



ALICE REZENDE RIBEIRO

**UMA ABORDAGEM EM CASCATA PARA PREDIÇÃO DE
FAIXA ETÁRIA DE AUTORES A PARTIR DOS SEUS TEXTOS**

LAVRAS – MG

2025

ALICE REZENDE RIBEIRO

**UMA ABORDAGEM EM CASCATA PARA PREDIÇÃO DE FAIXA ETÁRIA DE
AUTORES A PARTIR DOS SEUS TEXTOS**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós Graduação em Ciência da Computação, área de concentração em Inteligência Artificial e Otimização, para a obtenção do título de Mestre.

Prof. Dr. Luiz Henrique De Campos Merschmann
Orientador

**LAVRAS – MG
2025**

**Ficha Catalográfica elaborada pelo Sistema de Geração
de Ficha Catalográfica da Biblioteca Universitária da UFLA, com
dados informados pelo(a) próprio(a) autor(a).**

Rezende Ribeiro, Alice.

Uma Abordagem em Cascata para Predição de Faixa Etária de Autores a Partir
dos seus Textos / Alice Rezende Ribeiro. - 2025.
72 p.

Orientador: Luiz Henrique De Campos Merschmann

Dissertação (Mestrado Acadêmico) - Universidade Federal de Lavras, 2025.
Bibliografia.

1. Mineração Textual. 2. Faixa Etária. 3. Caracterização autora. I. De Campos
Merschmann, Luiz Henrique. II. Universidade Federal de Lavras. III. Título.

ALICE REZENDE RIBEIRO

**UMA ABORDAGEM EM CASCATA PARA PREDIÇÃO DE FAIXA ETÁRIA DE
AUTORES A PARTIR DOS SEUS TEXTOS**

**A CASCADE APPROACH FOR PREDICTING AUTHORS' AGE RANGE FROM
THEIR TEXTS**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós Graduação em Ciência da Computação, área de concentração em Inteligência Artificial e Otimização, para a obtenção do título de Mestre.

APROVADA em 30 de maio de 2025.

Profa. Dra. Elaine Cecília Gato UFLA

Prof. Dr. Reginaldo Cordeiro dos Santos Filho UFPA

Prof. Dr. Luiz Henrique De Campos Merschmann
Orientador

**LAVRAS – MG
2025**

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelos projetos CNPq Universal 406411/2021-2 e FAPEMIG Universal APQ-02176-21.

RESUMO

Com a difusão do uso da internet, ferramentas de mineração de textos têm se tornado importantes para a estratégia de negócio das empresas. Ao mesmo tempo, o crescente volume de textos anônimos disponíveis na internet representa um desafio à obtenção de informações úteis, por exemplo, para modelos de personalização para comércio e marketing digital e em áreas como a forense digital. Na área de Caracterização Autoral — área que busca prever características sociais e demográficas de um grupo com base em seus textos —, embora haja um grande número de estudos, eles apresentam desigualdades quanto às características analisadas, à linguagem utilizada e às abordagens de classificação empregadas. Dessa forma, o presente trabalho busca propor e avaliar um método de classificação em cascata, que combina formas de classificação tradicional com um modelo baseado em dicionários, para prever faixas etárias de autores com base em textos escritos na língua portuguesa. Os resultados encontrados mostram que o modelo baseado em dicionários em combinação com modelos tradicionais podem trazer melhoria na predição da faixa etária.

Palavras-chave: caracterização autoral; faixa etária; mineração de textos.

ABSTRACT

With the widespread use of the Internet, text mining tools have become important for companies' business strategies. At the same time, the growing volume of anonymous texts available on the Internet poses a challenge to obtaining useful information, for example, in the development of personalization models for digital commerce and marketing, as well as in fields such as digital forensics. In the area of Authorial Characterization — an area that seeks to predict social and demographic characteristics of a group based on their texts —, although there are a large number of studies, they present inequalities regarding the characteristics analyzed, the language used and the classification approaches employed. Thus, this work seeks to propose and evaluate a cascade classification method, which combines traditional classification methods with a dictionaries-based model, in order to predict author age ranges based on texts written in Portuguese. The results found show that the dictionaries-based model in combination with traditional models can bring improvements in age range classification.

Keywords: author profiling; age range; text mining.

INDICADORES DE IMPACTO

Este estudo propõe um procedimento de classificação em cascata visando à predição de intervalos etários a partir de textos em português, destacando impactos tanto tecnológicos quanto sociais.

No âmbito tecnológico, o estudo promove o avanço de sistemas inteligentes dedicados à caracterização autoral - uma subárea da mineração de textos com aplicações pertinentes em comércio eletrônico, marketing digital e forense digital. A integração de modelos tradicionais de aprendizado com técnicas baseadas em dicionários e as melhorias proporcionadas por essa estratégia podem potencializar sistemas de recomendação personalizados, principalmente em contextos que envolvam textos anônimos.

Do ponto de vista social, apesar de sua natureza acadêmica, a pesquisa sugere a possibilidade de criação de soluções que respeitem a privacidade dos usuários, enquanto possibilitam inferências valiosas para segmentação de campanhas informativas e estratégias em ambientes digitais.

Entre aqueles que podem se beneficiar estão profissionais de TI, pesquisadores em linguística e ciência de dados, além de usuários de plataformas digitais que utilizam modelos de personalização.

A pesquisa está em consonância com o Objetivo de Desenvolvimento Sustentável (ODS) nº 9 da ONU - Indústria, Inovação e Infraestrutura - ao sugerir inovações para a mineração textual.

IMPACT INDICATORS

This study proposes a cascade classification procedure for predicting age ranges from texts in Portuguese, highlighting both technological and social impacts.

In the technological field, the study promotes the advancement of intelligent systems dedicated to authorial characterization - a sub-area of text mining with relevant applications in e-commerce, digital marketing and digital forensics. The integration of traditional learning models with dictionary-based techniques and the improvements provided by this strategy can enhance personalized recommendation systems, especially in contexts involving anonymous texts.

From a social perspective, despite its academic nature, the research suggests the possibility of creating solutions that respect user privacy, while enabling valuable inferences for segmenting information campaigns and strategies in digital environments.

Among those who can benefit are IT professionals, researchers in linguistics and data science, as well as users of digital platforms that use personalization models.

The research is in line with UN Sustainable Development Goal (SDG) No. 9 - Industry, Innovation and Infrastructure - on innovation for text mining.

LISTA DE FIGURAS

Figura 2.1 – Exemplo de SVM	20
Figura 2.2 – Hiperplanos	21
Figura 2.3 – Arquitetura de uma rede neural	22
Figura 2.4 – <i>Holdout</i>	24
Figura 2.5 – Validação cruzada com 5 partições	25
Figura 4.1 – Abordagem cascata proposta	36
Figura 4.2 – Exemplo de dicionários por faixa etária - Palavras mais influentes (e-SIC) .	39
Figura 4.3 – Exemplo de mapeamento das instâncias e do hiperplano em um problema de classificação multiclasse	40
Figura 4.4 – Exemplo de mapeamento das instâncias e dos múltiplos hiperplanos em um problema de classificação multiclasse	42
Figura 4.5 – Exemplo de instâncias novas em relação aos múltiplos hiperplanos em um problema de classificação multiclasse	43

LISTA DE TABELAS

Tabela 2.1 – Exemplo BOW TF-IDF	29
Tabela 5.1 – Distribuição das instâncias na base B5 Corpus	46
Tabela 5.2 – Distribuição das instâncias na base BR Moral	46
Tabela 5.3 – Distribuição das instâncias na base BlogSet BR	46
Tabela 5.4 – Distribuição das instâncias na base e-SIC	47
Tabela 5.5 – Tamanho do vocabulário para cada base de dados	48
Tabela 5.6 – Hiperparâmetros do SVM	49
Tabela 5.7 – Redução do vocabulário para cada base	50
Tabela 5.8 – Resultados da reimplementação dos trabalhos de referência	52
Tabela 5.9 – Resultados do modelo baseado em dicionários em comparação com a abordagem de classificação tradicional	53
Tabela 5.10 – Resultados da Abordagem em Cascata e do <i>Baseline</i>	54
Tabela 5.11 – Resultados do Teste Estatístico entre a Abordagem em Cascata e do <i>Baseline</i>	54
Tabela 1 – Hiperparâmetros utilizados no classificador CNN	60
Tabela 2 – Hiperparâmetros utilizados no classificador LSTM	61
Tabela 3 – Hiperparâmetros utilizados no classificador Regressão Logística - BR Moral	61
Tabela 4 – Resultado das Execuções Individuais do Modelo <i>Baseline</i>	62
Tabela 5 – Resultado das Execuções Individuais do Modelo <i>Baseline</i>	62
Tabela 6 – Resultado das Execuções Individuais do Modelo Baseado em Dicionários	62
Tabela 7 – Resultado das Execuções Individuais do Modelo Baseado em Dicionários	63
Tabela 8 – Resultado das Execuções Individuais do Modelo Tradicional (Instâncias classificadas pelo dicionário)	63
Tabela 9 – Resultado das Execuções Individuais do Modelo Tradicional (Instâncias classificadas pelo dicionário)	63
Tabela 10 – Resultado das Execuções Individuais da Cascata	63
Tabela 11 – Resultado das Execuções Individuais da Cascata	64
Tabela 12 – Resultados do modelo baseado em dicionários em comparação com a abordagem de classificação tradicional - Por classe	71
Tabela 13 – Resultados da Abordagem em Cascata e do <i>Baseline</i> - Por classe	71

Tabela 14 – BR Moral - Resultados do modelo baseado em dicionários em comparação com a abordagem de classificação tradicional - Por classe e por partição . . 72

Tabela 15 – BR Moral - Resultados da Abordagem em Cascata e do *Baseline* - Por classe e por partição 73

LISTA DE QUADROS

Quadro 2.1 – Matriz de Confusão para a classe X	26
Quadro 3.1 – Resumo dos trabalhos relacionados - Predição da faixa etária como um problema de classificação em textos da língua portuguesa	35
Quadro 5.1 – Características da bases de dados	47
Quadro 5.2 – α_{ageX} adotado por base de dados	50
Quadro 5.3 – Configuração experimental dos trabalhos selecionados	51
Quadro .1 – BlogSet BR - Matriz de confusão para a faixa etária 1 (10-25), faixa etária 2 (26-40) e faixa etária 3 (>40) para a abordagem tradicional (somente instâncias classificadas pelo modelo baseado em dicionário).	65
Quadro .2 – BlogSet BR - Matriz de confusão para a faixa etária 1 (10-25), faixa etária 2 (26-40) e faixa etária 3 (>40) para modelo baseado em dicionários.	65
Quadro .3 – BlogSet BR - Matriz de confusão para a faixa etária 1 (10-25), faixa etária 2 (26-40) e faixa etária 3 (>40) para o <i>baseline</i>	65
Quadro .4 – BlogSet BR - Matriz de confusão para a faixa etária 1 (10-25), faixa etária 2 (26-40) e faixa etária 3 (>40) para a abordagem cascata.	66
Quadro .5 – BR Moral - Matriz de confusão para a faixa etária 1 (0-23), faixa etária 2 (24-30) e faixa etária 3 (31-99) para a abordagem tradicional (somente instâncias classificadas pelo modelo baseado em dicionário). Somatório das 10 partições.	66
Quadro .6 – BR Moral - Matriz de confusão para a faixa etária 1 (0-23), faixa etária 2 (24-30) e faixa etária 3 (31-99) para modelo baseado em dicionários. Somatório das 10 partições.	66
Quadro .7 – BR Moral - Matriz de confusão para a faixa etária 1 (0-23), faixa etária 2 (24-30) e faixa etária 3 (31-99) para o <i>baseline</i> . Somatório das 10 partições.	67
Quadro .8 – BR Moral - Matriz de confusão para a faixa etária 1 (0-23), faixa etária 2 (24-30) e faixa etária 3 (31-99) para a abordagem cascata. Somatório das 10 partições.	67
Quadro .9 – B5 Corpus - Matriz de confusão para a faixa etária 1 (18-20), faixa etária 2 (23-25) e faixa etária 3 (28-61) para a abordagem tradicional (somente instâncias classificadas pelo modelo baseado em dicionário).	67

