender da função.

Devido ao fato da BC trabalhar diretamente com dados contínuos, isto é, para calcular a distância de um salto a partir de um ninho, o algoritmo gera valores resultantes $x \in \mathbb{R}$.

Logo, é necessário convertê-los para valores binários. Neste trabalho, o tamanho do salto foi ajustado para ficar em um intervalo $I \in [0,1023]$, de tal modo que a parte decimal é descartada e o valor inteiro é convertido para um vetor na forma $V_i = \{1,1,1,1,1,1,1,1,1,0\}$, em que este valor representa o número 1022.

Dada uma matriz de atributos com um número d de colunas, a BCG gera uma população de vetores com este tamanho d. A cada dez posições percorridas, um valor é convertido de binário para decimal e, a partir dele, a BCG consegue efetuar os cálculos para gerar uma próxima possível boa solução. Isto foi feito para possibilitar a aplicação dos operadores genéticos posteriormente.

A última parte do vetor que eventualmente é convertida para o valor binário mais próximo, por exemplo, se existirem 53 dimensões, as três últimas posições são convertidas para valores no intervalo $I \in [0,7]$.

3.5.2 Geração de solução por BC

Após gerar a população e avaliar a aptidão de cada ninho, a próxima etapa é a geração de um novo ninho por meio dos voos de Levy. A BC possui duas formas de explorar o espaço, uma local e outra global, em que a principal diferença consiste na forma de calcular ambas e no tamanho do salto.

Segundo Yang (2017), uma das características da BC está no fato de que o algoritmo possui um balanço razoável entre exploração e explotação de soluções, isto é feito por meio da variação das equações de busca local e global, ambas foram aplicadas a cada geração ao acaso, com o intuito de gerar soluções distantes e próximas de uma boa solução descoberta na mesma proporção.

Com isto, a geração de um novo ninho é feita da seguinte forma: após a criação e avaliação da população inicial, os ninhos são ordenados de modo crescente e uma boa solução é escolhida como uma base para o novo ninho. Conforme explicado na Análise de Sentimentos, a abstração é que uma ave cuco tentará gerar um bom ninho a partir desta boa solução descoberta. Então, por meio das Equações (5 e (7, um novo ninho, isto é, conjunto de ovos que representam uma solução é gerado. Então se seleciona um ninho aleatório J, se a aptidão do novo ninho for melhor, então é substituído.

Desta forma, a última etapa relacionada à BC tem como objetivo substituir uma porção P_a da população por novos indivíduos gerados aleatoriamente, tal porção se trata das piores soluções até o momento. Adicionalmente, o restante da população, que possui uma aptidão

melhor, é mantido. Ao final da época, as soluções geradas aleatoriamente são avaliadas e a população é ordenada para se criar a porção P_a da próxima geração, assim, o melhor indivíduo é identificado e salvo.

3.5.3 Operadores genéticos

Com o intuito de preservar as vantagens da BC, os operadores genéticos são aplicados em certas épocas, isto é, há uma frequência em que são chamados para evitar que o algoritmo fique muito custoso, porém é suficiente para variar as soluções de modo a fazer buscas mais amplas. Assim, o valor desta frequência foi ajustado para que a cada dez gerações, em vez de executar as operações de substituir a porção P_a da população, os operadores são aplicados na população toda. O elitismo é aplicado inicialmente para não se perder a melhor solução encontrada até o momento, e então, os operadores são aplicados na seguinte ordem: seleção, cruzamento e mutação.

Primeiro, uma seleção por meio de torneio é utilizada, de modo que três indivíduos são selecionados ao acaso e o indivíduo com melhor aptidão é selecionado. Isto é repetido até que uma nova população temporária seja formada para sofrer os próximos modificadores. Todos os indivíduos têm a mesma chance de serem escolhidos e, a cada novo torneio, podem ser escolhidos novamente. Ao final, uma nova população foi selecionada.

Com isto, quaisquer soluções têm chance de continuar, mas a população no geral tende a melhorar. Este método foi utilizado porque permite que até as soluções ruins possam continuar (KRAMER, 2017), assim, o aspecto da exploração de novas soluções é destacado, uma vez que uma parte considerável da população fica estática durante a execução da BCG.

A partir disto, a operação de cruzamento ocorre da seguinte forma: cada par de indivíduos possui uma probabilidade P_c de sofrer o processo de cruzamento, caso contrário, os filhos são cópias idênticas dos pais. Os indivíduos pais são subdivididos a cada dez dimensões e então é aplicado um cruzamento com um ponto de divisão, tal ponto é escolhido ao acaso dentro do intervalo $I \in [2,8]$, e então cada combinação de dimensões é armazenada nos dois indivíduos filhos. Logo, os filhos são uma combinação de atributos feita a cada dez posições dos vetores pais, e uma nova população é gerada.

A etapa de mutação é feita de forma similar, em que cada indivíduo possui uma probabilidade P_m de sofrer mutações. Assim, se o indivíduo for selecionado, sofre alterações no valor da posição, de modo a substituir valores zeros por uns e uns por zeros. Cada posição do vetor pode receber a alteração, mas para evitar que o indivíduo seja muito alterado, existe

um valor máximo de até 40% de alterações que o indivíduo pode sofrer neste processo, isto é feito para evitar que uma solução fique deformada (KRAMER, 2017).

No final os indivíduos são avaliados e seus valores de aptidão são calculados. A avaliação é feita por meio da função de aptidão, utilizada em ambas as partes do algoritmo BCG. A função de aptidão levou em consideração a acurácia do subconjunto encontrado como aptidão, denominada função f_1, assim como feito por outros trabalhos (KUMAR; JAISWAL, 2019; HASSONAH et al., 2020). O classificador utilizado para tal foi o ME.

Para fins de verificação, uma segunda função de aptidão foi proposta com base nos trabalhos da literatura, cujo foco estava não somente em encontrar o subconjunto que atingisse o maior valor de acurácia, mas que também buscasse minimizar o número de atributos com o intuito de diminuir o custo computacional de análises posteriores.

A segunda função de aptidão está representada na Equação (13, denominada função f_2, em que X é o subconjunto encontrado, c_1 e $c_2 \in \mathbb{Z}$ e são constantes, $Ac(X)$ é a acurácia e $Na(X)$ é a quantidade de atributos com valor um, ou seja, os atributos selecionados.

Esta variação na função de aptidão foi proposta para se verificar se é possível atingir valores razoáveis de acurácia mesmo com o fator que pressiona uma redução no número de atributos.

$$Aptidão(X) = c_1 * Ac(X) + c_2 * \frac{1}{Na(X)} \tag{13}$$

3.5.4 Visão geral

Deste modo, a BCG proposta possui dois critérios de parada para garantir a finalização do programa. O primeiro critério de parada é feito por meio de um número máximo de gerações estabelecido, após a realização de testes, verificou-se que o valor 1000 seria suficiente para identificar uma possível boa solução.

Além disso, o segundo critério de parada ocorre caso a melhor solução encontrada fique estática por mais de 50 gerações, de modo que possivelmente uma boa solução já fora encontrada. Tal valor fora estipulado a partir de testes empíricos. Este modelo elaborado foi desenvolvido com a intenção de se fazer uma busca escalável e estável, enquanto variações mais abruptas são aplicadas nos dados com certa frequência.

Uma visão geral do funcionamento do algoritmo de seleção de atributos é ilustrada por meio da Figura 12, e por um pseudocódigo do método descrito na Figura 13. Recapitulando,

inicialmente uma população de n ninhos com tamanho d de dimensão é iniciada aleatoriamente com valores binários, e todos os ninhos são avaliados e recebem uma aptidão.

Os ninhos são ordenados de acordo com este parâmetro de qualidade e então se inicia o laço principal do algoritmo, que continua a executar até que a melhor solução fique estagnada após um determinado período ou até que o número máximo de épocas tenha sido alcançado.

Após a ordenação dos ninhos, a busca se inicia com a geração de uma nova possível boa solução que é feita por VL. Isto é executado a partir de um bom ninho previamente descoberto que funciona como um ponto de partida para a criação do ninho i, e este novo ninho pode ser gerado tanto por meio da busca local quanto da busca global.

Devido ao fato de que não há informações prévias sobre quais dimensões ou ovos devem ser mantidos, esta busca tem 50% de chance de ser tanto global quanto local. Para calcular a busca local, outros dois bons ninhos distintos da população são selecionados ao acaso, como mostrado na Equação (7).

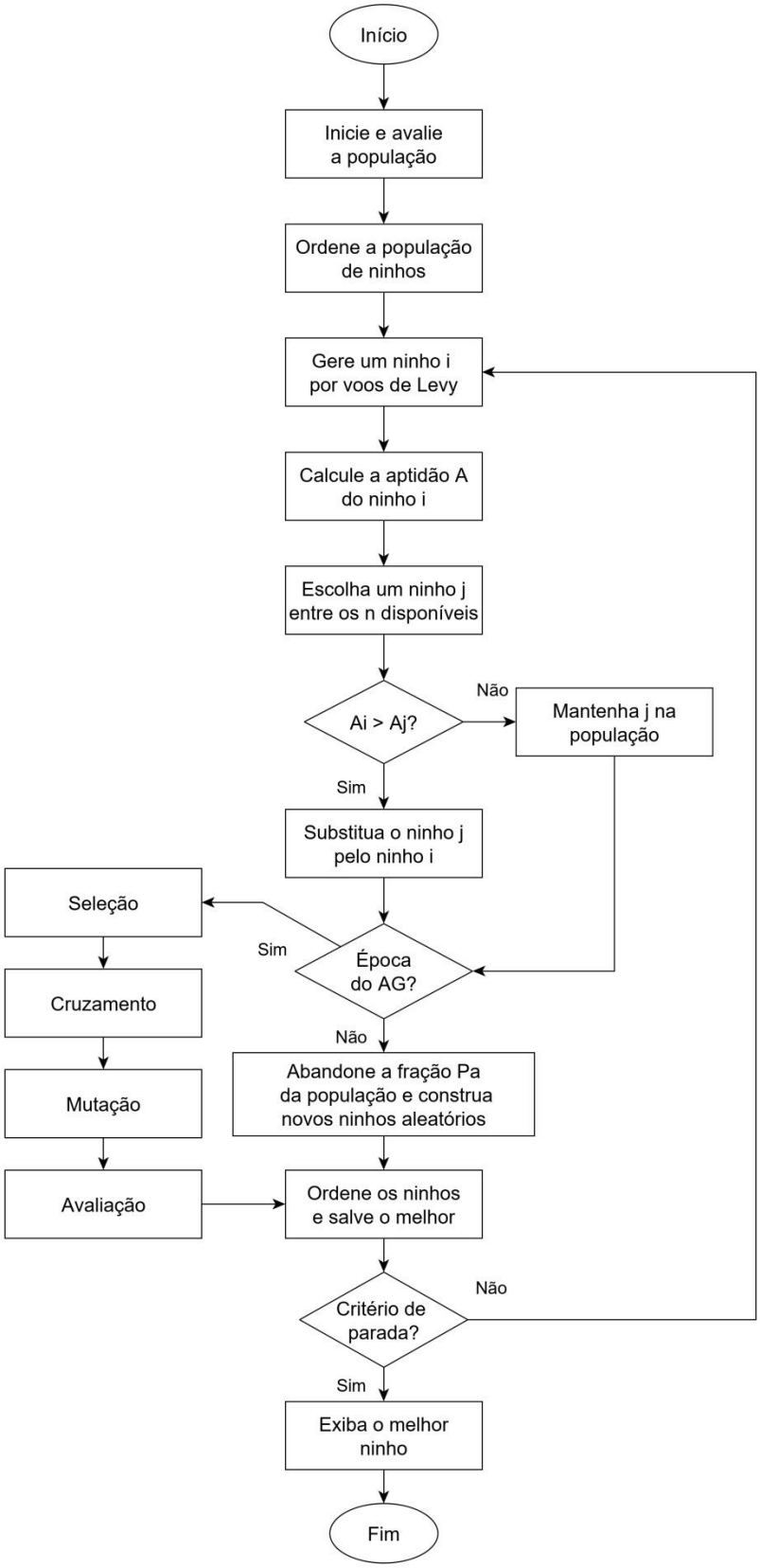
A aptidão A_i é calculada e então um ninho aleatório, seja ele o ninho j, da população é selecionado. Se $A_i > A_j$, o ninho i é inserido na população, caso contrário nada é feito. Então é verificado se a época t está contida na frequência de aplicação do AG, em caso afirmativo, os operadores genéticos são aplicados em toda a população e então o laço principal reinicia. Em caso negativo, o algoritmo segue a ordem de execução da BC, em que uma porção $p_a$dos dados é substituída por novas soluções aleatórias e o restante é mantido.