Quadro .10 – B5 Corpus - Matriz de confusão para a faixa etária 1 (18-20), faixa etária 2 (23-25) e faixa etária 3 (28-61) para modelo baseado em dicionários.	68
Quadro .11 – B5 Corpus - Matriz de confusão para a faixa etária 1 (18-20), faixa etária 2 (23-25) e faixa etária 3 (28-61) para o <i>baseline</i>	68
Quadro .12 – B5 Corpus - Matriz de confusão para a faixa etária 1 (18-20), faixa etária 2 (23-25) e faixa etária 3 (28-61) para a abordagem cascata.	68
Quadro .13 – e-SIC - Matriz de confusão para a faixa etária 1 (17-30), faixa etária 2 (31-42) e faixa etária 3 (>42) para a abordagem tradicional (somente instâncias classificadas pelo modelo baseado em dicionário).	69
Quadro .14 – e-SIC - Matriz de confusão para a faixa etária 1 (17-30), faixa etária 2 (31-42) e faixa etária 3 (>42) para modelo baseado em dicionários.	69
Quadro .15 – e-SIC - Matriz de confusão para a faixa etária 1 (17-30), faixa etária 2 (31-42) e faixa etária 3 (>42) para o <i>baseline</i>	69
Quadro .16 – e-SIC - Matriz de confusão para a faixa etária 1 (17-30), faixa etária 2 (31-42) e faixa etária 3 (>42) para a abordagem cascata.	70

LISTA DE SÍMBOLOS

F	Função que define um modelo de aprendizado de máquina em um problema de classificação
X	Classe genérica pertencente às classes do problema
i	Instância genérica pertencente aos dados
A, B, C	Exemplo de classes em um problema de classificação
n	Número de classes em um problema de classificação
k	Número de partições em uma validação cruzada
P	Métrica Precisão
R	Métrica Revogação
n	Conjunto de classes em um problema de classificação
t	Termo (palavra ou <i>token</i>) presente dentro de um conjunto textual
T	Número de instâncias que o termo t aparece dentro de um conjunto textual
ni	Número de instâncias em um conjunto de dados
age_{lex_X}	Índice calculado com base no dicionário da classe X que é utilizado para posicionar as instâncias em relação ao hiperplano da classe X e assim classificar a instância
α_{age_X}	Limiar adotado para o age_{lex_X} para o aumento da confiabilidade do modelo de classificação com base em dicionários

SUMÁRIO

1	INTRODUÇÃO	15
1.1	Objetivos	16
1.2	Estrutura do Documento	17
2	FUNDAMENTAÇÃO TEÓRICA	18
2.1	Classificação	18
2.1.1	Métodos de Classificação	18
2.1.1.1	Máquinas de Vetores de Suporte (SVM)	19
2.1.1.2	Regressão Logística	21
2.1.1.3	Redes Neurais	22
2.1.1.4	Avaliação	23
2.1.1.4.1	Particionamento do Conjunto de Dados	23
2.1.1.4.2	Métricas de Avaliação	24
2.2	Classificação Textual	27
2.2.1	Pré-Processamento	27
2.2.2	Representação Vetorial	28
2.3	Caracterização Autoral	30
3	TRABALHOS RELACIONADOS	32
4	METODOLOGIA	36
4.1	Modelo de Classificação Baseado em Dicionários	37
4.1.1	Criação dos Dicionários	38
4.1.2	Cálculo dos Limiares α_{ageX}	41
4.2	Modelo de Classificação Tradicional	43
5	EXPERIMENTOS COMPUTACIONAIS	45
5.1	Base de Dados	45
5.2	Configuração Experimental	47
5.2.1	Modelo Baseado em Dicionários	48
5.2.2	Modelos de Classificação Tradicionais	50
5.3	Resultados e Análises	51
6	CONCLUSÃO	56
	REFERÊNCIAS	57
	APÊNDICES	60

1 INTRODUÇÃO

A Internet continua a modificar a maneira com a qual as pessoas interagem, disponibilizam dados e compartilham seus pontos de vista. Em 2024, o número de usuários da Internet no mundo alcançou a marca de 5,44 bilhões, cerca de dois terços da população mundial (Statista, 2025). Assim, o volume de dados gerados, coletados e consumidos segue aumentando a cada ano. Estima-se que o uso da Internet tenha criado cerca de 149 zettabytes de dados até o fim de 2024 (DOMO, 2024). Esses dados têm se tornado essenciais para uma variedade de negócios, sendo que, em 2024, o valor do mercado mundial de soluções voltadas para a análise de dados e ferramentas de inteligência de negócios foi avaliado em US\$31,98 bilhões (Insights, 2025).

Apesar de diferentes tipos de dados veicularem na Internet atualmente, estima-se que até 95% deles apresentem um formato não estruturado (Adnan; Akbar, 2019), sendo grande parte disso formado por textos. Nesse cenário, a área de Mineração de Texto torna-se essencial para agregar valor a esses dados, permitindo o processamento dos mesmos para a extração de informações valiosas para diversas aplicações.

Além dos textos publicados na Internet, conhecer as características dos seus autores é de grande valor para diversas empresas e organizações, uma vez que essa informação pode ser utilizada para a criação de modelos de recomendação e personalização de produtos e serviços (Mastercard, 2020). Conhecer as características dos usuários também já se mostrou útil para melhorar o desempenho de tarefas como análise de sentimentos (Guimarães *et al.*, 2017). Outra área que se beneficia do conhecimento das características do usuário é a computação forense, que utiliza dados sobre autores de publicações em investigações de crimes cibernéticos (Yu, 2023).

No entanto, nem sempre as características dos autores de publicações na Internet são conhecidas. Por isso, para esse cenário, técnicas computacionais vêm sendo propostas por pesquisadores que atuam em uma área de pesquisa denominada Caracterização Autoral, cujo objetivo é inferir características dos autores a partir de seus textos. Nessa área, a partir de conceitos da linguística, da estatística e da computação, busca-se inferir características como gênero, faixa etária, religiosidade, entre outras (Dias, 2019).

Uma variedade de estudos na área de Caracterização Autoral têm suas raízes na competição PAN-CLEF (*Plagiarism, Authorship, and Near-Duplicate Detection - at the CLEF conference*), uma competição internacional que propõe desafios que estimulam a pesquisa em dife-

rentes áreas de análise de texto. Esse evento é conhecido por resultar em trabalhos que reúnem as técnicas estado-da-arte para os problemas da área e por gerar diversos conjuntos de dados textuais que se tornam referências. Tarefas como a predição de faixa etária, de gênero, a identificação de dialetos em um mesmo idioma, bem como a determinação de se o texto foi escrito por um humano ou um *bot* estão entre as diversas que já foram exploradas nessa competição (Bevendorff *et al.*, 2022).

Embora estudos tenham sido conduzidos nessa área de Caracterização Autoral, os progressos não são alcançados de forma homogênea com relação à característica predita, ao idioma dos textos analisados e ao desempenho das abordagens computacionais utilizadas. Por exemplo, existe uma grande concentração de trabalhos focados no idioma inglês (Hsieh; Dias; Paraboni, 2018), enquanto uma quantidade muito menor de pesquisas, recursos e ferramentas computacionais foi proposta para a língua portuguesa. Além disso, enquanto o desempenho preditivo de abordagens propostas para a tarefa de predição de gênero já alcançou índices de até 94% para a métrica F1 (Dias, 2019; Silva, 2020; Hsieh; Dias; Paraboni, 2018; Flores; Pavan; Paraboni, 2022; Moraes; Merschmann, 2022), estudos voltados para a predição da faixa etária (considerando-a como um problema de classificação multiclasse) alcançaram resultados muito inferiores, variando entre 40% e 67% de F1 (Guimarães *et al.*, 2017; Hsieh; Dias; Paraboni, 2018; Dias, 2019; Flores; Pavan; Paraboni, 2022; Neto, 2021).

Portanto, o foco deste trabalho é a predição da faixa etária de autores de textos escritos na língua portuguesa. No entanto, diferentemente da abordagem tradicional de classificação de texto utilizada na maioria dos trabalhos recentes apresentados na literatura, o presente estudo propõe avaliar uma abordagem alternativa que explora o domínio textual a partir da utilização de dicionários para realizar a tarefa de predição de faixa etária.

A hipótese na qual se baseia este trabalho é que explorar características específicas do domínio textual pode contribuir para a melhoria do desempenho preditivo da tarefa de classificação da faixa etária de autores de textos escritos na língua portuguesa.

1.1 Objetivos

Dada a hipótese apresentada anteriormente, este trabalho tem como objetivo propor e avaliar uma abordagem de classificação que leve em consideração o domínio do texto para

predizer a faixa etária do autor de um texto escrito na língua portuguesa utilizando somente o conteúdo textual. Esse objetivo geral pode ser desdobrado nos seguintes objetivos específicos:

- a) contribuir na área de Caracterização Autoral melhorando o desempenho obtido pelas abordagens apresentadas na literatura para a tarefa de predição de faixa etária de autores de textos escritos na língua portuguesa;
- b) explorar as características dos diferentes domínios textuais para auxiliar na tarefa de predição de faixas etárias;
- c) avaliar o uso de diferentes estratégias para criar e utilizar dicionários para capturar o domínio textual visando obter uma abordagem eficiente para a predição de faixas etárias.

1.2 Estrutura do Documento

Este capítulo introduziu o problema investigado nesta pesquisa, além de apresentar a motivação, a hipótese e os objetivos da mesma. No Capítulo 2, discorre-se sobre os fundamentos essenciais à compreensão deste estudo. O Capítulo 3 revisita estudos prévios da literatura que também exploraram a classificação da faixa etária de autores de textos na língua portuguesa. A metodologia proposta é detalhada no Capítulo 4. No Capítulo 5, são apresentados os experimentos computacionais realizados para avaliar a abordagem aqui proposta. Por fim, o Capítulo 6 resume as conclusões da pesquisa e aponta direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são discutidos conceitos teóricos importantes para o entendimento do trabalho. Primeiramente, na seção 2.1, são discutidos os classificadores, suas técnicas, formas de avaliação e diferentes abordagens. Depois, a Seção 2.2 fala sobre a classificação textual e suas técnicas. Por fim, a Seção 2.3 apresenta um resumo sobre a área de Caracterização Autoral.

2.1 Classificação

O problema da classificação no aprendizado de máquina consiste em categorizar uma instância em uma classe pertencente a um conjunto de classes previamente conhecidas a partir dos atributos que caracterizam essa instância (Han; Kamber; Pei, 2011).

O início do processo de classificação envolve a fase de treinamento do modelo, onde instâncias previamente rotuladas são empregadas para definir uma função F , $X = F(i)$, de forma a prever a classe X correspondente a uma instância i . Essa função, frequentemente chamada de modelo, é o resultado obtido via diferentes técnicas e algoritmos de aprendizado de máquina. Em seguida, na segunda etapa da classificação, conhecida como fase de teste, instâncias que não participaram do treinamento são submetidas ao modelo, permitindo assim avaliar sua eficácia e capacidade de generalização. O objetivo de um classificador é gerar um modelo que categorize um conjunto de dados dentro de classes predefinidas, minimizando erros de classificação em relação à classe verdadeira, caracterizando-se como uma forma de aprendizado supervisionado (Han; Kamber; Pei, 2011).

A seguir serão discutidos os algoritmos de classificação plana (ou tradicional) e a forma como esses algoritmos são avaliados.

2.1.1 Métodos de Classificação

Conforme mencionado anteriormente, o processo de classificação envolve a criação de um modelo para categorizar instâncias em classes pré-definidas. Diversas técnicas, cada uma com sua própria abordagem, podem ser empregadas na criação desses modelos (Han; Kamber; Pei, 2011). A seguir, são apresentadas algumas técnicas de classificação frequentemente empregadas na classificação plana.

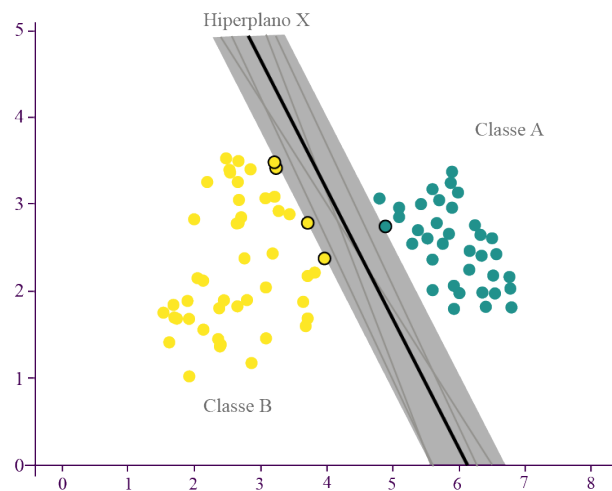
2.1.1.1 Máquinas de Vetores de Suporte (SVM)

Máquina de Vetores de Suporte (SVM) é uma técnica de aprendizado de máquina que busca representar as instâncias e seus atributos como pontos (ou vetores) em um espaço multi-dimensional. Após esse mapeamento, o método busca encontrar uma superfície de separação, ou hiperplano, que maximize a distância entre duas classes diferentes, buscando a máxima separação possível com base nos dados de treinamento. Assim, uma nova instância é classificada a partir da sua posição no espaço multidimensional em relação ao hiperplano de separação (Baeza-Yates; Ribeiro-Neto, 2013).

Em um problema de classificação binária, o treinamento de um classificador SVM resulta em um único hiperplano de separação entre as duas classes do problema. O hiperplano escolhido é aquele que maximiza a margem de separação entre as instâncias mais próximas do hiperplano, que são denominadas vetores de suporte. Definido o hiperplano de separação, a predição da classe de uma nova instância se dá verificando-se a posição da mesma em relação a esse. A definição dessa posição é obtida a partir do valor e do seu sinal (positivo ou negativo) obtidos ao aplicar a equação do hiperplano para a instância que se deseja classificar. Na Figura 2.1 apresenta um exemplo de um SVM Linear, utilizando dois atributos, no espaço 2-dimensional. Os vetores de suporte são representados pelas instâncias de contorno preto e o espaço cinza representa o espaço em que deve ser procurada a fronteira de separação, que deve ser escolhida de forma a maximizar a distância de instâncias de diferentes classes (A, B, C). As opções de hiperplano são representadas por linhas cinzas: essas não foram escolhidas pois não maximizam as distâncias dos vetores de suporte de ambas as classes. O hiperplano ou superfície de decisão escolhido é representado pelo Hiperplano X . Assim, o hiperplano X separa as instâncias da classe A e B . Caso uma nova instância esteja à esquerda do hiperplano, ela será considerada como pertencente à classe B e, caso ela esteja à direita, classe A .

Porém, o SVM somente consegue separar o espaço em duas partes, ou seja, só faz decisões binárias, e, por isso, para problemas de classificação multiclasse algumas estratégias podem ser adotadas. Na primeira *One vs. Rest* ou "Um-Contra-Todos", utilizado nesse trabalho, tem-se um classificador para cada classe, que define se a instância i pertence à classe X . O treinamento é realizado considerando todas as instâncias da classe X como positivas e todas as instâncias de outras classes, como negativas. Quando uma nova instância é classificada, são calculados os valores para os n classificadores, sendo n o número de classes do problema.

Figura 2.1 – Exemplo de SVM

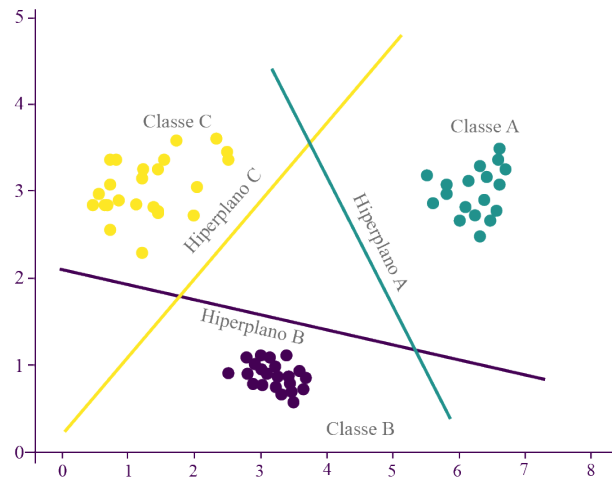


Fonte: Adaptado de Baeza-Yates & Ribeiro-Neto (2013)

A instância pode ser classificada em nenhuma, em uma ou em mais de uma classe, sendo a escolhida a que apresenta a maior distância em relação ao hiperplano (Baeza-Yates; Ribeiro-Neto, 2013). Veja o exemplo da Figura 2.2 para um problema de classificação com três classes (A, B, C). Nesse exemplo, o Hiperplano A separa as instâncias da classe A das instâncias das classes B e C; o Hiperplano B separa as instâncias da classe B das instâncias das classes A e C; e o Hiperplano C separa as instâncias da classe C das instâncias das classes A e B. Assim, a classificação de uma instância é feita olhando a sua posição em relação a todos os hiperplanos (A, B, C).

Outra abordagem é a *One vs. One* ou "Um-contra-Um", na qual são construídos classificadores entre duas classes do problema. Assim, todas as classes do problema são combinadas em pares e é construído um classificador que deve dizer se a instância i é pertencente à classe A ou à classe B. O treinamento do classificador das classes A e B é feito utilizando somente o exemplo dessas classes, ignorando as outras instâncias. Para classificar uma nova instância, todos os classificadores são executados e o resultado é comparado via votação simples ou outras estratégias mais complexas (Baeza-Yates; Ribeiro-Neto, 2013).

Figura 2.2 – Hiperplanos



Fonte: Criado pela autora

2.1.1.2 Regressão Logística

A regressão logística é uma técnica de classificação que se baseia na probabilidade de uma instância pertencer a uma classe, resultando sempre em uma saída entre 0 e 1. A variável a ser predita é denominada variável dependente, enquanto as variáveis de entrada são conhecidas como variáveis independentes ou preditoras. Os valores de probabilidade são determinados por uma função sigmóide aplicada à combinação linear das entradas (Hilbe, 2015). Inicialmente, calcula-se $g(x)$ conforme a Equação 2.1, em que o conjunto de pesos é estimado com base no conjunto de dados de treinamento, visando maximizar a probabilidade do evento ocorrer (Hilbe, 2015). Com o modelo estabelecido, para cada instância i a ser predita, a regressão logística retorna a probabilidade de i pertencer a uma classe, onde a probabilidade p_i é calculada pela Equação 2.2 (Hilbe, 2015) para todas classes possíveis e a instância é associada com a classe que apresenta o maior valor.

$$g(x) = B_0 + B_1X_1 + \dots + B_pX_p \quad (2.1)$$

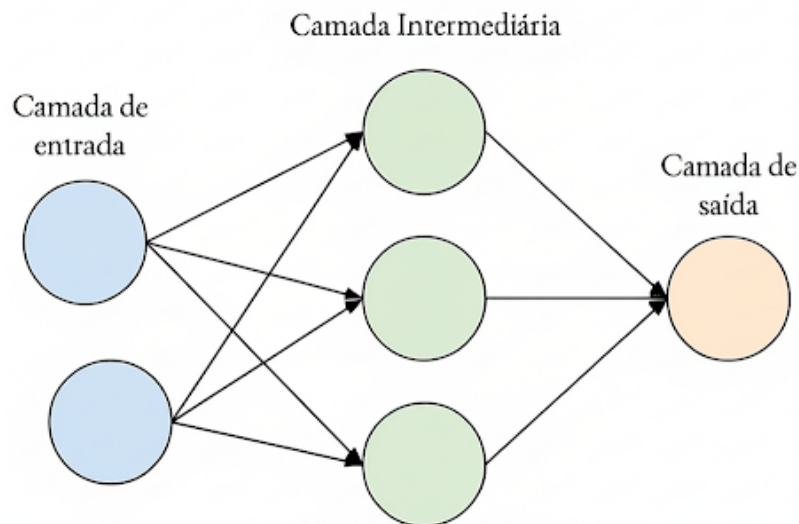
$$p_i = \frac{1}{1 + e^{-g(x)}} \quad (2.2)$$

2.1.1.3 Redes Neurais

As redes neurais constituem um modelo computacional inspirado no sistema neural humano (Mikheev; Gusynina; Shornikova, 2020). Compostas por nós interconectados através de conexões ponderadas, essas conexões refletem a influência de cada neurônio no resultado final. Ao receber uma entrada, cada nó realiza cálculos com base nos pesos das conexões, transmitindo o resultado para uma camada intermediária ou para a camada de saída, conforme discutido por Dongare, Kharde & Kachare (2012). Na camada de saída, o valor é processado por uma função de ativação, gerando uma saída mais interpretável. Através da retropropagação e da taxa de aprendizado durante cada iteração de treinamento, os pesos são ajustados para minimizar o erro com o valor real, permitindo que a rede aprenda e aprimore seus resultados (Mikheev; Gusynina; Shornikova, 2020).

A Figura 2.3 demonstra um exemplo da arquitetura de redes neurais. No caso de redes neurais profundas, essas possuem mais de uma camada intermediária (LeCun; Bengio; Hinton, 2015).

Figura 2.3 – Arquitetura de uma rede neural



Fonte: Criado pela autora

As Redes Neurais Convolucionais (CNNs) são um tipo de rede neural profunda projetada para processar dados estruturados em grades, como imagens e sequências temporais. Elas possuem a capacidade de capturar informações contextuais em diferentes escalas, ao mesmo

tempo em que reduzem a complexidade computacional por meio da aplicação de filtros compartilhados e operações de pooling. Cada conjunto de neurônios dentro da rede é responsável por detectar padrões específicos, como bordas, texturas e formas mais complexas à medida que as camadas avançam. Essa característica torna as CNNs altamente eficazes na extração automática de características relevantes dos dados, sendo amplamente utilizadas em tarefas como reconhecimento facial, reconhecimento de fala, processamento de linguagem natural e visão computacional. Além disso, sua estrutura hierárquica confere uma notável robustez a variações nos dados de entrada, garantindo melhor generalização em diferentes cenários (Zhang, 2018).

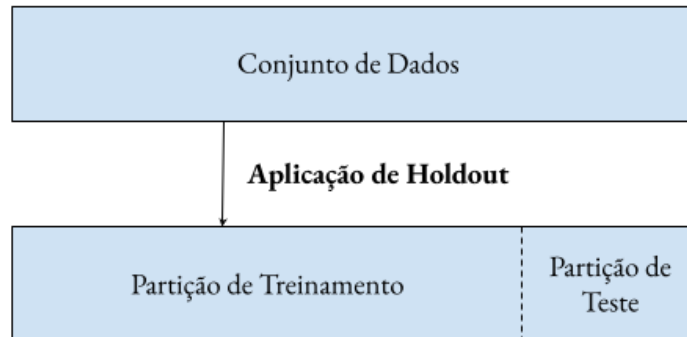
Em Hochreiter & Schmidhuber (1997), foi proposta a arquitetura da rede neural *Long Short-Term Memory* (LSTM), um modelo desenvolvido para superar a dificuldade das redes neurais em aprender informações de longo prazo. Esse problema ocorre devido ao desaparecimento ou explosão dos gradientes ao longo do treinamento. A LSTM resolve essa limitação por meio de um mecanismo de controle do fluxo de informações baseado em portas (*gates*). Cada célula LSTM possui uma porta de entrada, uma porta de esquecimento e uma porta de saída, que regulam a passagem e retenção de informações, além de um estado de célula que preserva dados relevantes ao longo do tempo. Esse design reduz o impacto do desaparecimento ou explosão do gradiente e torna a LSTM altamente eficiente para tarefas sequenciais, como reconhecimento de fala, previsão de séries temporais e aplicações em processamento de linguagem natural (NLP).

2.1.1.4 Avaliação

Após a criação dos modelos de classificação, é necessário calcular quão precisos eles são. Esta seção começa apresentando como os dados podem ser divididos em conjuntos de treinamento e teste e, em seguida, as métricas utilizadas para avaliar o desempenho dos modelos são apresentadas.

2.1.1.4.1 Particionamento do Conjunto de Dados

O método de *holdout* consiste em dividir aleatoriamente o conjunto de dados em duas partes independentes, denominadas de treinamento e teste, como ilustrado na Figura 2.4. O conjunto de treinamento serve para ajustar o modelo, enquanto o conjunto de teste é utilizado para calcular as métricas de desempenho.