Ao final, os ninhos precisam ser ordenados novamente e o melhor ninho é identificado e salvo para a próxima geração. Se o critério de parada foi alcançado, este melhor ninho é apresentado. Caso contrário, o algoritmo volta para a etapa de geração de uma nova solução por VL.

Tal estratégia foi desenvolvida com o intuito de ampliar a descoberta de potenciais soluções boas, ao mesmo tempo em que se busca evitar um excessivo custo computacional. Devido ao fato de que os operadores genéticos podem ser custosos, eles são aplicados pontualmente para perturbar as soluções e eventualmente chegar a novos pontos no espaço de busca.

Adicionalmente, se os operadores genéticos forem acionados, isto quer dizer que toda a população é submetida a uma avaliação. Assim, não seria interessante aplicar a substituição da porção P_a da população, uma vez que esta perturbação aplicada se trata dos operadores genéticos e não há necessidade de se alterar a população duas vezes na mesma geração.

Figura 12 - Fluxograma do algoritmo BCG



Fonte: Elaborado pelo autor

Figura 13 - Pseudocódigo do algoritmo BCG

Início
Iniciar a população P com n ninhos
Avaliar a população e calcular a aptidão A de cada ninho
Ordene a população P de acordo com a aptidão de cada ninho
Enquanto (t<MaxGerações) ou (Critério)
Selecione um cuco aleatoriamente por VL e avalie sua aptidão F_i
Escolha um ninho j aleatório entre os n possíveis
Se ($A_i > A_j$**)**
Substitua a solução do ninho j por i
Fim_se
Se (t ∈ Época de AG)
Salvar o melhor ninho em I_{melhor}
Aplicar SELEÇÃO em P
Aplicar CRUZAMENTO em P com P_c
Aplicar MUTAÇÃO em P com P_m
Avaliar a população P e calcular a aptidão de cada ninho
Inserir I_{melhor} em P
Senão
Uma fração (P_a) dos piores ninhos é abandonada e novas soluções são geradas
Manter as melhores soluções ou os melhores ninhos
Ordene as soluções e salve a melhor solução
Fim_se
Fim_enquanto
Exibir a melhor solução
Fim

Fonte: Elaborado pelo autor

3.6 Considerações sobre o trabalho desenvolvido

Neste capítulo discorreu-se a respeito da implementação desenvolvida para lidar com os dados textuais de mídias sociais. Uma visão geral do conjunto de técnicas utilizado foi apresentada, assim como detalhes de cada uma das etapas. O modelo contempla desde a etapa de aquisição de dados brutos até o processo de classificação final.

Após a limpeza inicial dos dados, a estratégia de tradução e de transformação dos dados por meio do contexto ao qual cada documento pertence e alguns exemplos foram apresentados. Por fim, ambos os métodos de extração de características e o algoritmo bioinspirado foram descritos, desde sua abstração até a forma como a BCG almeja alcançar um equilíbrio entre exploração e explotação de soluções.

Capítulo 4

Testes e resultados

Neste capítulo é apresentada a metodologia de experimentação, e os detalhes de execução dos testes de qualidade da abordagem proposta, assim como discute os resultados obtidos e discorre sobre as possíveis implicações dos testes.

4.1 Metodologia de experimentação

O intuito dos testes realizados é averiguar o algoritmo quanto às métricas de qualidade, como os valores de acurácia, e analisar os resultados sobre o número de atributos selecionados em média, de forma a se considerar a execução de algoritmos de classificação sem uma seleção de atributos prévia e algoritmos de seleção tradicionais.

Além disso, uma análise quanto à qualidade do método inspirado pela natureza em relação a outras técnicas heurísticas foi feita com o objetivo de verificar as vantagens do método desenvolvido em relação a outras abordagens. Portanto, para que fosse possível analisar os dados gerados pelo algoritmo desenvolvido, todos os testes não determinísticos foram executados dez vezes de forma independente e os resultados representam a média e desvio-padrão obtidos, isto foi feito para garantir consistência estatística. As especificações do ambiente de testes estão na Tabela 3.

Tabela 3 - Especificações do ambiente de teste

Especificação	**Valor**
Processador	Intel core i5 7200U
Memória	DDR4 8Gb a 1066 MHz
Disco rígido	SSD 480GB
Sistema operacional	Windows 10

Fonte: Elaborado pelo autor

A seguir, estão descritos os materiais e métodos utilizados nos testes, assim como especificações sobre os parâmetros configurados nos algoritmos e informações mais detalhadas sobe a base dados.

4.2 Materiais e métodos

Como informado previamente, este trabalho foi desenvolvido na linguagem de programação Python, assim, toda a etapa de pré-processamento, seleção e classificação de dados foi implementada por meio das bibliotecas explicadas na seção Abordagem híbrida bioinspirada.

As métricas de qualidade consideradas foram as mesmas utilizadas por trabalhos correlatos, tais métricas são a acurácia, que mede a quantidade de instâncias classificadas corretamente, independente do número de classes, e a quantidade de atributos selecionados (KUMAR; JAISWAL, 2019; AKTHAR et al., 2017; HASSONAH et al., 2020).

4.2.1 Bases de dados

Os dados utilizados foram obtidos com a junção das quatro bases de dados apresentadas previamente. Tal base de dados, após uma remoção de amostras irrelevantes, resultou em um conjunto com 12.000 amostras das quais é possível gerar vários subconjuntos para testes. No entanto, quatro bases de dados menores foram criadas a partir de uma seleção aleatória dos dados presentes no conjunto total, de modo que estas bases foram nomeadas de A até D. A base de dados “A” possui dados de todos os três diferentes contextos, enquanto que a base de dados “B” contém dados relativos à tecnologia, a base “C” detém dados sobre transporte aéreo e a base “D” abrange os dados de noticiários e conteúdo político.

A base “A” possui amostras cujos idiomas originais eram português e inglês em igual quantidade, enquanto que as bases “B” e “C” possuem frases originalmente em inglês e a base “D” em português.

Em relação à quantidade de amostras, as bases de dados “A”, “B” e “C” possuem 600 amostras cada, e a base “D” possui 900 amostras. Estas e outras especificações estão representadas na Tabela 4.

Tabela 4 - Especificações das bases de testes

Base	Contexto	Amostras	Número de atributos original	Número de atributos pós limpeza	Idioma original
A	Empresas de tecnologia, empresas de transporte e noticiários	600	1.783	300	Inglês e português
B	Empresas de tecnologia	600	1.708	346	Inglês
C	Empresas de transporte aéreo	600	1.711	327	Inglês
D	Noticiários e conteúdo político	900	2.361	453	Português

Fonte: Elaborado pelo autor

O principal intuito destas bases de dados foi de que as amostras fossem escolhidas de modo que as bases ficassem balanceadas, isto é, os dados em português e inglês foram dispostos em 50% na base A, e os dados das classes positiva, negativa e neutra estão dispostos na mesma quantidade em todas as bases. A maior discrepância ocorre na proporção do contexto, em que o contexto de noticiários e política ocupa metade da base A, enquanto os outros dois contextos sobre empresas de tecnologia e empresas aéreas ocupam a outra metade.

A base de dados A teve seus documentos igualmente distribuídos quanto a suas classes e seus idiomas, isto foi feito para que não houvesse um viés sobre estes aspectos. Inicialmente as bases de dados apresentam um número elevado de atributos, em que as bases A, B, C e D possuem 1.783, 1.708, 1.711 e 2.361 atributos, respectivamente. Tais atributos representam as palavras com ocorrência única após a realização da etapa de pré-processamento.

Depois da transformação em uma matriz de termos, o número de atributos de cada base de dados foi reduzido para 300, 346, 327 e 453. Isto ocorreu por conta dos valores associados aos pesos TF-IDF. Nota-se, portanto, que o pré-processamento manteve a quantidade de atributos final similar às outras bases de igual tamanho. Os algoritmos classificadores empregados foram os seguintes: ME, NB, SVM e RF.

4.2.2 Especificações dos testes

O método de validação adotado foi o *hold-out*, feito na proporção 70-30, em que os valores representam um porcentual do conjunto de treino e do conjunto de teste, respectivamente. Com o intuito de verificar possíveis problemas na ordem dos dados, uma seleção aleatória com cinco estados de dispersão dos dados foi analisada em cinco execuções independentes.

Isto foi aplicado em todos os algoritmos estocásticos, de forma a evitar uma possível ordenação inicial boa ou ruim para a classificação. Assim, as médias foram calculadas e estão representadas nos resultados juntamente com o desvio-padrão.

4.3 Avaliação do método

Antes de iniciar os testes na base de dados propriamente dita, é interessante apresentar os resultados de uma análise prévia do modelo desenvolvido aplicado a um diferente contexto. Isto foi feito com o objetivo de verificar bons parâmetros iniciais e constatar a qualidade do método em um caso genérico como feito por Yang e Deb (2009).