Figura 2.4 – *Holdout*

Fonte: Criado pela autora

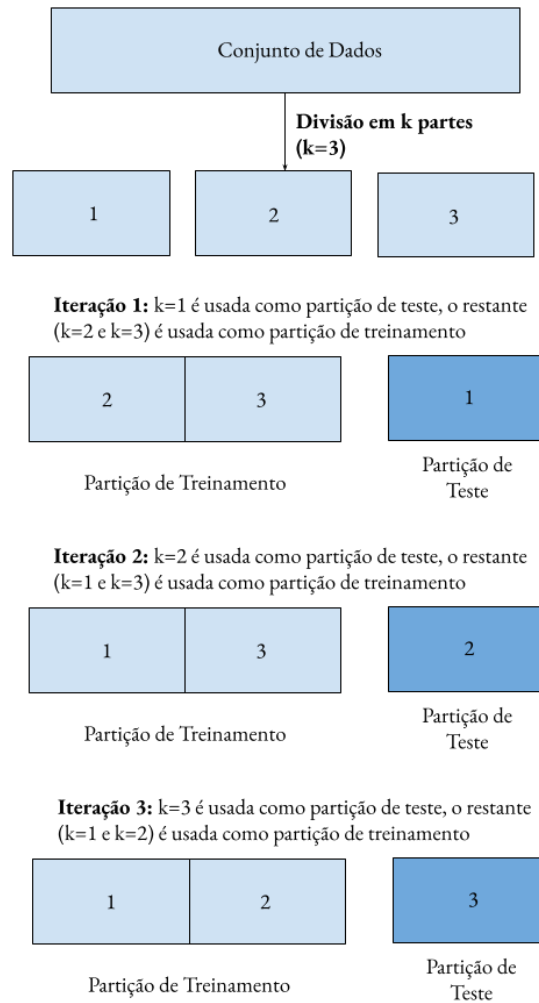
Na de validação cruzada, uma base de dados é dividida aleatoriamente em k partes (partições), e o procedimento de avaliação do modelo é executado k vezes. Em cada execução, uma das partições funciona como conjunto de teste, enquanto as restantes $k - 1$ partições são usadas para o treinamento do modelo (Han; Kamber; Pei, 2011), como ilustra a Figura 2.5. Quando se realiza validação cruzada de forma estratificada, as proporções de cada classe presente na base original são preservadas em todas as partições.

2.1.1.4.2 Métricas de Avaliação

Quando o resultado de um classificador é analisado para uma determinada classe X , podem ocorrer os seguintes casos:

- a) **verdadeiro positivo ou VP_X** : quando uma instância de uma classe X foi corretamente classificada na classe X ;
- b) **verdadeiro negativo ou VN_X** : quando uma instância que não é da classe X é corretamente classificada como não sendo da classe X ;
- c) **falso positivo ou FP_X** : quando uma instância que não é da classe X é identificada como classe X ;
- d) **falso negativo ou FN_X** : quando uma instância da classe X é identificada como não sendo da classe X .

Figura 2.5 – Validação cruzada com 5 partições



Fonte: Criado pela autora

A ocorrência e frequência desses casos podem ser úteis para a avaliação de um classificador e são resumidas na matriz de confusão para a classe X exemplificada na Tabela 2.1.

A acurácia mede a porcentagem do conjunto de dados que foi classificada corretamente pelo classificador considerando todas as classes (Equação 2.3). Embora amplamente utilizada, ela não é uma boa métrica para casos em que há desbalanceamento na quantidade de instâncias por classe (Han; Kamber; Pei, 2011).

$$Acurcia = \frac{VP + VN}{VP + VN + FP + FN} \quad (2.3)$$

Quadro 2.1 – Matriz de Confusão para a classe X

		Classe Predita	
		Classe X	Outras classes
Classe real	Classe X	VP_X	FN_X
	Outras classes	FP_X	VN_X

Fonte: Criado pela autora

A precisão mede a fração das instâncias de uma classe que foi classificada corretamente em relação ao total de instâncias classificadas nessa classe (Equação 2.4) , enquanto a revocação mede a fração dos dados que foram classificados corretamente em uma classe em relação à quantidade real de instâncias pertencentes a essa classe (Equação em 2.5) (Baeza-Yates; Ribeiro-Neto, 2011).

$$Precisao_X = \frac{VP_X}{VP_X + FP_X} \quad (2.4)$$

$$Revocao_X = \frac{VP_X}{VP_X + FN_X} \quad (2.5)$$

A métrica F1 combina os valores de precisão e revocação, buscando equilibrar suas importâncias (Silva, 2020). Quando calculada por classe, segue a Equação 2.6. Frequentemente, para melhor interpretação e comparação, esses valores são consolidados em uma única métrica, resultante da média dos valores de F1 por classe. A métrica macro F1 é adotada quando se deseja conferir igual importância a cada classe, conforme a Equação 2.7. Por outro lado, a micro F1 é empregada quando se busca atribuir igual importância a cada instância, seguindo a Equação 2.8. Neste contexto, P e R representam a média de precisão e revocação entre as classes (Baeza-Yates; Ribeiro-Neto, 2011).

$$F_{1X} = \frac{2 * P_X * R_X}{P_X + R_X} \quad (2.6)$$

$$macroF_1 = \frac{\sum_{c=1}^X F_{1X}}{|c|} \quad (2.7)$$

$$microF_1 = \frac{2 * P * R}{P + R} \quad (2.8)$$

2.2 Classificação Textual

A atividade de classificação textual envolve a atribuição de rótulos previamente definidos a textos, utilizando técnicas de aprendizado de máquina supervisionadas. Em um contexto onde se tem um conjunto de documentos e um conjunto de classes, um classificador textual é representado por uma função F , que é derivada dos dados de treinamento. Essa função irá determinar se um documento i , pertencente ao conjunto de documentos do problema, pertence à classe X , pertencente às classes n do problema.

2.2.1 Pré-Processamento

Para possibilitar o processamento computacional de um texto, é essencial realizar uma fase preliminar de processamento, que tem como objetivo extrair informações importantes do conteúdo textual, além de remover ruídos e informações não relevantes para a tarefa em questão (Eisenstein, 2019). As tarefas de pré-processamento variam de acordo com as necessidades do modelo textual sendo criado. Dentre essas tarefas, tem-se a tokenização, transformação das letras de um texto para minúsculas, a remoção de caracteres especiais, remoção de caracteres específicos do domínio (tags HTML, hashtags, URLs), eliminação de pontuações, entre outras (Eisenstein, 2019).

A tokenização, geralmente utilizada nas primeiras etapas de pré-processamento textual, subdivide o texto em partes menores, como palavras, caracteres ou sequências de alguma das duas (Eisenstein, 2019). Por exemplo, a frase "Minha terra tem palmeiras onde canta o sabiá", ao passar pelo processo de tokenização, poderia ser dividida das seguintes formas:

- a) **tokenização utilizando uma palavras:** Minha, terra, tem, palmeiras, onde, canta, o, sabiá;
- b) **tokenização utilizando duas palavras:** Minha terra, tem palmeiras, onde canta, o sabiá;

- c) **tokenização utilizando letras:** M, i, n, h, a, t, e, r, r, a, t, e, m, p, a, l, m, e, i, r, a, s, o, n, d, e, c, a, n, t, a, o, s, a, b, i, á.

Outro pré-processamento frequente é a remoção de palavras muito e/ou pouco frequentes no texto, as quais geralmente agregam pouca informação para o processo de classificação. As *stopwords* são exemplos de palavras que aparecem com frequência nas frases e geralmente não contribuem significativamente para o seu significado. A definição específica de quais palavras são consideradas *stopwords* pode depender da tarefa em questão, mas classes gramaticais como artigos, pronomes e conjunções são exemplos de stopwords. A remoção dessas palavras é comumente realizada para reduzir o tamanho do texto, manter a informação concisa e sem ruídos, além de diminuir o tempo de processamento (Eisenstein, 2019). Por exemplo, para a frase "No meio do caminho tinha uma pedra", removendo as *stopwords* tem-se "No meio caminho pedra" (resultado obtido utilizando o pacote de *stopwords* em português da biblioteca *nltk*¹).

2.2.2 Representação Vetorial

A linguagem humana, apesar de parecer simples para os seres humanos, possui regras, ambiguidades e relações que podem ser incompreensíveis para as máquinas. Assim, o texto em linguagem natural deve ser convertido para uma representação vetorial mais adequada para o processamento por um computador. Essas representações buscam extrair o máximo de significado, estrutura e relações presentes no texto, tornando o processamento computacionalmente mais eficiente (Eisenstein, 2019).

Uma técnica comumente utilizada é a *Bag of Words* (BOW), que emprega representação vetorial em conjunto com métricas estatísticas. As métricas utilizadas podem ser diversas, como a simples contagem de palavras ou, mais comumente, a métrica TF-IDF (*Term Frequency Inverse Document Frequency*).

A abordagem *Bag of Words* TF-IDF cria vetores que mensuram a frequência e relevância dos termos t em um conjunto de documentos, sendo T quantas vezes o termo t aparece dentre a quantidade de documentos ni . O *Term Frequency* (TF) destaca a frequência no documento (Equação 2.9), enquanto o *Inverse Document Frequency* (IDF) ajusta para a importância do termo no conjunto de documentos (Equação 2.10). Esses valores são multiplicados para combinar frequência e raridade, gerando uma métrica de importância que leva em consideração tanto o

¹ https://www.nltk.org/howto/portuguese_en.html#stopwords

texto individualmente quanto suas palavras em relação ao conjunto de documentos (Eisenstein, 2019) (Equação 2.11). Embora simples, o TF-IDF apresenta bons resultados em diversas tarefas, sendo uma escolha interessante quando se considera o custo computacional em comparação com representações vetoriais mais complexas.

$$tf_{t,i} = \begin{cases} 1 + \log(count(t,i)), & \text{se } count(t,i) > 0 \\ 0, & \text{senão} \end{cases} \quad (2.9)$$

$$idf_t = \log \frac{T}{ni} \quad (2.10)$$

$$TF-IDF_{t,i} = tf_{t,i} * idf_t \quad (2.11)$$

A Tabela 2.1 mostra um exemplo da transformação de um conjunto de frases (documentos) em vetores a partir da técnica BOW utilizando TF-IDF. Nessa figura, podemos ver que cada documento é representado por um vetor que contém um valor para cada palavra pertencente ao conjunto de documentos e que a mesma palavra pode ter diferentes valores dependendo da sua frequência no documento.

Tabela 2.1 – Exemplo BOW TF-IDF

	Documento 1: o gato está no telhado									
	Documento 2: o cachorro está no quintal									
	Documento 3: o gato e o cachorro brincam juntos									
	brincam	cachorro	e	está	gato	juntos	no	o	quintal	telhado
Documento 1	0.000	0.000	0.000	0.433	0.433	0.000	0.433	0.336	0.000	0.569
Documento 2	0.000	0.433	0.000	0.433	0.000	0.000	0.433	0.336	0.569	0.000
Documento 3	0.424	0.323	0.424	0.000	0.323	0.424	0.000	0.501	0.000	0.000

Fonte: Criado pela autora

Mais recentemente, outras formas de representações vetoriais, baseadas em processamento semântico e redes neurais, têm sido amplamente adotadas, como a técnica de *Word Embeddings*, além dos modelos BERT e GPT. Esses modelos, ao capturarem a semântica de uma frase de maneira mais aprimorada, destacam-se também por serem pré-treinados em conjuntos de dados extensos, conferindo-lhes um conhecimento mais abrangente.

Conforme mencionado anteriormente, a abordagem de *Word Embeddings* utiliza redes neurais para gerar uma matriz na qual cada instância, representada por uma linha, é expressa por um vetor de valores. Esse vetor mapeia o significado da instância em relação a determi-

nados atributos e outras instâncias no corpus, com base na hipótese distribuída. Dessa forma, termos reconhecidos pelo algoritmo como possuindo significados semelhantes exibem valores aproximados (Mikolov *et al.*, 2013).

2.3 Caracterização Autoral

O termo "socioleto" ou "dialeto social" foi cunhado pela sociolinguística para descrever as formas como as características sociais e demográficas podem influenciar o uso da linguagem dentro de um grupo que compartilha determinada característica (Nguyen *et al.*, 2015). Assim, a partir da aplicação de métodos quantitativos e estatísticos para analisar o estilo linguístico de textos, é possível identificar o autor de um texto ou certas características do autor de um texto.

Na computação, uma área de estudo denominada Caracterização Autoral, busca prever características específicas de autores com base em seus textos, utilizando para isso técnicas de aprendizado de máquina e estatística (Neal *et al.*, 2017). Para realizar essas previsões, são consideradas características de estilo textual, informações léxicas, sintáticas, estruturais e do domínio, além de atributos baseados em caracteres (Mekala; Bulusu; Reddy, 2018). Algumas das características previstas incluem avaliação da religiosidade, determinação do posicionamento político e identificação da formação em Tecnologia da Informação (TI) (Dias, 2019).

A PAN CLEF é uma competição anual sobre estilometria e análise forense digital de textos, que não apenas apresenta uma variedade de trabalhos, mas também introduz diferentes técnicas relacionadas à caracterização autorial. Além disso, disponibiliza conjuntos de dados em diversas línguas que são amplamente utilizados em pesquisas (Bevendorff *et al.*, 2022). Desde 2013, a competição apresenta diversas tarefas na área, como predição de gênero, predição de idade, predição se um autor de um texto é humano ou um *bot*, predição da variedade linguística utilizada, entre outras (CLEF, 2024).

Nos anos de 2015 (CLEF, 2015) e 2016 (CLEF, 2016), foi introduzida a tarefa de predição de faixas etárias. Nesse caso, considera-se que as características de um texto variam conforme a idade, agrupando-se em faixas etárias a serem previstas. A premissa é que idades similares tenham características de escrita semelhantes, transformando o problema em uma tarefa de classificação. Outros trabalhos que abordam a predição dessa característica em textos em português como um problema de classificação incluem Guimarães *et al.* (2017), Silva

(2020), Dias (2019), Hsieh, Dias & Paraboni (2018), Neto (2021) e Flores, Pavan & Paraboni (2022).

3 TRABALHOS RELACIONADOS

A pesquisa realizada por Guimarães *et al.* (2017) investiga a influência da faixa etária na análise de sentimentos em plataformas de redes sociais. O estudo apresenta um classificador binário que diferencia entre adolescentes e adultos, utilizando dados não textuais dos perfis dos usuários, e explora uma variedade de métodos de classificação. Os resultados destacam a elevada eficácia do método fundamentado em redes neurais profundas e convolucionais, alcançando uma medida F1 de 94%. Além disso, o estudo indica que prever a faixa etária, mesmo sem acesso direto a essa informação, aprimora a análise de sentimentos, atingindo desempenho quase equivalente ao de modelos que utilizam a idade real.

Já o estudo realizado por Hsieh, Dias & Paraboni (2018) compara diversas representações textuais para a classificação de gênero, faixa etária, grau de religiosidade e formação em Tecnologia da Informação (TI) dentro da base de textos B5 Corpus (Ramos *et al.*, 2018)¹. Utilizando o algoritmo de regressão logística *LibLinear*, são analisadas as representações textuais *Bag of Words* (BOW), n-gramas de caracteres, frequência de palavras (TF-IDF), LIWC+P e *Word Embeddings* (*Word2vec*). Os resultados mais promissores são predominantemente obtidos com a representação TF-IDF, alcançando os melhores desempenhos, especialmente na classificação de faixa etária, com medidas F1 variando de 56% a 61%.

Em Dias (2019), os resultados de diversos algoritmos, abrangendo tanto métodos tradicionais quanto de aprendizagem profunda, são examinados para um conjunto variado de dados em várias línguas e tarefas. Destacam-se as bases textuais B5 Corpus (Ramos *et al.*, 2018)², BR Moral (Santos; Paraboni, 2019) e BlogSet BR (Santos; Woloszyn; Vieira, 2018)³, todos em português, possuindo informações sobre a faixa etária dos autores, alinhando-se com os critérios deste estudo. Diferentes combinações entre os classificadores (incluindo regressão logística e vários tipos de redes neurais profundas) e as representações textuais TF-IDF, *word embeddings* e *char embeddings* são testadas, sendo que o modelo que utiliza regressão logística com TF-IDF serve como *baseline*. A rede neural convolucional associada à vetorização TF-IDF apresenta o melhor desempenho para os conjuntos B5 Corpus e BlogSet BR, superando o *baseline*, enquanto, para o BR Moral, nenhuma das abordagens ultrapassa o *baseline*.

¹ Disponível em: <http://metatext.io/datasets/b5-corpus>

² Disponível em: <http://metatext.io/datasets/b5-corpus>

³ Disponível em: <https://www.inf.pucrs.br/linatural/wordpress/>

Expansão de Guimarães *et al.* (2017), o trabalho de Silva (2020) examina um problema de múltiplos rótulos envolvendo a classificação simultânea dos grupos etários e dos gêneros dos autores. Neste estudo, foram aplicados tanto os algoritmos *Label Powerset* quanto *Classifier Chains*, em conjunto com cinco modelos adicionais: *Decision Trees*, *Random Forests*, *Gradient Boosting*, *Support Vector Classifier* e *Complement Naive Bayes*. Os achados indicaram que os modelos de classificação baseados em regras superaram os modelos empregados em estudos anteriores. Contudo, nenhum dos modelos analisados conseguiu superar o desempenho máximo documentado em investigações passadas. Adicionalmente, foi constatado que uma seleção criteriosa de características nos modelos pode resultar em uma redução dos custos computacionais associados à análise.

Em Neto (2021), são testados diferentes modelos para a predição de gênero e idade em conjunto de dados interdomínio. Para isso, são treinados quatro modelos diferentes: regressão logística em TF-IDF utilizando n-gramas ou as palavras do texto, dependendo do conjunto de dados (esse modelo foi utilizado como baseline); um LSTM com embeddings pré-treinados NILC; um modelo *BERT*, baseado no modelo pré-treinado BERTimbau; e o modelo *bert-add* que adiciona ao *bert* a técnica de adaptação adversária com destilação. Também é testado um modelo de *stacking* de classificadores, combinando a saída do *bert-add* com a regressão logística. As bases de dados utilizadas são B2W-Reviews01, B5 Corpus, BlogSet BR, e-SIC e TwiSty e, em sua maioria, o melhor modelo encontrado é o *bert-add*. O estudo também conclui que o uso de dados de treinamento de diversas fontes é vantajoso para a categorização autoral interdomínio ao invés do treinamento com uma única fonte e teste com múltiplas fontes, que o uso do modelo de *stacking* apresenta melhoras significativas na predição e que o maior volume de dados, presentes em um modelo interdomínio, reduz as perdas de acurácia, quando comparados com um modelo de domínio único.

No estudo apresentado em Flores, Pavan & Paraboni (2022), são analisadas diversas tarefas de caracterização autoral, como gênero, faixa etária, educação e profissão, juntamente com uma análise da satisfação dos autores. Os dados utilizados são provenientes do conjunto de dados e-SIC⁴, que consiste em textos de um sistema governamental brasileiro onde os usuários fazem solicitações de documentos ao governo. Vários modelos são experimentados, desde aqueles que empregam LIWC com regressão logística até modelos como multi-layer perceptron, LSTM, BiLSTM e combinações destes com regressão logística utilizando TF-IDF de n-grams.

⁴ Disponível em: <https://drive.google.com/file/d/12sFdgiPuK2d1QyrTlnv5QwFj1Gs5mdnI/view?usp=sharing>

Além disso, modelos pré-treinados como ELMo e BERT também são avaliados. Um modelo de referência é estabelecido com base em métricas de qualidade de texto e regressão logística. O modelo LSTM se destaca como o mais eficaz na predição da faixa etária e em outras tarefas de categorização dos autores, enquanto o modelo BERT se destaca na avaliação da satisfação do usuário, ambos resultando em melhorias estatisticamente significativas.

Na Tabela 3.1, os pontos principais desses trabalhos estão resumidos.

Por fim, em Sap *et al.* (2014), é proposta a criação de um vocabulário/dicionário como uma forma de predição de gênero, como um problema de classificação, e idade, como problema de regressão. Utilizando um vocabulário consistente de mensagens do Facebook de usuários da aplicação *MyPersonality* na língua inglesa. O dicionário é composto por palavras utilizadas por pelo menos 1% dos usuários, cada qual contém um peso: para o problema de classificação de gênero (binário), o peso é obtido dos pesos de um SVM Linear com penalização L1, e, para o problema da regressão da idade, a partir dos coeficientes da regressão *Ridge* com penalização L2. O peso do dicionário indica a influência da palavra no valor ou classe final. A partir do dicionário, é, então, calculado para a instância, o *usage_lex*, que é a soma da frequência relativa da palavra em um documento vezes o peso daquela palavra para todas as palavras dentro do dicionário. Os resultados mostram-se acima dos baselines considerados, sendo a acurácia 0,92 para gênero e 0,83 para idade.

O trabalho proposto expande o trabalho de Sap *et al.* (2014), adaptando-o para lidar com um problema de faixa etária como classificação multiclasse, ao invés de considerar a idade como uma tarefa de regressão. A premissa é que, dentro de determinadas faixas etárias, autores de idades similares exibem semelhanças em seu vocabulário, as quais podem ser reveladas por meio de seus respectivos dicionários. Assim, ao contrário da classificação binária de gênero, que emprega apenas um dicionário, a classificação multiclasse de faixa etária resultará na criação de múltiplos dicionários. Estes dicionários precisam ser aplicados e interpretados adequadamente para alcançar uma atribuição dentro do modelo proposto.

Quadro 3.1 – Resumo dos trabalhos relacionados - Predição da faixa etária como um problema de classificação em textos da língua portuguesa

Trabalho	Tarefa	Dados	Técnica
Guimarães <i>et al.</i> (2017)	Faixa etária (binário)	Elementos não textuais e características textuais (Twitter)	Redes neurais artificiais, decision trees, random forest e SVM
Hsieh, Dias & Paraboni (2018)	Gênero, faixa etária , grau de religiosidade e formação de TI	Texto (b5)	Regressão logística com BOW, n-gramas, TF-IDF, LIWC+P e Word Embeddings
Dias (2019)	Gênero, faixa etária , grau de religiosidade e formação de TI	Texto (b5, BRMoral e BlogSet-BR)	Regressão logística e redes neurais com TF-IDF, word embeddings e char embeddings
Silva (2020)	Multirrotulo gênero e faixa etária (binário)	Elementos não textuais e características textuais (Twitter)	Label Powerset ou o Classifier Chains + Decision Tree, Random Forest, Gradient Boosting, SVM, Complement Naive Bayers
Neto (2021)	Predição de gênero e faixa etária interdomínio	Texto (B2W-Reviews01, b5-post, BlogSetBR, E-SIC e TwiSty)	Regressão logística com TF-IDF e n-gramas; LSTM com embeddings pré-treinados NILC; BERTimbau; bert-add
Flores, Pavan & Paraboni (2022)	Predição de gênero, faixa etária , educação e profissão. Satisfação do usuário	Texto (E-SIC)	LIWC e regressão logística, Perceptron, LSTM, BiLSTM e combinações com regressão logística e TF-IDF de n-grams, ELMo e BERT

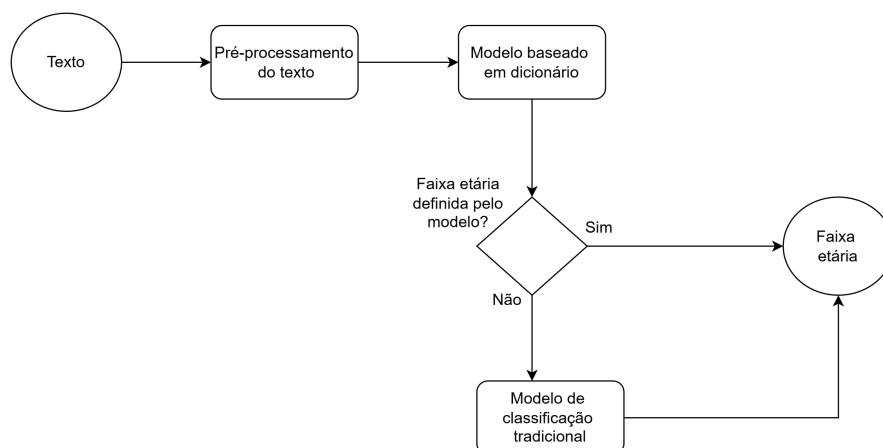
Fonte: Criado pela autora

4 METODOLOGIA

Como visto no Capítulo 3, trabalhos na literatura já exploraram o problema de predição de faixa etária de autores a partir de seus textos escritos na língua portuguesa, alguns tratando-o como um problema de classificação binária (por ex., com classes adulto e adolescente) e outros como um problema de classificação multiclasse (com cada classe representando uma faixa etária diferente). Porém, todos os trabalhos analisados lidam com o problema da predição da faixa etária utilizando técnicas tradicionais de classificação plana e processamento de linguagem natural, variando apenas o algoritmo de classificação, a base de dados e os pré-processamentos realizados nos textos. O presente trabalho propõe uma nova abordagem em cascata para classificação de faixas etárias, utilizando, além de um classificador plano tradicional, um módulo de classificação baseado em dicionários.