Desta forma, o código apresentado na Figura 13 foi implementado para resolver um problema de minimização de uma função com dez dimensões. Devido ao fato de que existem boas estimativas iniciais para os algoritmos BC e AG, torna-se importante determinar qual a taxa de ocorrência de aplicação dos operadores genéticos.

Os parâmetros iniciais foram escolhidos com base em outros estudos (SHEHAB; KHADER; AL-BETAR, 2017) e testes iniciais, eles são: $P = 100$, $P_a = 0{,}1$, $\alpha = 1$, $\beta = 2$, $P_c = 0{,}7$, $P_m = 0{,}1$ e foi escolhido como critério de parada 500 gerações sem melhoria no melhor indivíduo.

Os valores do espaço de busca foram aplicados no intervalo $[-1000; 1000]$, os indivíduos podem assumir valores $x \in \mathbb{Z}$ e foram consideradas $n = 10$ dimensões para se encontrar a solução da Equação (14):

$Minimizar$

$$Z = f(X) = 10n + \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i)] \tag{14}$$

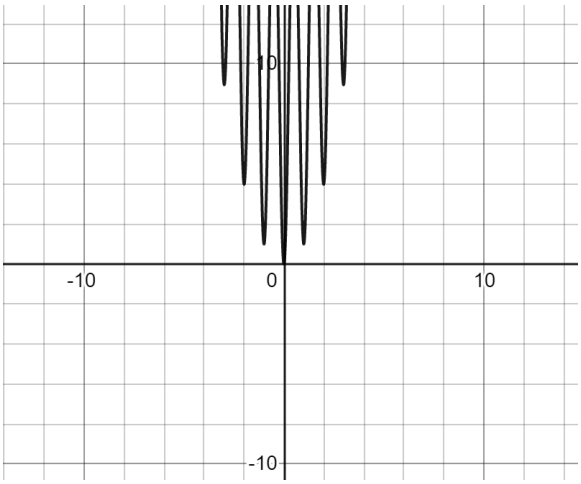
$sujeito\ a$

$$-1000 \leq X \leq 1000, \quad i = 1{,}2, \ldots, 10$$

Neste exemplo, a função é conhecida como “Rastrigin” e pode ser calculada para qualquer $n \in \mathbb{N}$ (YANG; DEB, 2009). Tal função possui diversos pontos ótimos locais, porém possui um ótimo global no ponto $X = [0,0,\ldots,0]$ e $f(X) = 0$, o que pode ser visualizado por meio da Figura 14 que representa a imagem desta função em um intervalo $[-10; 10]$. Se quaisquer valores de x assumirem valores altos, Z se torna um valor grande.

Tal função foi selecionada para aferir se o algoritmo desenvolvido seria capaz de lidar com uma função cuja imagem é bastante variável, ou seja, pontos bons estão próximos de pontos ruins. Assim, esta uma função que visa simular um cenário difícil ao colocar vários pontos ótimos locais, com foco em determinar quais os parâmetros do método BCG para um caso com dados reais de redes sociais.

Figura 14 - Função de teste Rastrigin

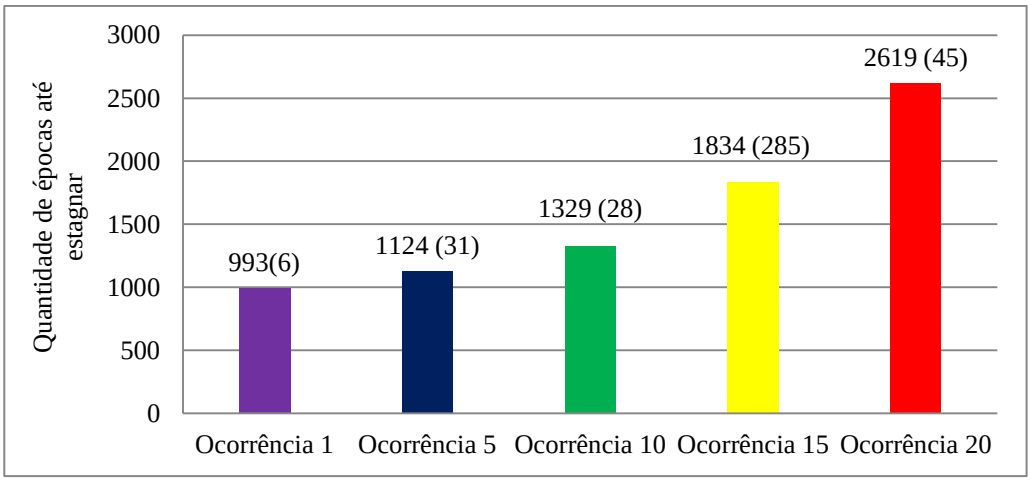


Fonte: Elaborado pelo autor

A partir disto, testes foram executados a fim de verificar os parâmetros mencionados e, sobretudo, qual a melhor frequência de aplicação dos operadores genéticos, tal parâmetro foi variado com os seguintes valores: 1; 5; 10; 15 e 20. O objetivo fora identificar qual o valor mais elevado que impactasse menos a qualidade e número de gerações do método, afinal, os operadores genéticos são relativamente custosos e sua chamada deve ser minimizada sem que isto interfira na capacidade de encontrar uma boa solução com poucas épocas.

A seguir, um gráfico foi construído para apresentar a velocidade e forma com que as soluções foram obtidas. Com o intuito de sintetizar o comportamento do algoritmo, o gráfico que apresenta a variação da taxa de ocorrência de 1 até 20 está representado por meio da Figura 15. Os resultados estão representados da seguinte forma: a altura da barra é a quantidade de épocas executadas até a solução convergir e o valor entre parênteses é a melhor aptidão.

Figura 15 – Comparação do desempenho de BCG com variação da ocorrência



Fonte: Elaborado pelo autor

Ao se analisar o gráfico, é possível notar que a aptidão média da população tende piorar conforme a taxa de ocorrência de aplicação de operadores genéticos aumenta, assim como é necessário escolher um valor próximo da ocorrência 1. Devido ao fato de que é necessário escolher um valor de ocorrência que seja o mais alto e, ao mesmo tempo, com baixa interferência na qualidade, o valor mediano que passa a ser o mais adequado, visto que apresenta um equilíbrio entre custo computacional e velocidade de conversão. A configuração com taxa 1 atingiu uma aptidão $f(X) = 6$, em 993 gerações, portanto é possível considerar este o melhor cenário inicial em termos de qualidade, mas é o mais custoso em termos de recursos, assim é importante buscar uma configuração com qualidade similar.

O próximo apresenta uma aptidão igual a 31 em 1.124 gerações. Isto significa que houve uma pequena piora na aptidão, e um aumento de aproximadamente 13,19% no número de épocas. Logo, não há uma diferença notável e tanto a aptidão quanto o número de gerações são similares. A execução com taxa de ocorrência 10 também apresenta um comportamento similar, em que o valor de aptidão atingido foi de 28 e o número de iterações foi de 1.329, o que indica um aumento de 33,83% em relação ao primeiro caso. Como pode ser observado, o valor da aptidão ficou próximo dos casos anteriores, ao passo que o número de gerações também não sofreu um aumento drástico.

Os casos seguintes de taxas configuradas para 15 e 20, respectivamente, mostram um aumento considerável em relação ao número de épocas. Estes valores foram para 1.824 e 2.619, o que representa um aumento arredondado de 84,69% e 163,74%, enquanto que os valores de aptidão ficaram em 285 e 45, respectivamente. Isto indica que a qualidade do algoritmo pouco se altera em relação à variável ocorrência, porém o número de iterações necessárias para convergir praticamente dobra quando a frequência é ajustada para 15.

Por meio de sua análise, nota-se um comportamento não linear quanto ao número de gerações e que não seria interessante continuar a verificar outros valores maiores de ocorrência.

Além disso, devido ao fato de a taxa de ocorrência 15 praticamente dobrar o número de gerações até a convergência, a taxa de ocorrência 10 apresenta um resultado mais satisfatório, pois ela representa um aumento de cerca de 30%, em comparação à frequência 1, enquanto que a qualidade da solução se mantém. Logo, a taxa igual a 10 foi selecionada para os próximos testes.

4.4 Testes com BCG

Com o objetivo de verificar e validar a eficiência do algoritmo BCG em relação à classificação de polaridade de documentos, testes foram realizados para analisar os valores de qualidade atingidos pelo método nas bases de dados descritas anteriormente.

Os testes foram executados de modo a comparar o algoritmo desenvolvido com: classificadores com técnica TF-IDF; estratégias tradicionais e estratégias meta-heurísticas de seleção de atributos, como a BC e o AG. A partir disto, é possível identificar quais foram as melhorias e o quanto o método desenvolvido melhorou o processo de classificação.

Inicialmente, é necessário estabelecer quais foram os parâmetros adotados para a execução do algoritmo. A partir dos estudos da literatura (SHEHAB; KHADER; AL-BETAR, 2017) e de testes feitos, os melhores valores foram selecionados e estão exibidos na Tabela 5.

Tabela 5 - Parâmetros do algoritmo BCG

Parâmetro	**Valor**
Tamanho da população (P)	100
Porção de piores ninhos (P_a)	10(%)
Escala do salto (α)	1
Controle da perturbação (β)	2
Probabilidade de cruzamento (P_c)	70(%)
Probabilidade de mutação (P_m)	10(%)
Número máximo de gerações	1000
Número máximo de gerações sem melhoria	50
Frequência de aplicação de AG	10

Fonte: Elaborado pelo autor

Além disso, é interessante apresentar os resultados de acurácia base das quatro bases de dados para averiguar quais classificadores apresentam os melhores resultados. Tais valores obtidos somente com o método TF-IDF estão ilustrados na Tabela 6. Com isto, é possível afirmar que os melhores classificadores são o ME e o SVM, mas o algoritmo ME se sobressai na maioria dos casos, além de ser o mais rápido dos quatro classificadores.

As bases de dados A e B apresentaram os maiores valores de acurácia, enquanto C e D apresentaram os piores. Isto mostra que não há uma relação entre os dados traduzidos e os que são originalmente do idioma inglês. Por outro lado, o contexto dos dados parece influenciar negativamente a qualidade dos resultados em relação ao contexto de transporte aéreo.