A Figura 4.1 ilustra o funcionamento da abordagem em cascata aqui proposta. Ela começa com o processo de pré-processamento do texto, onde são removidas pontuações, caracteres específicos do domínio (tags HTML para blogs, *hashtags* para textos extraídos de redes sociais etc.), as letras são transformadas para caixa baixa e as *stopwords* são removidas. Logo em seguida, tem-se o módulo baseado em dicionários, em que a faixa etária é definida de acordo com múltiplos dicionários de palavras. Por fim, as instâncias que os dicionários não foram capazes de classificar são encaminhadas para um classificador plano tradicional.

Figura 4.1 – Abordagem cascata proposta



Fonte: Criado pela autora

As etapas que compõem a abordagem em cascata aqui proposta são detalhadas nas Seções 4.1 e 4.2.

4.1 Modelo de Classificação Baseado em Dicionários

O modelo de classificação baseado em dicionários utilizado neste trabalho parte do pressuposto de que usuários de faixas etárias diferentes utilizam vocabulários parcialmente distintos para expressar suas ideias. Portanto, a partir das palavras utilizadas em um texto, pode-se tentar identificar a faixa etária de seu autor. O modelo aqui proposto é baseado na abordagem apresentada em Sap *et al.* (2014), onde um classificador SVM com *kernel* linear foi treinado a partir de uma base de dados textual para a criação de um dicionário de palavras associadas a pesos que indicavam a predominância das palavras em determinadas classes.

Enquanto no trabalho de Sap *et al.* (2014) um único dicionário é obtido a partir de um modelo SVM linear treinado para um problema de classificação binária, neste trabalho, uma adaptação foi necessária pelo fato de estarmos lidando com um problema de classificação multiclasse, o que faz com que mais de um dicionário tenha que ser utilizado para a definição da classe de uma instância.

O problema de classificação abordado neste trabalho envolve três classes. Portanto, a partir de uma base de dados com textos associados a três diferentes faixas etárias, um modelo SVM Linear (usando a estratégia "*one-versus-rest*") é treinado para dar origem a três dicionários de palavras que possuem pesos associados, onde esses pesos indicam a relevância das palavras para determinada(s) faixa(s) etária(s). Mais especificamente, cada dicionário contém palavras associadas a pesos positivos e negativos, sendo que o sinal e a magnitude do peso são os indicadores utilizados para definir que a palavra é predominantemente relacionada com textos escritos por pessoas de determinada(s) faixa(s) etária(s).

A definição da faixa etária do autor de um texto utilizando os dicionários construídos é feita a partir do cálculo de três índices denominados age_lex_1 , age_lex_2 e age_lex_3 (similares ao $usage_{lex}$ em (Sap *et al.*, 2014)), sendo que cada um deles é calculado a partir de um dicionário diferente. A abordagem aqui proposta adota valores limiares α_{age_1} , α_{age_2} e α_{age_3} para os índices age_lex_1 , age_lex_2 e age_lex_3 , respectivamente, de modo que um texto só é considerado pertencente a uma determinada faixa etária X se $(age_lex_X - \alpha_{age_X}) * \alpha_{age_X} \geq 0$ para $X = 1, 2$ e 3 , o que significa que:

$$\begin{cases} \alpha_{age_X} > 0 \text{ e } age_{lex_X} \geq \alpha_{age_X} \\ \text{ou} \\ \alpha_{age_X} < 0 \text{ e } age_{lex_X} \leq \alpha_{age_X} \end{cases} \quad \forall X \in \{1, 2, 3\} \quad (4.1)$$

Ao calcular esses índices (age_{lex_X}) para um determinado texto, pode ocorrer de nenhuma das três faixas etárias atender à condição especificada na Equação 4.1. Nesse caso, o texto não é classificado pelo módulo baseado em dicionários e, portanto, é encaminhado para a etapa seguinte (classificador plano tradicional).

A Equação 4.2 apresenta a como um índice age_{lex_X} , ($X = 1, 2$ ou 3) é calculado para um determinado texto. Observe que o cálculo do índice leva em conta o peso associado a cada uma de suas palavras presentes no dicionário X e o valor TF-IDF das mesmas no texto em questão.

$$age_{lex_X} = \left(\sum_{f \in lex_X} w_f * x_f \right) + w_0, \quad (4.2)$$

onde w_f representa o peso associado a uma palavra presente no dicionário X , x_f é o valor TF-IDF da palavra no texto e w_0 é uma constante denominada *intercept*.

As seções a seguir detalham o processo de criação dos dicionários (Seção 4.1.1) e o cálculo dos limiares (α_{age_X}) utilizados na Equação 4.2 (Seção 4.1.2).

4.1.1 Criação dos Dicionários

De maneira semelhante à abordagem proposta em (Sap *et al.*, 2014), um classificador SVM com *kernel* linear é utilizado para a criação de dicionários de palavras que caracterizem as faixas etárias dos autores dos textos. Neste trabalho, diferentemente de Sap *et al.* (2014), que utiliza a frequência relativa das palavras contidas no conjunto de textos de treinamento, o classificador SVM é treinado utilizando como atributos preditivos o valor TF-IDF dessas palavras. Além disso, no presente trabalho, o atributo classe pode assumir três valores distintos, cada um representando uma faixa etária diferente.

O treinamento do classificador SVM Linear resulta na criação de dicionários de palavras associadas a pesos, onde as palavras correspondem aos atributos que foram utilizados no seu treinamento. Já os pesos correspondem aos coeficientes que definem (juntamente com a

constante denominada *intercept* – w_0) cada hiperplano de separação criado pelo classificador SVM no espaço n -dimensional (onde n é o número de atributos da base de dados de treinamento). O raciocínio é que, quanto maior for o valor absoluto do peso (coeficiente na equação do hiperplano) associado a uma determinada palavra, mais relevante essa palavra é para determinada(s) classe(s). Como no problema abordado neste trabalho tem-se três classes (faixas etárias), o SVM Linear utilizando a estratégia *one vs rest* gera três hiperplanos de separação, o que significa que são gerados três dicionários de palavras associadas a pesos.

A Figura 4.2 ilustra parte dos dicionários obtidos a partir de uma base de dados real denominada e-SIC (Flores; Pavan; Paraboni, 2022). Nessa figura, no dicionário construído a partir do hiperplano que separa as instâncias da faixa etária 17 – 30 das instâncias de outra(s) classe(s) (Figura 4.2 a), as palavras associadas aos pesos positivos (marcados com cor vermelha) caracterizam a faixa etária 17 – 30. Já para o dicionário gerado a partir do hiperplano responsável pela separação das instâncias da faixa etária 31 – 42 das instâncias de outra(s) classe(s) (Figura 4.2 b), as palavras associadas aos pesos positivos (marcados com cor vermelha) caracterizam a faixa etária 31 – 42. Por fim, para o dicionário construído a partir do hiperplano que separa as instâncias da faixa etária > 42 das instâncias de outra(s) classe(s) (Figura 4.2 c), as palavras associadas aos pesos negativos (marcados com cor azul) caracterizam a faixa etária > 42 .

Figura 4.2 – Exemplo de dicionários por faixa etária - Palavras mais influentes (e-SIC)

Dicionário	17-30 Contra Todos	Dicionário	31-42 Contra Todos	Dicionário	>42 Contra Todos
estudante	3,0679	gostaria	4,0153	gostaria	3,8542
políticas	2,4931	olá	3,1929	entro	3,3359
requer	2,4859	requisito	3,0995	solicito	3,0235
005	2,4275	políticas	2,9892	sere	2,9225
ola	2,3841	prezada	2,7321	prezada	2,7449
...		...		...	
fundamenta	-3,0636	vem	-2,8622	auditoria	-2,4884
doa	-3,1396	informar	-3,0860	senhores	-2,6158
antecipadamente	-3,2596	antecipadamente	-3,6205	informar	-2,7308
senhores	-3,4990	ouvidor	-3,6280	solicitamos	-2,8185
load	-8,2057	senhores	-4,4959	ouvidor	-3,9450

a

b

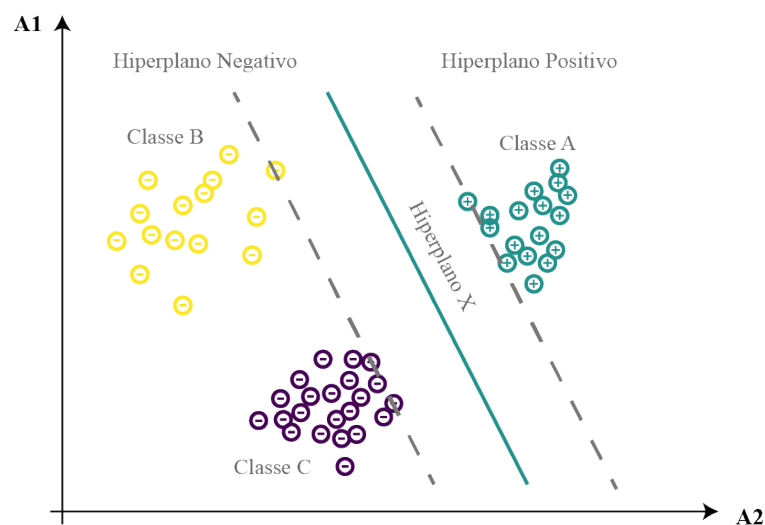
c

Fonte: Criado pela autora

A definição de qual grupo de palavras do dicionário (com pesos positivos ou negativos) caracteriza uma faixa etária é realizada verificando-se a posição das instâncias de treinamento

daquela faixa etária em relação ao hiperplano de separação. No exemplo apresentado na Figura 4.3, que contempla instâncias das classes A, B e C, num dicionário criado a partir do *Hiperplano X*, as palavras associadas a pesos positivos caracterizam instâncias da classe A, enquanto palavras associadas a pesos negativos caracterizam instâncias das classes B e C.

Figura 4.3 – Exemplo de mapeamento das instâncias e do hiperplano em um problema de classificação multiclasse



Fonte: Criado pela autora

Nesses dicionários, vale observar que, quanto maior for o valor (em módulo) do peso associado a uma palavra, mais forte é a relação dessa palavra com alguma(s) faixa(s) etária(s). Portanto, diferentemente da proposta apresentada em (Sap *et al.*, 2014), adotou-se um limiar (definido empiricamente) em relação ao peso associado às palavras para manter nos dicionários somente aquelas palavras que possuem o maior poder de discriminação para a predição das faixas etárias.

Duas estratégias foram utilizadas para essa redução da quantidade de palavras de cada um dos dicionários gerados a partir do classificador SVM Linear. A primeira estratégia mantém, em cada dicionário originalmente gerado, somente um percentual das palavras com os maiores pesos em módulo. Observe que, nessa estratégia, dependendo dos pesos associados às palavras, podem ser removidas quantidades diferentes de palavras em cada um dos conjuntos de palavras associadas aos pesos positivos e negativos. Para evitar esse desequilíbrio na quantidade de palavras removidas associadas aos pesos positivos e negativos, a segunda estratégia aqui considerada para redução de cada dicionário mantém, separadamente para cada um dos

dois conjuntos de palavras (com pesos positivos e negativos), um percentual das palavras com os maiores pesos em módulo. Portanto, nessa segunda estratégia, o mesmo percentual de palavras é removido em cada um dos dois conjuntos de palavras (associadas aos pesos positivos e negativos).

4.1.2 Cálculo dos Limiares α_{age_X}

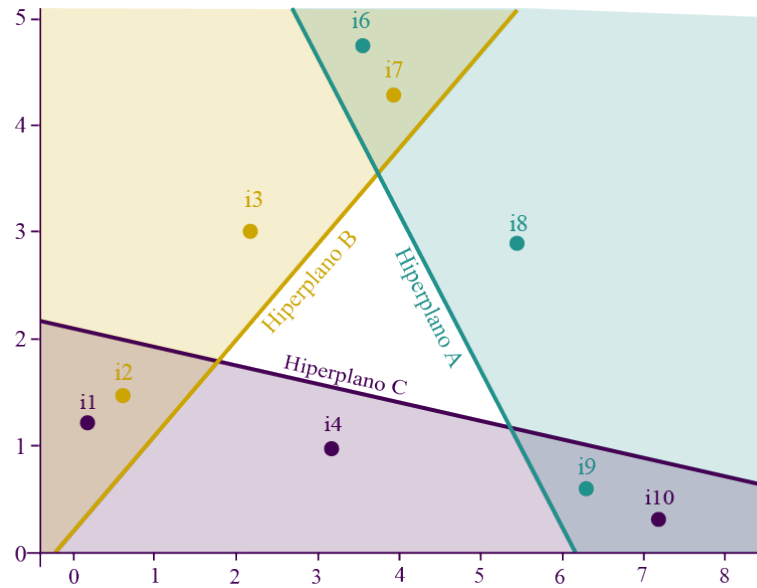
No exemplo apresentado na Figura 4.3, enquanto a posição das instâncias da classe *A* com relação ao *Hiperplano X* faz com que o valor do age_{lex_X} (ver Equação 4.2) para as instâncias dessa classe seja sempre maior do que zero ($age_{lex_X} > 0$), a posição das instâncias das classes *B* e *C* em relação a esse mesmo hiperplano faz com que o valor do age_{lex_X} (ver Equação 4.2) para as instâncias dessas duas classes seja sempre menor do que zero ($age_{lex_X} < 0$).

Assim como explicado no Capítulo 2, verificando-se a posição de uma nova instância com relação a um hiperplano de separação, define-se a sua classe. Portanto, no caso do exemplo da Figura 4.3, isso significa que se o valor do age_{lex_X} para uma nova instância for maior do que zero, então essa instância será atribuída à classe *A*. Caso contrário ($age_{lex_X} < 0$), significa que essa instância é da classe *B* ou *C*.

Considerando que o problema apresentado se caracteriza como multiclasse, a atribuição de uma instância à classe *A* requer a análise não apenas de age_{lex_A} , mas também dos valores de age_{lex_B} e age_{lex_C} . A Figura 4.4 ilustra essa questão. Observando o hiperplano *A*, nota-se que todas as instâncias da classe *A* estão posicionadas acima dele (*i6*, *i8*, *i9*), indicando que, para uma instância pertencer à classe *A*, deve-se ter $age_{lex_A} > 0$. Contudo, instâncias pertencentes à classe *B*, como por exemplo *i7*, podem também mostrar $age_{lex_A} > 0$. Assim, a atribuição da classe *A* para uma instância requer uma análise conjunta dos valores de age_{lex_A} , age_{lex_B} e age_{lex_C} . No caso considerado, observa-se que a maioria das instâncias da classe *A* apresenta $age_{lex_B} < 0$ e $age_{lex_C} > 0$. Desse modo, são definidas as seguintes regras para a atribuição de cada classe:

$$\text{Faixa etária} = \begin{cases} A, & \text{se } age_{lex_A} > 0 \text{ e } age_{lex_B} < 0 \text{ e } age_{lex_C} > 0 \\ B, & \text{se } age_{lex_A} < 0 \text{ e } age_{lex_B} > 0 \text{ e } age_{lex_C} > 0 \\ C, & \text{se } age_{lex_A} < 0 \text{ e } age_{lex_B} < 0 \text{ e } age_{lex_C} < 0 \end{cases} \quad (4.3)$$

Figura 4.4 – Exemplo de mapeamento das instâncias e dos múltiplos hiperplanos em um problema de classificação multiclasse



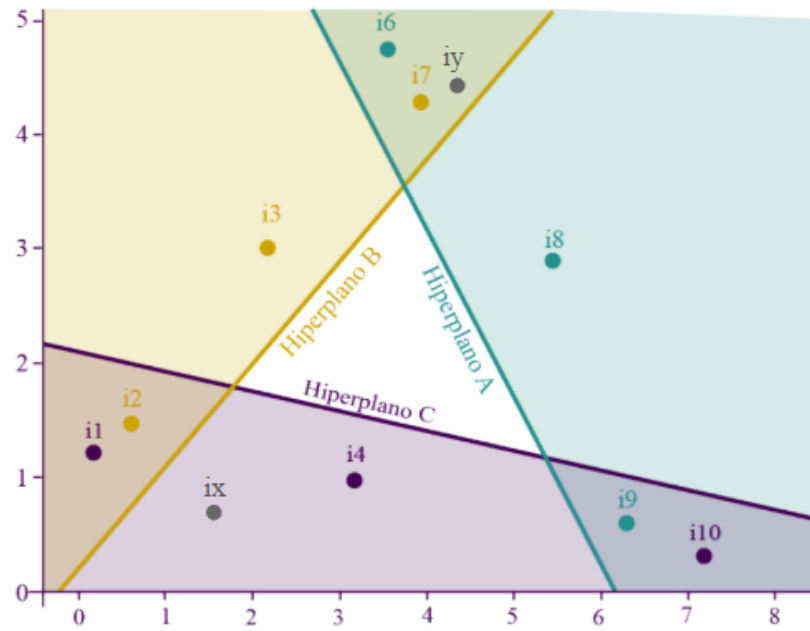
Fonte: Criado pela autora

As regras demonstradas na Equação 4.3 são formuladas com base nas instâncias de treinamento de cada classe, resultando em variações conforme a base de dados utilizada. Ademais, é importante ressaltar que uma nova instância pode não atender a nenhuma das regras criadas, impedindo que a mesma possa ser classificada por esse módulo baseado em dicionários.

Na Figura 4.5, temos exemplos de novas instâncias colocadas no espaço bidimensional com os três hiperplanos. Seguindo as regras definidas pela Equação 4.3, a instância ix seria classificada como classe C, pois está abaixo do hiperplano A ($age_lex_A < 0$), abaixo do hiperplano B ($age_lex_B < 0$) e abaixo do hiperplano C ($age_lex_C < 0$). Porém, a instância iy não consegue ser classificada em nenhuma classe, pois está acima do hiperplano A ($age_lex_A > 0$), acima do hiperplano B ($age_lex_B > 0$) e acima do hiperplano C ($age_lex_C > 0$) e, nesse caso, não atende a nenhuma das regras apresentadas na Equação 4.3.

Sendo assim, dada uma instância para ser classificada, a partir do cálculo dos valores dos seus índices age_lex_X é possível tentar definir a sua classe. No entanto, ao invés de utilizar apenas o sinal (positivo ou negativo) do valor do age_lex_X para definir a classe da instância, neste trabalho, para aumentar o grau de certeza da classificação, adota-se um limiar denominado α_{age_X} de modo que, para uma instância ser considerada como pertencente a uma deter-

Figura 4.5 – Exemplo de instâncias novas em relação aos múltiplos hiperplanos em um problema de classificação multiclasse



Fonte: Criado pela autora

minada classe, os valores dos seus índices age_lex_X devem ser maiores ou iguais a α_age_X , se $\alpha_age_X > 0$, ou menores ou iguais a α_age_X , se $\alpha_age_X < 0$ (ver Equação 4.1).

Esses limiares α_age_X são definidos a partir dos índices age_lex_X calculados para as instâncias de treinamento da classe X , ou seja, a partir das mesmas instâncias que foram utilizadas na construção do dicionário. Mais especificamente, a partir do conjunto de valores age_lex_X calculados para as instâncias de treinamento da classe X , tanto a média quanto os quartis (separatrizes que dividem o conjunto em quatro partes iguais) são considerados como possíveis valores para α_age_X . Desse modo, o valor do α_age_X pode ser igual a zero ou à média ou ao primeiro quartil ou ao segundo quartil (mediana) ou ao terceiro quartil do conjunto de valores age_lex_X das instâncias de treinamento da classe X . Dadas essas possibilidades de valores para α_age_X ($X = 1, 2$ e 3), os melhores valores para cada base de dados foram definidos empiricamente, escolhendo os melhores de acordo com o resultado obtido pela aplicação desses.

4.2 Modelo de Classificação Tradicional

A última etapa da abordagem em cascata faz uso de um modelo de classificação tradicional, seguindo o que a literatura apresenta para classificação de faixa etária como um pro-

blema multiclasse para textos encontrados na língua portuguesa. Nessa etapa, os textos pré-processados são vetorizados e utilizados no treinamento de um modelo de classificação. Assim, nessa abordagem em cascata, todas as instâncias que não são classificadas pelo modelo baseado em dicionários são classificadas por esse modelo de classificação tradicional.

5 EXPERIMENTOS COMPUTACIONAIS

Para avaliação da metodologia proposta, uma série de experimentos computacionais foi realizada. Os detalhes dos experimentos são organizados da forma descrita a seguir. Na Seção 5.1 são apresentadas as bases de dados textuais utilizadas no trabalho. Em seguida, na Seção 5.2, são descritas as configurações utilizadas na implementação da proposta, com a construção de um modelo baseado em dicionário e de um modelo de classificação tradicional baseado em trabalhos encontrados na literatura, para ser usado tanto na abordagem em cascata quanto para uma avaliação comparativa com a abordagem proposta. Por fim, na Seção 5.3, são discutidos os resultados obtidos pela abordagem proposta e o comparativo com os modelos de classificação tradicionais.

5.1 Base de Dados

Para a realização dos experimentos computacionais, foram selecionadas quatro bases de dados textuais já utilizadas em outros trabalhos da literatura. Essas bases de dados são diversificadas com relação à origem dos dados (*websites*, *blogs* e redes sociais), ao conteúdo dos textos e ao tamanho dos mesmos. Segue uma breve descrição de cada uma das bases de dados.

B5 Corpus¹: base que reúne 516 textos extraídos de postagens de usuários distintos no Facebook, com cada texto reunindo até 1000 postagens de um usuário. Neste trabalho, utilizamos somente o subconjunto dos textos da base original que continha as informações de faixa etária de seu autor. Criada por Ramos *et al.* (2018), essa base faz parte de um corpus com textos informais sobre diversos temas e foi utilizada em Hsieh, Dias & Paraboni (2018) e em Dias (2019). A Tabela 5.1 apresenta a distribuição das instâncias de cada faixa etária.

BR Moral²: base que reúne textos opinativos gerados por 510 autores distintos. Os textos abordam temas como religião, política, legalização de drogas, pena de morte e outros. Essa base de dados, disponibilizada por Santos & Paraboni (2019), foi utilizada em Dias (2019) com a distribuição das instâncias nas faixas etárias apresentada na Tabela 5.2.

¹ Disponível em: <http://metatext.io/datasets/b5-corpus>

² Santos & Paraboni (2019)

Tabela 5.1 – Distribuição das instâncias na base B5 Corpus

Faixa Etária	# Instâncias
18 - 20	182
23 - 25	189
28 - 61	145
# Total	516

Fonte: Criado pela autora

Tabela 5.2 – Distribuição das instâncias na base BR Moral

Faixa Etária	# Instâncias
0 - 23	187
24 - 30	182
31 - 99	141
# Total	510

Fonte: Criado pela autora

BlogSet BR³: base formada por 2604 textos de *blogs* publicados por diferentes usuários. Cada texto reúne uma ou mais publicações de um autor. Criada por Santos, Woloszyn & Vieira (2018), essa base foi extraída da plataforma Blogspot a partir de mais de 7 milhões de textos que abordam temas diversos, indo desde cuidados pessoais até política internacional. Essa base, que já foi utilizada em Dias (2019) e Neto (2021), possui a distribuição das instâncias nas faixas etárias mostrada na Tabela 5.3.

Tabela 5.3 – Distribuição das instâncias na base BlogSet BR

Faixa Etária	# Instâncias
10 - 25	669
26 - 40	1.021
> 40	914
# Total	2.604

Fonte: Criado pela autora

³ <https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/blogset-br/>

e-SIC⁴: base de textos obtida do Sistema Eletrônico do Serviço de Informações ao Cidadão, disponibilizado pelo governo. Criada e utilizada por Flores, Pavan & Paraboni (2022), se destaca pela grande quantidade de instâncias. Ela é estruturada de forma a conter, além da requisição feita pelo cidadão ao governo, diversas informações sobre o mesmo, incluindo a sua faixa etária (Flores; Pavan; Paraboni, 2022). A distribuição das instâncias nas faixas etárias é mostrada na Tabela 5.4.