Tabela 6 - Acurácia base com quatro com classificadores

Base	ME (%)	NB (%)	SVM (%)	RF (%)
A	**73,77** $\pm$ 2,20	60,66 $\pm$ 1,13	73,33 $\pm$ 2,16	71,77 $\pm$ 1,37
B	70,77 $\pm$ 2,59	58,11 $\pm$ 1,51	**72,44** $\pm$ 3,07	72,33 $\pm$ 1,96
C	**52,55** $\pm$ 1,24	49,44 $\pm$ 3,04	**52,55** $\pm$ 2,77	50,00 $\pm$ 2,30
D	**62,00** $\pm$ 1,77	60,07 $\pm$ 2,35	60,44 $\pm$ 2,54	59,70 $\pm$ 1,48

Fonte: Elaborado pelo autor

4.4.1 Comparações das funções de aptidão

Com isto, o algoritmo BCG foi executado com tais parâmetros e com as configurações adicionais descritas no capítulo anterior. Inicialmente, o objetivo deste teste é realizar uma entre as funções de aptidão denominadas f_1 e f_2, f_1 leva em consideração apenas a acurácia do modelo encontrado e $f_2$considera a acurácia e o inverso da quantidade de atributos selecionados, isto é, a Equação (13). Tal teste considerou quatro algoritmos classificadores executados com a base de dados A. Isto foi feito com o intuito de avaliar qual função de aptidão seria utilizada nos próximos testes.

Na Tabela 7 estão descritos os resultados obtidos na base de dados A por meio da execução do método desenvolvido com f_1. A função de aptidão utilizou a acurácia do modelo de acordo com o classificador ME que apresentou os melhores resultados iniciais. Como se trata de uma média, os valores apresentados possuem os seus respectivos valores de desvio-padrão ao lado.

Nesta mesma linha, os valores relativos à quantidade de atributos selecionados pelo método juntamente com a acurácia obtida pelo classificador ME e a porcentagem de redução

estão disponíveis na Tabela 8, em que os valores apresentados em relação aos cinco estados mencionados. Os valores destacados representam os melhores resultados em ambas as tabelas.

Tabela 7 – Acurácia com f_1 na base A com quatro classificadores

Classificador	Abordagem TF-IDF (%)	Abordagem TF-IDF+BCG (%)	Aumento médio na acurácia (%)
ME	$73{,}77 \pm 2{,}20$	$\mathbf{84{,}31 \pm 2{,}35}$	**10,54**
NB	$60{,}66 \pm 1{,}13$	$64{,}20 \pm 4{,}21$	4,46
SVM	$73{,}33 \pm 2{,}16$	$78{,}77 \pm 2{,}65$	5,44
RF	$71{,}77 \pm 1{,}37$	$76{,}65 \pm 3{,}16$	4,88

Fonte: Elaborado pelo autor

Tabela 8 – Quantidade de atributos com f_1 na base A com classificador ME em cinco estados

Estado	Acurácia com ME e TF-IDF+BCG (%)	Abordagem TF-IDF	Abordagem TF-IDF+BCG	Atributos selecionados (%)
1	$84{,}38 \pm 1{,}52$	300	151	$50{,}29 \pm 1{,}94$
2	$80{,}88 \pm 0{,}90$	300	151	$50{,}06 \pm 2{,}66$
3	$\mathbf{86{,}61 \pm 1{,}12}$	300	147	$49{,}76 \pm 3{,}81$
4	$85{,}83 \pm 1{,}49$	300	151	$50{,}63 \pm 1{,}95$
5	$83{,}83 \pm 1{,}22$	300	147	$\mathbf{46{,}99 \pm 3{,}55}$

Fonte: Elaborado pelo autor

Por meio da análise da Tabela 7, é possível notar que a acurácia do método ME e SVM foram as maiores obtidas com o método TF-IDF, contudo, após a execução do algoritmo bioinspirado, o classificador ME conseguiu atingir um valor de acurácia mais discrepante em relação aos outros classificadores, o que resultou em um aumento de até 10,54% na acurácia do modelo.

De um modo geral, independente do método de classificação aplicado a posteriori, o algoritmo desenvolvido conseguiu elevar a acurácia média em pelo menos 4,88% quando técnicas tradicionais são levadas em consideração.

Adicionalmente, estes resultados confirmam hipóteses levantadas na subseção Seleção de atributos, em que o algoritmo criado apresenta resultados razoáveis de acurácia mesmo que o conjunto de dados possua idiomas e contextos distintos.

Neste contexto, os resultados apresentados na Tabela 8 confirmam a estabilidade do modelo do ponto de vista de identificar soluções com aptidões similares com exceção do

estado 2 de configuração inicial dos dados, uma vez que a média da acurácia foi de aproximadamente 84,30%.

Além disso, o algoritmo chegou a uma redução de atributos bastante similar em relação aos estados iniciais dos dados, porque é possível verificar que a quantidade de atributos selecionados ficou, em média, em 49,54% com 3,71% de desvio-padrão.

Somado a isto, é visível que a quantidade de atributos parece não ter uma grande interferência na acurácia, pois o maior valor de acurácia calculado no estado 3 foi encontrado com 147 atributos selecionados, ao passo que o menor valor de acurácia do estado 2 foi encontrado com 151 atributos, em média.

A fim de analisar possíveis melhorias na função de aptidão e forçar uma maior redução dos atributos selecionados, testes foram executados de modo a se levar em consideração a função de aptidão f_2 definida anteriormente.

Na Tabela 9 estão expostos os resultados de acurácia dos quatro classificadores em relação à função f_2, e na Tabela 10 estão os resultados relativos à diminuição dos atributos selecionados.

Tabela 9 - Acurácia com f_2 em quatro classificadores

Classificador	Abordagem TF-IDF (%)	Abordagem TF-IDF+BCG (%)	Aumento médio na acurácia (%)
ME	$73{,}77 \pm 2{,}20$	**$84{,}58 \pm 2{,}32$**	**10,81**
NB	$60{,}66 \pm 1{,}13$	$63{,}40 \pm 4{,}82$	2,74
SVM	$73{,}33 \pm 2{,}16$	$79{,}13 \pm 3{,}03$	5,80
RF	$71{,}77 \pm 1{,}37$	$77{,}23 \pm 2{,}94$	5,46

Fonte: Elaborado pelo autor

Desta forma, é possível observar que os valores de acurácia são bastante similares ao caso anterior, de forma que o algoritmo classificador ME ainda apresenta a maior acurácia e o maior aumento médio deste parâmetro.

Os classificadores SVM e RF tiveram valores um tanto maiores, enquanto que a acurácia do método NB diminuiu. No geral, é possível concluir que não houve uma melhora significativa, pelo menos em termos relativos aos valores de acurácia, apesar da pequena melhora em valores absolutos.

Tabela 10 - Quantidade de atributos com f_2 com classificador ME em cinco estados

Estado	Acurácia com ME e TF-IDF+BCG (%)	Abordagem TF-IDF	Abordagem TF-IDF+BCG	Atributos selecionados (%)
1	$84{,}72 \pm 1{,}03$	300	**105**	**$35{,}29 \pm 3{,}47$**
2	$80{,}94 \pm 1{,}13$	300	118	$39{,}56 \pm 2{,}46$
3	**$86{,}66 \pm 0{,}99$**	300	117	$39{,}03 \pm 2{,}83$
4	$86{,}00 \pm 1{,}68$	300	116	$38{,}70 \pm 4{,}14$
5	$84{,}61 \pm 1{,}13$	300	106	$35{,}43 \pm 2{,}27$

Fonte: Elaborado pelo autor

Por outro lado, houve uma mudança evidente na quantidade de atributos selecionados. Enquanto a média de atributos escolhidos no caso de f_1 foi de 49,54%, a média no caso de f_2 foi de 37,60%, o que representa uma boa melhora sobre este aspecto. Por meio da Tabela 10 é possível informar que os estados iniciais 1 e 5 foram os melhores para diminuir a quantidade de características selecionadas. Contudo, os estados iniciais 3 e 4 obtiveram os maiores valores médios de acurácia, isto pode indicar que para atingir os maiores valores de acurácia, ainda é necessária uma quantidade de atributos próxima de 40% do total nesta base de dados.

Deste modo, é importante verificar qual das funções é mais adequada para os próximos casos. Primeiramente, é sabido que os valores de acurácia são similares, logo este critério não é considerado. Em segundo lugar, devido ao fato de que o número de atributos selecionados com f_2 foi consideravelmente menor, tal função poderia ser interessante, especialmente em casos em que o número de atributos totais é grande.

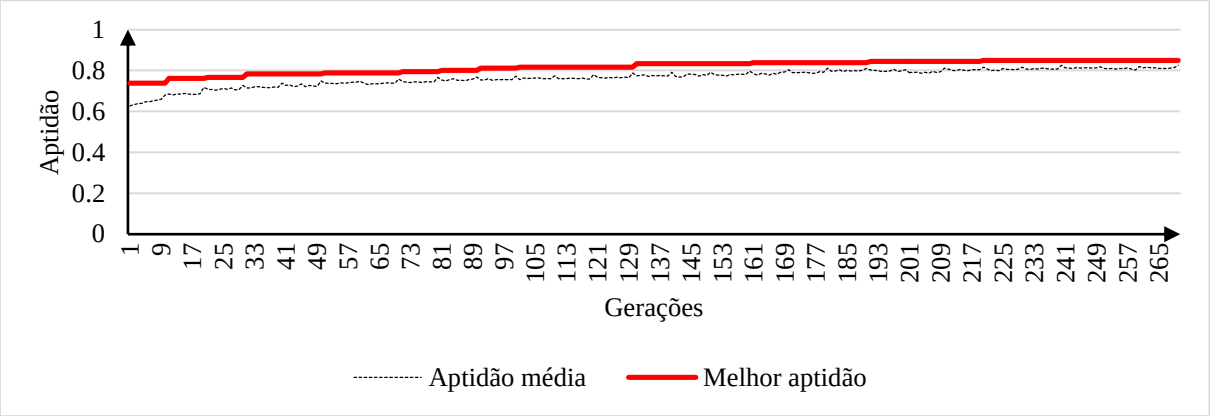
Porém, o tempo de execução e o número de gerações não foram levados em consideração nas análises, assim, é importante verificar quais foram, de modo que a escolha da melhor função fica a cargo do especialista e do tomador de decisão.

Com isto, é interessante analisar qual o comportamento da população com ambas as funções, de modo a averiguar a velocidade de convergência. Em primeiro lugar, é necessário avaliar o comportamento de f_1 e de f_2, isto está representado na Figura 16 e Figura 17, respectivamente, de forma que representam uma execução arbitrária com estado inicial 1.

A convergência é bastante similar e a população apresenta uma tendência de chegar cada vez mais próxima da melhor solução conforme avança, com a diferença de a execução com f_2 apresenta mais melhorias por conta de que a redução no número de atributos também aumenta sua aptidão.

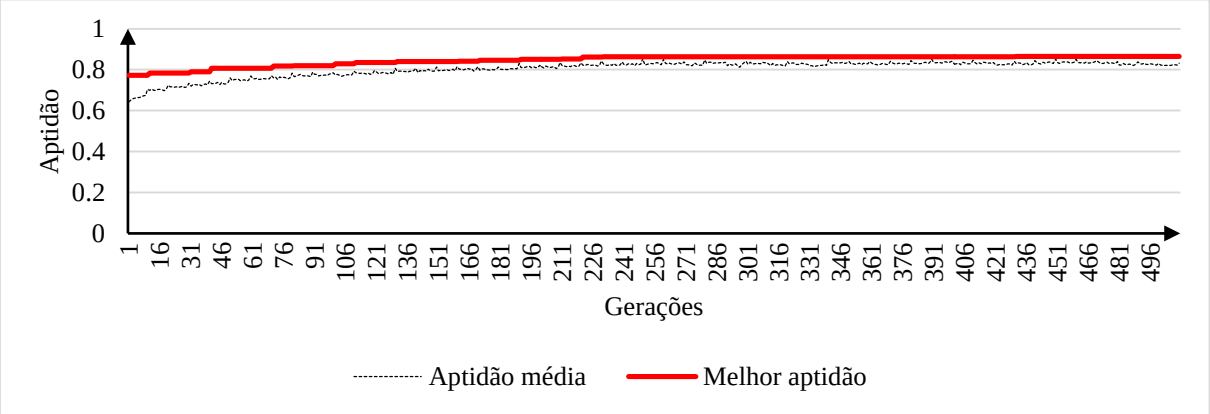
Apesar disto, a principal diferença está no número de gerações que salta de 269 para 509, isto representa um aumento de 89,21% de f_2 em relação à f_1, ao passo que a acurácia final em ambos os casos é a mesma.

Figura 16 - Desempenho da BCG com f_1



Fonte: Elaborado pelo autor

Figura 17 - Desempenho da BCG com f_2



Fonte: Elaborado pelo autor

Outro fator importante para se considerar é o tempo de execução resultante a partir de cada execução. Os resultados relativos ao tempo de execução, em segundos, estão resumidos na Tabela 11. Por meio da análise desta, é notável que a quantidade de gerações interferiu fortemente no tempo de execução.

Como mostrado na Tabela 11, a maior diferença entre as funções ocorreu no estado inicial 1, em que o aumento no tempo de execução ultrapassa cerca de 50 minutos, ao passo que a menor ocorre no estado 5 com um aumento de aproximadamente 18 minutos. Com base na média, é possível afirmar que a melhoria foi de aproximadamente 32 minutos, isto é, a função f_1 consegue atingir os mesmos valores de acurácia em cerca de 23 minutos, ao passo que f_2 precisa de 55 minutos, aproximadamente.

Tabela 11 – Comparação de tempo de execução das funções f_1 e f_2 com BCG

Estado	Método TF-IDF+BCG com f_1 (s)	Método TF-IDF+BCG com f_2 (s)	Aumento médio no tempo de execução (s)
1	1.545,22 ± 314,69	4.615,75 ± 4.470,00	3.070,53
2	1.372,52 ± 356,33	3.174,46 ± 1.754,77	1.801,94
3	1.386,87 ± 412,90	3.312,20 ± 1.139,49	1.925,33
4	1.329,06 ± 347,43	3.142,50 ± 899,80	1.813,44
5	1.288,89 ± 363,94	2.400,90 ± 869,17	1.112,01
Média	1.384,51 ± 370,90	3.329,16 ± 2.387,45	1.944,65

Fonte: Elaborado pelo autor

É válido ressaltar que a primeira função também possui uma maior estabilidade em comparação com f_2, de acordo com o desvio-padrão apresentado. Também é possível afirmar que não existe um comportamento linear neste algoritmo, já que o tempo de execução não cresce linearmente conforme o número de gerações sobe.

Logo, cabe ao tomador de decisão optar por uma estratégia que atinge bons valores de acurácia em menor tempo, mas não se preocupa em reduzir fortemente o número de atributos, ou então, optar pela abordagem que atinge os mesmos valores de acurácia, porém demanda um tempo de execução maior porque se propõe a encontrar o menor subconjunto possível de atributos. Nos seguintes testes, a função f_1 foi utilizada como base de comparação.

Finalmente, é importante apresentar quais são os resultados obtidos por meio do algoritmo BCG em todas as bases de dados, e assim analisar as suas implicações. Os valores de acurácia obtidos com a função f_1 estão expostos na Tabela 12, e a quantidade média de características selecionadas está apresentada na

Tabela 13.

Tabela 12 - Acurácia do método BCG com f_1 nos quatro classificadores

Base	ME (%)	NB (%)	SVM (%)	RF (%)
A	**84,31 ± 2,35**	64,20 ± 4,21	78,77 ± 2,65	76,65 ± 3,16
B	**85,04 ± 2,03**	57,88 ± 2,54	77,84 ± 2,61	72,66 ± 3,32
C	**69,86 ± 2,44**	55,80 ± 3,54	60,57 ± 3,19	58,68 ± 4,24
D	**74,16 ± 1,97**	65,24 ± 3,23	68,91 ± 2,39	66,40 ± 3,87

Fonte: Elaborado pelo autor

Tabela 13 - Quantidade de atributos selecionados com f_1 nos quatro classificadores

Base	Abordagem TF-IDF	Abordagem TF-IDF+BCG	Atributos selecionados (%)
A	300	147	$49{,}54 \pm 3{,}17$
B	346	148	$43{,}26 \pm 2{,}96$
C	327	134	$41{,}29 \pm 2{,}90$
D	453	217	$48{,}02 \pm 2{,}82$

Fonte: Elaborado pelo autor

Os valores de acurácia da Tabela 12 mostram a evolução da taxa de acerto em comparação ao método base. O destaque é o classificador ME que atingiu os maiores resultados em todas as bases de dados. Observa-se, portanto, um crescimento nos valores de acurácia de aproximadamente 10,54%, 12,60%, 17,31% e 12,16% em relação às bases A, B, C e D, respectivamente.

Em relação à seleção de atributos, os resultados da

Tabela 13 mostram uma média de 45,52% de uso de atributos, isto é, houve uma redução de 54,48% na quantidade de características necessárias para classificar os documentos. Devido ao fato de que os valores de seleção são próximos, é possível afirmar que o método BCG manteve sua estabilidade independente da base de dados. As reduções em relação à quantidade de atributos original variaram de 91% a 92%.

4.4.2 Comparação com métodos determinísticos

Na literatura existem alguns algoritmos de seleção de atributos como os métodos filtro e *wrapper*, tais métodos são normalmente empregados para auxiliar processos de classificação. Neste caso, dois métodos de seleção de atributos foram selecionados e os resultados foram comparados com o algoritmo desenvolvido.

Os algoritmos são: o método *K-Best* com seleção por meio de teste ANOVA e o método de eliminação de atributos recursivo. A primeira abordagem é um método filtro que associa um valor a cada atributo e somente aqueles com valores mais elevados são selecionados. A segunda se trata de uma técnica *wrapper* que utiliza remoção de características. Tal método é conhecido como *recursive feature elimination* (RFE), ou eliminação de atributos recursiva do inglês. Ele consiste em remover os atributos menos

relevantes até que atinja um número de características pré-estabelecido. Isto é feito por meio de um algoritmo classificador, neste caso, o algoritmo usado foi o ME.

Os testes necessitam, neste caso, que o número de atributos a serem selecionados seja fornecido previamente. Assim, três valores prévios foram definidos arbitrariamente, com base nos resultados obtidos nos testes anteriores, de modo que estes valores são aproximadamente 33%, 50% e 66% do total de atributos iniciais. Somado a isto, cada porcentagem de atributos foi verificada com cada um dos quatro classificadores. Os resultados destacados indicam a maior acurácia com a determinada quantidade de atributos. O algoritmo que atingiu os melhores resultados, em geral, foi o classificador ME. Entretanto, é informado no teste em questão qual algoritmo atingiu o melhor valor de acurácia.

A seguir estão apresentados os resultados associados à acurácia do melhor classificador que cada método de seleção atingiu. Os resultados do método filtro *K-best* estão na Tabela 14; e os resultados do método RFE estão na Tabela 15.

Tabela 14 - Acurácia (%) com método filtro *K-Best*

Base	Quantidade de atributos selecionados		
	33%	50%	66%
A	**79,22 ± 2,18**	78,11 ± 1,70	77,88 ± 1,28
B	76,33 ± 1,51	76,88 ± 1,97	**77,00 ± 2,44***
C	**63,33 ± 4,37**	60,88 ± 1,43	60,00 ± 1,53
D	62,37 ± 3,47**	**69,33 ± 3,31****	68,22 ± 1,77**

*obtido por meio do SVM

**obtido por meio do NB

Fonte: Elaborado pelo autor

Tabela 15 - Acurácia (%) com método *wrapper* RFE

Classificador	Quantidade de atributos selecionados		
	33%	50%	66%
A	78,44 ± 2,41	**79,11 ± 2,44**	78,22 ± 2,36
B	77,44 ± 1,74	**78,11 ± 1,94**	76,66 ± 2,30*
C	**62,66 ± 2,81**	61,33 ± 1,08	59,00 ± 1,83*
D	66,51 ± 2,37	**68,22 ± 1,39**	67,25 ± 1,63

*obtido por meio do SVM

Fonte: Elaborado pelo autor

Com base nestes resultados, foi possível averiguar que os métodos *K-Best* e RFE chegaram a resultados similares. Na maioria dos casos o algoritmo ME atingiu os melhores valores, enquanto que os algoritmos SVM e NB o ultrapassaram poucas vezes. É possível afirmar, portanto, que não houve um classificador ideal, assim como é visível que houve uma melhoria significativa com os métodos de redução tradicionais. Em geral, a maioria dos melhores resultados pertence à técnica *K-Best*.

Isto vai ao encontro com os resultados apresentados na subseção anterior, em que a porcentagem de atributos final do método elaborado chegou a valores em torno de 33% e 50% de atributos selecionados. No entanto, os valores de acurácia ainda ficam abaixo dos valores encontrados pelo algoritmo BCG, mesmo no caso de métodos *wrapper* que usam o classificador ME como estratégia de ajuste.

Com o intuito de integrar as informações aqui apresentadas e comparar com os valores obtidos com a técnica BCG, os melhores resultados foram selecionados juntamente com a melhoria obtida, isto está ilustrado na Tabela 16.

Tabela 16 – Valores de aumento na acurácia com métodos tradicionais

Base	Aumento com abordagem *K-Best* (%)	Aumento com abordagem RFE (%)	Aumento com abordagem BCG (%)
A	5,45	5,34	10,54
B	4,56*	5,67	12,60
C	10,78	10,11	17,31
D	7,33**	6,22	12,16

*obtido por meio do SVM

**obtido por meio do NB

Fonte: Elaborado pelo autor

Por meio desta tabela, é possível afirmar que a abordagem BCG possui uma vantagem visível em relação aos métodos tradicionais. O método *K-Best* apresenta uma ligeira melhoria em relação ao RFE, porém ambos os métodos são superados pelo BCG.

Tais resultados reforçam a ideia de que o ambiente com dados reais, apesar de ser relativamente pequeno, já representa uma grande dificuldade para técnicas tradicionais que, apesar de apresentar uma melhoria interessante, ainda são inferiores a abordagens recentes.

Em geral, a melhoria apresentada pelo método BCG chega a dobrar o valor de acurácia nas bases A, B e D, o que mostra a importância de se implementar os métodos bioinspirados.

4.4.3 Comparação com métodos meta-heurísticos

Finalmente, é importante realizar uma comparação entre o algoritmo desenvolvido e as técnicas meta-heurísticas AG e BC, de forma a identificar quais são as vantagens de se mesclar estes métodos e qual o ganho de fato. A função de aptidão escolhida foi a f_1 por conta do seu desempenho mais veloz em relação à f_2, logo f_1 foi aplicada em todos os algoritmos meta-heurísticos. Tal comparação visa identificar, inicialmente, quais os valores de acurácia e o número de atributos selecionados que os algoritmos conseguiram atingir nas bases de dados alvo, de modo a se considerar os cinco estados iniciais.

Desta forma, o primeiro método a ser comparado é a BC. Este método se mostra eficaz para análise de sentimentos segundo a literatura em comparação com outras técnicas (YADAV; VISHWAKARMA, 2020), ao passo que ele também apresenta a limitação de ficar preso em ótimos locais, logo tende a encontrar uma solução não tão boa de forma rápida. É possível visualizar os resultados relativos à acurácia por meio da Tabela 17, enquanto que as outras métricas como quantidade de atributos selecionados e tempo de execução em cada base de dados estão disponíveis na Tabela 18.

Tabela 17 – Comparação de acurácia com método ME entre BC e BCG

Bases	Abordagem TF-IDF (%)	Acurácia com método BC (%)	Acurácia com BCG (%)	Diferença entre BCG e BC
A	$73{,}77 \pm 2{,}20$	$76{,}58 \pm 2{,}41$	$84{,}31 \pm 2{,}35$	7,73
B	$70{,}77 \pm 2{,}59$	$75{,}74 \pm 2{,}44$	$85{,}04 \pm 2{,}03$	9,30
C	$52{,}55 \pm 1{,}24$	$56{,}66 \pm 1{,}49$	$69{,}86 \pm 2{,}44$	13,20
D	$62{,}00 \pm 1{,}77$	$63{,}08 \pm 2{,}11$	$74{,}16 \pm 1{,}97$	11,08

Fonte: Elaborado pelo autor

Tabela 18 – Métricas de qualidade e desempenho entre BC e BCG

Bases	Atributos selecionados com BCG (%)	Atributos selecionados com BC (%)	Tempo de execução com BCG (s)	Tempo de execução com BC (s)
A	$49{,}54 \pm 3{,}17$	$46{,}51 \pm 5{,}29$	$1.384{,}51 \pm 370{,}90$	$495{,}06 \pm 289{,}83$
B	$43{,}26 \pm 2{,}96$	$39{,}34 \pm 3{,}87$	$1.682{,}14 \pm 560{,}30$	$924{,}29 \pm 406{,}79$
C	$41{,}29 \pm 2{,}90$	$38{,}00 \pm 2{,}81$	$2.036{,}84 \pm 586{,}11$	$517{,}66 \pm 198{,}60$
D	$48{,}02 \pm 2{,}82$	$42{,}48 \pm 3{,}83$	$5.061{,}28 \pm 1.124{,}63$	$1.407{,}01 \pm 623{,}71$

Fonte: Elaborado pelo autor

A partir destas informações, foi possível confirmar que o algoritmo BC melhorou a qualidade da classificação em todas as bases de dados. No entanto, a melhoria é pouco expressiva uma vez que os métodos tradicionais e o algoritmo desenvolvido apresentaram valores melhores de acurácia. Como apresentado pelos dados presentes na Tabela 17, o método BC atingiu valores cuja diferença de acurácia do método BCG varia de 7% até 13%, aproximadamente. Assim, estes valores de acurácia não justificam sua utilização.

Por outro lado, é importante destacar que a quantidade de atributos selecionados do algoritmo BC foi melhor. Os resultados confirmam que há uma diminuição de 3% até 6%, em valores aproximados, o que significa que a redução de atributos responde melhor ao algoritmo BC. Apesar disto, a acurácia em cada base de dados não melhorou tanto quanto poderia, de modo que a função f_2 também atingiu bons valores de acurácia com poucos atributos usados.

Somado a isto, o tempo de execução da BC foi o mais baixo entre os métodos bioinspirados. Em média, houve um aumento de 182% até 360% no tempo de execução, a depender da base de dados. A maior diferença ocorreu na base D, em que o método BCG levou cerca 84 minutos para executar, enquanto o método BC levou 23. Isto revela que uma execução mais rápida não necessariamente chega a bons valores de qualidade, assim como reforça a ideia de que há uma conversão prematura neste algoritmo, e o algoritmo não identificou melhores soluções somente com sua configuração básica.

A seguir, o AG foi devidamente implementado por meio dos operadores genéticos descritos no Capítulo 2 com o objetivo de verificar se a abordagem BCG se apresentaria benéfica em relação a ele. O algoritmo foi desenvolvido com tais premissas e os resultados relativos à acurácia podem ser vistos na Tabela 19, enquanto que o restante das métricas está representado na Tabela 20.

Tabela 19 – Comparação de acurácia com método ME entre AG e BCG

Bases	Abordagem TF-IDF (%)	Acurácia com método AG (%)	Acurácia com BCG (%)	Diferença entre BCG e AG
A	$73{,}77 \pm 2{,}20$	$84{,}74 \pm 2{,}22$	$84{,}31 \pm 2{,}35$	-0,43
B	$70{,}77 \pm 2{,}59$	$85{,}46 \pm 1{,}97$	$85{,}04 \pm 2{,}03$	-0,42
C	$52{,}55 \pm 1{,}24$	$70{,}04 \pm 2{,}53$	$69{,}86 \pm 2{,}44$	-0,18
D	$62{,}00 \pm 1{,}77$	$74{,}27 \pm 2{,}24$	$74{,}16 \pm 1{,}97$	-0,11

Fonte: Elaborado pelo autor

Tabela 20 – Métricas de qualidade e desempenho entre AG e BCG

Bases	Atributos selecionados com BCG (%)	Atributos selecionados com AG (%)	Tempo de execução com BCG (s)	Tempo de execução com AG (s)
A	49,54 ± 3,17	50,58 ± 3,45	1.384,51 ± 370,90	2.289,89 ± 939,60
B	43,26 ± 2,96	44,47 ± 2,15	1.682,14 ± 560,30	3.497,47 ± 1.317,81
C	41,29 ± 2,90	43,82 ± 2,16	2.036,84 ± 586,11	3.223,18 ± 1.195,29
D	48,02 ± 2,82	49,75 ± 1,90	5.061,28 ± 1.124,63	8.239,48 ± 2.693,48

Fonte: Elaborado pelo autor

Diferentemente da BC, os valores obtidos por meio do AG relevam que esta técnica é bastante promissora em termos de qualidade, pois é possível notar que todos os resultados de acurácia foram maiores, em média, em todas as bases, como mostrado na Tabela 19. No entanto, é válido ressaltar que os valores estão bastante próximos dos valores atingidos pela BCG, de forma que as maiores diferenças ocorrem nas bases de dados A e B com vantagem para o AG de 0,43% e 0,42%, respectivamente.

Por meio da análise da Tabela 20, a quantidade de atributos selecionados ficou bastante próxima em todas as bases de dados, de modo que há um aumento de aproximadamente 1% com o método AG. Se considerássemos apenas as métricas de qualidade, o AG seria o mais interessante, no entanto, o tempo de execução do AG é de 158% até 207% maior que o tempo de execução do método BCG. A execução mais demorada ocorreu na base D, em que o BCG atingiu o critério de parada com cerca de 84 minutos, enquanto que o AG levou 137 minutos.

A fim de verificar se havia diferenças estatísticas nas amostragens dos algoritmos AG e BCG, pois os seus valores de acurácia estão bastante próximos, testes estatísticos foram aplicados para averiguar isto. A hipótese de nulidade considerada foi a seguinte: não há diferença de acurácia entre as abordagens e a hipótese alternativa considera que há diferença.

Os testes estatísticos foram realizados pela ferramenta BioStat[10] e os resultados estão ilustrados na Tabela 21. Devido ao fato de que os valores obtidos pelo AG não apresentaram distribuição normal, mas ainda apresentam certa homogeneidade. Assim, um teste paramétrico e um não-paramétrico foram utilizados, eles são os testes “Student” e “Mann-Whitney” e o valor de α considerado foi de 5%.

[10] https://www.mamiraua.org.br/downloads/programas/

Tabela 21 – Testes estatísticos sobre acurácia entre AG e BCG

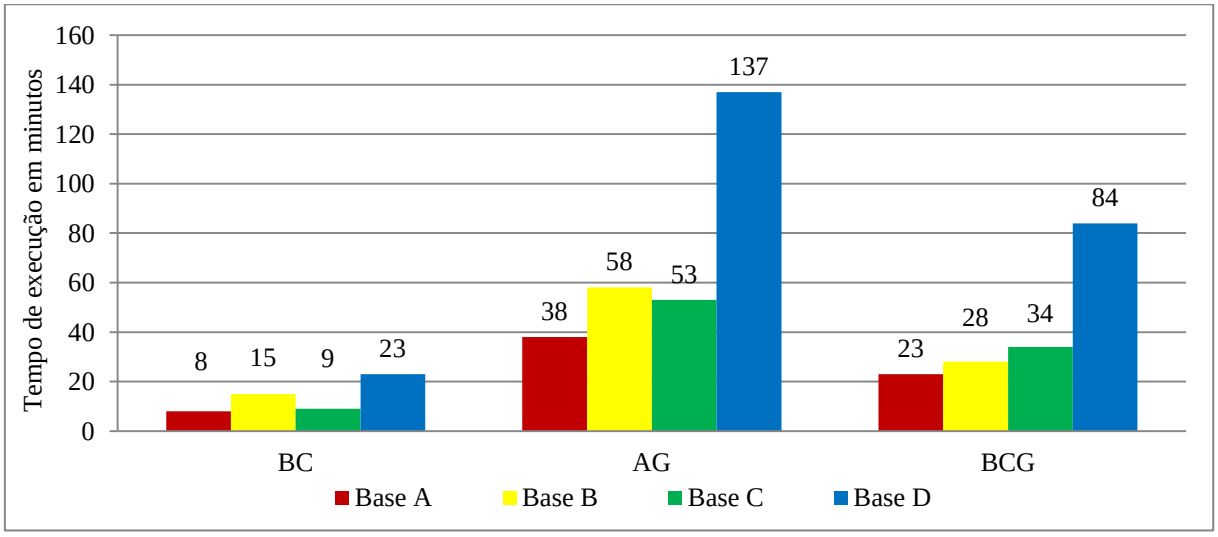
Bases	Teste Student			Teste Mann-Whitney		
	t	p-valor (unilateral)	p-valor (bilateral)	U	p-valor (unilateral)	p-valor (bilateral)
A	-0,9376	0,1754	0,3507	1092,00	0,1380	0,2761
B	-0,8429	0,2010	0,4020	540,00	0,1697	0,3394
C	-0,2476	0,4028	0,8055	303,00	0,4269	0,8538
D	0,2606	0,3975	0,7950	1243,00	0,4808	0,9615

Fonte: Elaborado pelo autor

Como demonstrado por meio da Tabela 21, ambos os testes estatísticos inferem que não há diferença estatística entre os valores d acurácia obtidos pelos algoritmos AG e BCG. Logo, a hipótese de nulidade não é rejeitada. Isto implica que apesar dos valores de acurácia do método AG estarem nominalmente maiores, ambos os métodos chegaram ao mesmo resultado em relação à taxa de acerto.

Ao final, vale ressaltar que a maior diferença entre as técnicas se mostra no tempo de execução. Analisar o custo computacional das técnicas heurísticas é interessante para demonstrar qual técnica é mais apropriada. Para que seja possível visualizar o tempo de execução de cada algoritmo em cada base de dados, tais valores foram convertidos em minutos, de forma aproximada, e foram dispostos em uma mesma escala conforme apresentado por meio da Figura 18.

Figura 18 – Comparação do desempenho da BCG com BC e AG



Fonte: Elaborado pelo autor

Desta forma, é visível que o tempo de execução é bastante variado a depender da abordagem meta-heurística. A BC possui os menores tempos, de modo que sua execução mais demorada levou 23 minutos aproximadamente, ao contrário do AG levou, em média, 137 minutos para finalizar sua execução na base D. Devido ao fato de que a BC apresentou uma dificuldade considerável em sair de ótimos locais, acabou por convergir prematuramente, de forma que sua execução rápida resultou nos piores valores de acurácia.

Em contrapartida, o AG apresentou os melhores resultados em termos de acertos na classificação, porém também apresentou os maiores tempos de execução. Em comparação com a BCG, o AG apresentou um tempo de execução maior com 65%, 107%, 55% e 63%, aproximadamente, em relação às bases de dados A, B, C e D, respectivamente.

Com isto, é possível afirmar que a BCG, que atingiu os mesmos valores de acurácia em termos estatísticos, é o método mais balanceado entre as três opções, de modo que não é o mais rápido, mas atinge os melhores resultados com um tempo de execução médio. Além disso, a BCG também atingiu os valores médios da quantidade de atributos selecionados, de forma que estes valores podem ser mais reduzidos se a função de aptidão f_2 for utilizada.

4.5 Considerações finais sobre os testes

Diversos experimentos foram realizados para analisar métricas de qualidade e de desempenho do algoritmo desenvolvido. Por meio dos resultados foi possível averiguar o comportamento da abordagem proposta, e ainda verificar o comportamento de duas funções de aptidão, assim como foi possível analisar outros algoritmos empregados nesta área e confirmar as informações apresentadas na literatura.

Com isto, foi constatado que o algoritmo desenvolvido conseguiu atingir um equilíbrio entre buscar novas soluções, ao mesmo tempo em que também lida com as boas soluções previamente encontradas para gerar novas, quando comparado às técnicas citadas neste trabalho. É observável que o BCG se apresenta como uma estratégia balanceada entre os métodos meta-heurísticos.

No entanto, ao se considerar o tempo de execução, o elevado custo computacional do AG em comparação à BCG se sobressai, o que mostra a efetividade do modelo desenvolvido. Além disso, ao se adicionar o pré-processamento descrito, este que considera tanto diferentes contextos quanto dois idiomas, o número de atributos sofre uma redução significativa e uma análise em um conjunto de dados com diversas particularidades é executável com uma acurácia cerca 12% maior do que sem a seleção de atributos.

Capítulo 5

Conclusão

A análise de sentimentos é uma área cuja importância é crescente e suas aplicações se ampliam conforme a sociedade evolui e se torna cada vez mais necessário entender opiniões e sentimentos de pessoas em relação aos mais variados tópicos. Este estudo pode auxiliar áreas comerciais, políticas e de biociências ao analisar o comportamento de usuários. Porém, os desafios existentes relacionados, dentre outras coisas, ao volume, à variedade, ao valor dos dados, e especialmente à quantidade de atributos a serem analisados são pontos cruciais a serem tratados, pois lacunas na literatura são resolvidas e geradas a cada novo trabalho.

Nos últimos anos, diversas estratégias inspiradas na natureza foram propostas para resolver este desafio que estratégias tradicionais não são capazes de lidar, assim como não pode ser feito por humanos porque excede sua capacidade de compreensão. Tais estratégias, em particular, foram aplicadas a diversas áreas como educação, robótica, finanças, diagnóstico de doenças, agricultura e até previsão do tempo, mas somente no período recente é que estes métodos passaram a ser utilizados no domínio de análise de sentimentos devido ao seu potencial de reduzir os problemas associados a isto.

Com isto, diversas técnicas bioinspiradas como AG, PSO, ACO, BC, dentre outros foram utilizadas amplamente para problemas de seleção de atributos, classificação e agrupamento com dados em inglês e de contextos únicos, em sua maioria.

No entanto, os estudos presentes na literatura geralmente não possuem foco em trabalhos que lidem diretamente com diferentes idiomas, diferentes contextos e que ainda proponham uma abordagem inspirada pela natureza que seja capaz de lidar com estes desafios e ainda atingir valores de qualidade semelhantes. Assim, o objetivo deste trabalho foi buscar contornar estes problemas de modo a apresentar um modelo que realize o tratamento inicial nos dados e ainda fazer um algoritmo que tenha um bom equilíbrio de exploração e explotação de soluções, a fim de encontrar um bom resultado em comparação a outros métodos tradicionais e meta-heurísticos.

Na fundamentação teórica foram apresentados os conceitos relacionados à análise de sentimentos, em particular, foram apresentados os possíveis estudos e métodos de análise, assim como definiu o problema da maldição da dimensionalidade e sua necessidade se ser processado por meios não tradicionais. Dois métodos meta-heurísticos aplicados neste contexto foram descritos. O trabalho, então, apresentou o modelo desenvolvido, este se divide em uma parte voltada para o pré-processamento e outra parte voltada para o algoritmo inspirado pela natureza que mescla as técnicas de AG e BC.

O foco deste trabalho está na confecção de um método multi-idiomas que considera o contexto dos dados textuais, e que realiza a seleção de atributos com um método híbrido inspirado na natureza que se mostra interessante para a área de AS de mídias sociais.

5.1 Contribuição científica

A contribuição científica deste trabalho se baseia na abordagem que une tanto um pré-processamento que considera dois idiomas, diferentes contextos, quanto um algoritmo inspirado pela natureza que mescla os benefícios das técnicas BC e AG. São observadas na Tabela 22 as principais características dos trabalhos correlatos e que fazem parte deste. Como representado na Tabela 22, o presente trabalho fez uso de algoritmos de ML e algoritmos meta-heurísticos, além de aplicar técnicas de padronização dos dados com dados reais de mídias sociais, assim como feito pela maioria dos trabalhos correlatos. Somado a isto, o trabalho desenvolvido realizou uma classificação ternária que considerou uma análise multi-idiomas embasada em contexto, diferentemente do que a literatura produziu até então.

Por meio dos resultados obtidos, foi possível concluir que a BCG atingiu os melhores valores de acurácia em relação aos métodos tradicionais, tanto sem quanto com seleção de atributos. Com isto, é possível concluir que o algoritmo desenvolvido possui melhores qualidades em relação aos algoritmos BC e AG quando estes são aplicados individualmente.

Isto se deu por conta dos baixos valores de acurácia encontrados com a BC e o elevado tempo de execução do algoritmo AG, o que faz da abordagem proposta um meio balanceado e efetivo para resolver os desafios da seleção de atributos.

O estudo ainda revela que o método elaborado aumenta a acurácia média de 10% até 17%, enquanto que a redução de atributos pode variar entre 50% e 60%, a depender da função de aptidão escolhida. A redução de atributos pode atingir até 92% se considerarmos o conjunto de dados sem aplicação da técnica TF-IDF. Devido ao fato de que este trabalho considerou diferentes características nos dados textuais, a presença de um algoritmo de seleção de atributos mais eficiente fora necessária para se chegar a um resultado mais interessante em tempo hábil e com boa qualidade.

Tabela 22 – Comparação de aspectos de correlatos com o trabalho feito

	(AKHTAR et al., 2017)	(APPEL et al., 2018)	(KUMAR et al., 2019)	(KUMAR; JAISWAL et al., 2019)	(HASSONAH et al., 2020)	Este trabalho
Uso de ML	Sim	Sim	Sim	Sim	Sim	Sim
Uso de meta-heurística	Sim	Não	Sim	Sim	Sim	Sim
Pré-processamento	Sim	Sim	Sim	Sim	Sim	Sim
Cenário de mídias sociais	Não	Sim	Sim	Sim	Sim	Sim
Classificação ternária	Sim	Não	Não	Sim	Sim	Sim
Tratamento de contexto	Não	Não	Não	Não	Sim	Sim
Tratamento de diferentes idiomas	Não	Não	Não	Não	Não	Sim

Fonte: Elaborado pelo autor

5.2 Trabalhos futuros

Por fim, alguns possíveis trabalhos futuros incluem considerar outros idiomas na análise, assim como aplicar diferentes técnicas léxicas que façam uma redução inicial no número de atributos, como considerar sarcasmo, subjetividade e negação. Outros aspectos como análise multimodal, isto é, considerar outras características não textuais como imagens e vídeos ampliaria as possibilidades de análise de mais comentários de mídias sociais.

Em relação ao algoritmo, foram levadas em consideração apenas as principais operações genéticas e o básico da BC. Logo, recomenda-se aplicar diferentes estratégias de seleção, cruzamento e mutação, assim como aplicar diferentes formas de inicialização de dados, assim como aplicar técnicas que filtrem os atributos antes de submetê-los aos métodos heurísticos. É válido ressaltar que técnicas de paralelismo são interessantes, uma vez que o tempo de execução não foi o foco deste trabalho, mas considerar tal métrica pode ser bastante benéfico em casos de bases de dados massivas.

Referências

ABD EL AZIZ, M.; HASSANIEN, A. E.. Modified cuckoo search algorithm with rough sets for feature selection. **Neural Computing and Applications**, v. 29, n. 4, p. 925-934, 2018.

AHUJA, R. et al. The impact of features extraction on the sentiment analysis. **Procedia Computer Science**, v. 152, p. 341-348, 2019.

AKHTAR, M. S. et al. Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. **Knowledge-Based Systems**, v. 125, p. 116-135, 2017.

ALARIFI, A. et al. A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. **The Journal of Supercomputing**, v. 76, n. 6, p. 4414-4429, 2020.

APPEL, O. et al. A hybrid approach to the sentiment analysis problem at the sentence level. **Knowledge-Based Systems**, v. 108, p. 110-124, 2016.

_____. et al. Successes and challenges in developing a hybrid approach to sentiment analysis. **Applied Intelligence**, v. 48, n. 5, p. 1176-1188, 2018.

ARAÚJO, M.; PEREIRA, A.; BENEVENUTO, F. A comparative study of machine translation for multilingual sentence-level sentiment analysis. **Information Sciences**, v. 512, p. 1078-1102, 2020.

BALAZS, J. A.; VELÁSQUEZ, J. D. Opinion mining and information fusion: a survey. **Information Fusion**, v. 27, p. 95-110, 2016.

BATRINCA, B.; TRELEAVEN, P. C. Social media analytics: a survey of techniques, tools and platforms. **Ai & Society**, v. 30, n. 1, p. 89-116, 2015.

CAI, J. et al. Feature selection in machine learning: A new perspective. **Neurocomputing**, v. 300, p. 70-79, 2018.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers & Electrical Engineering**, v. 40, n. 1, p. 16-28, 2014.

CHANG, J. et al. Novel feature selection approaches for improving the performance of sentiment classification. **Journal of Ambient Intelligence and Humanized Computing**, p. 1-14, 2020.

CIRQUEIRA, D. et al. A literature review in preprocessing for sentiment analysis for Brazilian Portuguese social media. In: **2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)**. IEEE, 2018. p. 746-749.

DATAREPORTAL. GLOBAL SOCIAL MEDIA STATS. **Datareportal**, 2022. Disponível em: < https://bit.ly/3L38zFd >. Acesso em: 14 mar. 2022.

DE CASTRO, L. N. **Fundamentals of natural computing: basic concepts, algorithms, and applications**. CRC Press, 2006.

DE MOURA MENESES, A. A. et al. Application of Cuckoo Search algorithm to Loading Pattern Optimization problems. **Annals of Nuclear Energy**, v. 139, p. 107214, 2020.

EL ANSARI, O.; ZAHIR, J.; MOUSANNIF, H. Context-based sentiment analysis: a survey. In: **International Conference on Model and Data Engineering**. Springer, Cham, 2018. p. 91-97.

GHANI, N. A. et al. Social media big data analytics: A survey. **Computers in Human Behavior**, v. 101, p. 417-428, 2019.

HASSONAH, M. A. et al. An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter. **Knowledge-Based Systems**, v. 192, p. 105353, 2020.

HEMMATIAN, F.; SOHRABI, M. K. A survey on classification techniques for opinion mining and sentiment analysis. **Artificial Intelligence Review**, v. 52, n. 3, p. 1495-1545, 2019.

HERRICK, R. To the virgins, to make much of time. **The Complete Poetry of Robert Herrick**, v. 1, p. 13-14, 1963.

KANAGARAJ, G.; PONNAMBALAM, S. G.; JAWAHAR, N. A hybrid cuckoo search and genetic algorithm for reliability–redundancy allocation problems. **Computers & Industrial Engineering**, v. 66, n. 4, p. 1115-1124, 2013.

KRAIEM, M. B. et al. OLAP operators for social network analysis. **Cluster Computing**, p. 1-28, 2019.

KRAMER, O. **Genetic algorithm essentials**. Springer, 2017.

KO, N. et al. Identifying product opportunities using social media mining: application of topic modeling and chance discovery theory. **IEEE Access**, v. 6, p. 1680-1693, 2017.

KUMAR, A.; GARG, G. Systematic literature review on context-based sentiment analysis in social multimedia. **Multimedia tools and Applications**, p. 1-32, 2019.

KUMAR, A. et al. Sentiment analysis using cuckoo search for optimized feature selection on Kaggle tweets. **International Journal of Information Retrieval Research (IJIRR)**, v. 9, n. 1, p. 1-15, 2019.

KUMAR, A.; JAISWAL, A. Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter. **Multimedia Tools and Applications**, v. 78, n. 20, p. 29529-29553, 2019.

LI, J. et al. Feature selection: A data perspective. **ACM Computing Surveys (CSUR)**, v. 50, n. 6, p. 1-45, 2017.

LI, Z. et al. A survey on sentiment analysis and opinion mining for social multimedia. **Multimedia Tools and Applications**, v. 78, n. 6, p. 6939-6967, 2019.

LIMA, A. C. ES; DE CASTRO, L. N.; CORCHADO, J. M. A polarity analysis framework for Twitter messages. **Applied Mathematics and Computation**, v. 270, p. 756-767, 2015.

LIU, W.; WANG, J. A brief survey on nature-inspired metaheuristics for feature selection in classification in this decade. In: **2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)**. IEEE, 2019. p. 424-429.

MÄNTYLÄ, M. V.; GRAZIOTIN, D.; KUUTILA, M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. **Computer Science Review**, v. 27, p. 16-32, 2018.

MOALLA, I.; NABLI, A.; HAMMAMI, M. Integration of a multidimensional schema from different social media to analyze customers' opinions. In: **2017 11th International Conference on Research Challenges in Information Science (RCIS)**. IEEE, 2017. p. 391-400.

MUHAMMAD, Aminu; WIRATUNGA, Nirmalie; LOTHIAN, Robert. Contextual sentiment analysis for social media genres. **Knowledge-based** systems, v. 108, p. 92-101, 2016.

NAZIR, F. et al. Social media signal detection using tweets volume, hashtag, and sentiment analysis. **Multimedia Tools and Applications**, v. 78, n. 3, p. 3553-3586, 2019.

OKTAVIA, T. et al. The influence of social media to support learning process in higher education institution: A survey perspective. In: **2017 International Conference on ICT For Smart Society (ICISS)**. IEEE, 2017. p. 1-5.

OLIVEIRA, D. J. S.; BERMEJO, P. H. de S.; DOS SANTOS, P. A. Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. **Journal of Information Technology & Politics**, v. 14, n. 1, p. 34-45, 2017.

PANDEY, A. C.; RAJPOOT, D. S.; SARASWAT, M. Twitter sentiment analysis using hybrid cuckoo search method. **Information Processing & Management**, v. 53, n. 4, p. 764-779, 2017.

PANDEY, A. C.; RAJPOOT, D. S. Spam review detection using spiral cuckoo search clustering method. **Evolutionary Intelligence**, v. 12, n. 2, p. 147-164, 2019.

PANDEY, A. C.; RAJPOOT, D. S.; SARASWAT, M. Feature selection method based on hybrid data transformation and binary binomial cuckoo search. **Journal of Ambient Intelligence and Humanized Computing**, v. 11, n. 2, p. 719-738, 2020.

PEREIRA, D. A. A survey of sentiment analysis in the Portuguese language. **Artificial Intelligence Review**, p. 1-29, 2020.

RODRIGUES, D. et al. BCS: A binary cuckoo search algorithm for feature selection. In: **2013 IEEE International Symposium on Circuits and Systems (ISCAS)**. IEEE, 2013. p. 465-468.

ROUT, J. K. et al. A model for sentiment and emotion analysis of unstructured social media text. **Electronic Commerce Research**, v. 18, n. 1, p. 181-199, 2018.

SHARMA, M.; KAUR, P. A Comprehensive Analysis of Nature-Inspired Meta-Heuristic Techniques for Feature Selection Problem. **Archives of Computational Methods in Engineering**, p. 1-25, 2020.

SHEHAB, M.; KHADER, A. T.; AL-BETAR, M. A. A survey on applications and variants of the cuckoo search algorithm. **Applied Soft Computing**, v. 61, p. 1041-1059, 2017.

SHU, K. et al. Fake news detection on social media: A data mining perspective. **ACM SIGKDD Explorations Newsletter**, v. 19, n. 1, p. 22-36, 2017.

SIVARAJAH, U. et al. Critical analysis of Big Data challenges and analytical methods. **Journal of Business Research**, v. 70, p. 263-286, 2017.

STATISTA. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. **Statista**, 2022. Disponível em: < https://bit.ly/37Fh9vl >. Acesso em: 14 mar. 2022.

STOREY, V. C.; SONG, Il-Y. Big data technologies and Management: What conceptual modeling can do. **Data & Knowledge Engineering**, v. 108, p. 50-67, 2017.

TRIPATHY, A.; AGRAWAL, A.; RATH, S. K. Classification of sentiment reviews using n-gram machine learning approach. **Expert Systems with Applications**, v. 57, p. 117-126, 2016.

UYSAL, A. K. An improved global feature selection scheme for text classification. **Expert systems with Applications**, v. 43, p. 82-92, 2016.

_____. On two-stage feature selection methods for text classification. **IEEE Access**, v. 6, p. 43233-43251, 2018.

VALÊNCIO, C. R. et al. Data Warehouse Design to Support Social Media Analysis in a Big Data Environment. **Journal of Computer Science,** v. 16, n. 2, p. 126-136, 2020.

VIOULÈS, M. J. et al. Detection of suicide-related posts in Twitter data streams. **IBM Journal of Research and Development**, v. 62, n. 1, p. 7: 1-7: 12, 2018.

YADAV, A.; VISHWAKARMA, D. K. A comparative study on bio-inspired algorithms for sentiment analysis. **Cluster Computing**, p. 1-21, 2020.

YANG, X.; DEB, S. Cuckoo search via Lévy flights. In: **2009 World congress on nature & biologically inspired computing (NaBIC)**. IEEE, 2009. p. 210-214.

YANG, X. Cuckoo search and firefly algorithm: overview and analysis. In: **Cuckoo search and firefly algorithm**. Springer, Cham, 2014. p. 1-26.

YANG, X. (Ed.). **Nature-inspired algorithms and applied optimization**. Springer, 2017.

YUE, L. et al. A survey of sentiment analysis in social media. **Knowledge and Information Systems**, p. 1-47, 2018.

ZAINUDDIN, N.; SELAMAT, A.; IBRAHIM, R.. Hybrid sentiment classification on twitter aspect-based sentiment analysis. **Applied Intelligence**, v. 48, n. 5, p. 1218-1232, 2018.