Tabela 5.4 – Distribuição das instâncias na base e-SIC

Faixa Etária	# Instâncias
17-30	16.769
31-42	16.978
> 42	14.015
# Total	47.762

Fonte: Criado pela autora

A Tabela 5.1 apresenta um resumo das principais características das bases utilizadas nos experimentos.

Quadro 5.1 – Características da bases de dados

Base	Domínio	Instâncias	# Palavras/Texto
BlogSet BR	Blogs	2.604	5.390,18
BR Moral	Textos Opinitivos	510	427,81
B5 Corpus	Redes Sociais (Facebook)	516	2.846
e-SIC	Requisição no sistema do governo (e-Gov)	47.762	73,64

Fonte: Criado pela autora

5.2 Configuração Experimental

A implementação da abordagem em cascata aqui proposta envolve o pré-processamento dos textos das bases de dados, a construção do modelo baseado em dicionário e o treinamento de modelos de classificação tradicionais. Neste trabalho, toda implementação de código foi

⁴ Disponível em: <https://drive.google.com/file/d/12sFdgiPuK2d1QyrTlnv5QwFj1Gs5mdnI/view?usp=sharing>

Tabela 5.5 – Tamanho do vocabulário para cada base de dados

Base de dados	# vocabulário (palavras)
BlogSet BR	42.215
BR Moral	de 3.733 a 3.815 (varia com as partições)
B5 Corpus	13.869
e-SIC	2.769

Fonte: Criado pela autora

realizada utilizando a linguagem *Python*, em conjunto com módulos e bibliotecas da linguagem, tais como *nlTK*, *scikit-learn*, *keras*, entre outras. O código foi disponibilizado no *GitHub* ⁵.

Conforme mencionado no Capítulo 4, o pré-processamento do texto incluiu a remoção de *stopwords*, de pontuações e de caracteres específicos do domínio, como *hashtags* para redes sociais ou *tags* HTML para páginas da web. Além disso, as letras foram convertidas para caixa baixa. A decisão de remoção de *stopwords* foi tomada a partir da observação de que elas não contribuíam para a distinção das faixas etárias. Para a identificação e remoção das *stopwords* utilizou-se a lista de *stopwords* para a língua portuguesa presente na biblioteca *nlTK*.

Os detalhes das implementações do modelo baseado em dicionários e da construção dos modelos de classificação tradicionais são apresentados nas Seções 5.2.1 e 5.2.2.

5.2.1 Modelo Baseado em Dicionários

Após a etapa de pré-processamento dos textos, assim como realizado em (Sap *et al.*, 2014), por não agregarem valor ao processo de distinção das classes, as palavras que apareciam em menos de 1% dos textos foram removidas e, consequentemente, permitiram a redução do tamanho do vocabulário utilizado no treinamento do classificador SVM com kernel linear. A Tabela 5.5 apresenta o tamanho do vocabulário de cada base de dados após esse processo de remoção de palavras.

Em seguida, o módulo *TfidfVectorizer* da biblioteca *scikit-learn* é utilizado para transformar cada documento de texto em um vetor numérico com base na métrica TF-IDF (*Term Frequency–Inverse Document Frequency*). Esses vetores numéricos foram utilizados no trei-

⁵ https://github.com/ciceribeiroo/age_predict_dict

Tabela 5.6 – Hiperparâmetros do SVM

Base de dados	C	<i>max_iter</i>	<i>tol</i>
BlogSet BR	1	-1	0,001
BR Moral	1	-1	0,001
B5 Corpus	1	1.000	0,0001
e-SIC	1	-1	0,001

Fonte: Criado pela autora

namento do classificador SVM Linear, para o qual foi utilizado o módulo SVC da biblioteca *scikit-learn*, com o parâmetro *kernel* definido como ‘linear’ e *decision_function_shape* como ‘ovr’ (para a utilização da estratégia ‘one versus rest’). Além disso, os valores dos hiperparâmetros *C*, *max_iter* e *tol* foram definidos utilizando-se a estratégia de calibração de parâmetros denominada *Grid Search*. A Tabela 5.6 apresenta, para cada base de dados, os parâmetros que resultaram no melhor resultado de macro F1 alcançado pelo *Grid Search* utilizando validação cruzada com 10 partições para a base BR Moral e 5 partições para as demais (assim garantindo a mesma porcentagem de treinamento e teste de um holdout 80/20).

O treinamento do classificador SVM Linear resultou na criação dos dicionários de palavras associadas a pesos que são utilizados na classificação de novas instâncias. No entanto, assim como explicado no Capítulo 4, visando manter nesses dicionários somente as palavras com maior poder de discriminação das faixas etárias, duas estratégias de redução da quantidade de palavras dos dicionários foram avaliadas. Portanto, experimentos foram conduzidos mantendo 100%, 50% e 25% das palavras originalmente contidas nos dicionários. A melhor estratégia e o respectivo percentual de redução adotados para cada base de dados são apresentados na Tabela 5.7.

Por fim, a definição da faixa etária do autor de um texto, utilizando os dicionários resultantes da etapa anterior, é feita somente se os valores dos índices *age_lex_X* respeitarem os limiares α_{age_X} . Esses limiares têm como objetivo maximizar o desempenho preditivo do modelo baseado em dicionário, ainda que, para isso, tenhamos uma redução na quantidade de instâncias para as quais esse modelo consegue realizar a predição da faixa etária.

Os limiares α_{age_X} são definidos a partir dos valores dos índices *age_lex_X* calculados para as instâncias utilizadas no treinamento do classificador SVM. Mais especificamente,

Tabela 5.7 – Redução do vocabulário para cada base

Base de dados	Estratégia	% de palavras removidas
BlogSet BR	-	0%
BR Moral	-	0%
B5 Corpus	-	0%
e-SIC	Maiores valores em módulo	50%

Fonte: Criado pela autora

avaliou-se os seguintes valores para α_{age_X} : zero, a média, o primeiro quartil, o segundo quartil (mediana) e o terceiro quartil do conjunto de valores age_lex_X das instâncias de treinamento da classe X . Dadas essas possibilidades de valores para α_{age_X} ($i = 1, 2$ e 3), os melhores valores encontrados para cada base de dados estão apresentados na Tabela 13.

Quadro 5.2 – α_{age_X} adotado por base de dados

Base de Dados	Faixa Etária	α_{age_1}	α_{age_2}	α_{age_3}
BlogSet BR	Faixa Etária 1	0	0	0
	Faixa Etária 2	0	0	0
	Faixa Etária 3	0	0	0
BR Moral	Faixa Etária 1	0	0	0
	Faixa Etária 2	0	0	0
	Faixa Etária 3	0	0	0
B5 Corpus	Faixa Etária 1	Terceiro Quartil	Primeiro Quartil	Primeiro Quartil
	Faixa Etária 2	Primeiro Quartil	Terceiro Quartil	Primeiro Quartil
	Faixa Etária 3	Primeiro Quartil	Primeiro Quartil	Terceiro Quartil
e-SIC	Faixa Etária 1	Mediana	0	0
	Faixa Etária 2	0	Mediana	0
	Faixa Etária 3	0	0	Mediana

Fonte: Criado pela autora

5.2.2 Modelos de Classificação Tradicionais

Dado que a avaliação da abordagem em cascata proposta é realizada a partir de uma comparação com modelos de classificação tradicionais já apresentados na literatura, para cada base de dados utilizada, toda a configuração experimental (métodos de vetorização do texto, algoritmos de classificação e as técnicas de avaliação dos classificadores) foi exatamente a mesma

utilizada nos trabalhos adotados como referência, cujos detalhes encontram-se descritos no Capítulo 3.

A Tabela 5.3 apresenta um resumo da configuração experimental utilizada para cada base de dados, a saber: o trabalho da literatura utilizado como referência (segunda coluna), a técnica utilizada na avaliação da tarefa de classificação (terceira coluna), o método utilizado para converter o texto em um vetor numérico (quarta coluna) e o classificador adotado para a classificação da faixa etária.

Quadro 5.3 – Configuração experimental dos trabalhos selecionados

Base de Dados	Trabalho de Referência	Técnica de Avaliação	Método de Vetorização	Classificador
BlogSet BR	Dias (2019)	<i>Holdout</i> 80/20	<i>Bag of Words</i> com TF-IDF	Rede Neural Convolucional
BR Moral	Dias (2019)	10-Validação Cruzada	<i>Bag of Words</i> com TF-IDF	Regressão Logística
B5 Corpus	Dias (2019)	<i>Holdout</i> 80/20	<i>Bag of Words</i> com TF-IDF	Rede Neural Convolucional
e-SIC	Flores, Pavan & Paraboni (2022)	<i>Holdout</i> 80/20	<i>Tokenizer</i> (<i>keras</i>) ⁶	LSTM

Fonte: Criado pela autora

Além disso, ao invés de se realizar a avaliação comparativa tomando como base os resultados reportados nos trabalhos de referência, para se ter uma comparação mais justa, todas as abordagens da literatura foram reimplementadas de modo que as mesmas partições de dados utilizadas no treinamento e no teste dos modelos fossem empregadas na comparação das abordagens.

Vale ressaltar que os experimentos realizados, tanto com as abordagens reimplementadas como com a abordagem proposta, incluíram uma etapa de calibração de hiperparâmetros dos algoritmos utilizando a estratégia *Grid Search*. Os melhores hiperparâmetros encontrados e, portanto, utilizados nos experimentos, encontram-se descritos no Apêndice A. Para a reimplimentação dos trabalhos de referência, foi utilizada a biblioteca *sklearn*, para a regressão logística, e as bibliotecas *keras*, *tensorflow* e *scikeras* para as redes neurais convolucionais e a LSTM.

5.3 Resultados e Análises

Nesta seção, inicialmente serão apresentados os resultados obtidos a partir da reimplimentação das abordagens propostas na literatura, as quais serviram como base de comparação

Tabela 5.8 – Resultados da reimplementação dos trabalhos de referência

Base de Dados	Métrica	Reportado	<i>Baseline</i>
BlogSet BR	Macro F1	0,45	0,48
BR Moral	Macro F1	0,45	0,43
B5 Corpus	Macro F1	0,62	0,61
e-SIC	F1 ponderada	0,67	0,57

Fonte: Criado pela autora

com a abordagem proposta neste trabalho. Em seguida, serão apresentados e discutidos os resultados alcançados com a abordagem aqui proposta de maneira detalhada, ou seja, primeiro apresentando os desempenhos obtidos somente a partir do primeiro módulo da abordagem (modelo baseado em dicionários) e, posteriormente, o resultado final, considerando também o desempenho do modelo de classificação existente na abordagem em cascata.

Como para a base BR Moral, seguindo o trabalho de referência, o modelo é avaliado a partir de uma validação cruzada com 10 partições, os resultados apresentados nesse capítulo são a média do obtido pelas 10 partições. Os resultados individuais de cada partição estão no Apêndice B. As matrizes de confusão obtidas pelos experimentos estão no Apêndice C e os desempenhos obtidos pelos experimentos para as métricas Precisão, Revogação e F1 por classe se encontram no Apêndice D.

A Tabela 5.8 apresenta os resultados obtidos a partir da reimplementação dos trabalhos da literatura. Nessa tabela, a coluna ‘Métrica’ apresenta a métrica adotada em cada trabalho de referência. Em seguida, os resultados reportados nos artigos de referência são mostrados na coluna ‘Reportado’. Por fim, a coluna ‘*Baseline*’ apresenta os resultados obtidos a partir da reimplementação desses trabalhos de referência, conforme detalhado na Seção 5.2.2. Uma vez que o protocolo experimental adotado foi o mesmo dos trabalhos de referência, as diferenças existentes entre os resultados das reimplementações e aqueles reportados na literatura devem-se à aleatoriedade inerente ao processo de particionamento dos dados para geração dos conjuntos de treinamento e teste dos modelos em avaliação e ao processo de calibração de hiperparâmetros dos algoritmos.

A Tabela 5.9 reporta nas colunas ‘Dicionários’ o desempenho (macro F1 e acurácia) obtido pelo primeiro módulo da abordagem em cascata proposta, ou seja, o modelo baseado

em dicionários. Como esse modelo pode conseguir classificar apenas parte das instâncias da base de dados de teste, a coluna ‘Cobertura’ mostra o percentual das instâncias de teste que o modelo baseado em dicionários conseguiu classificar. Por fim, para se entender a contribuição do uso desse módulo no resultado final da abordagem em cascata, as colunas ‘Classificador Tradicional’ apresentam qual seria o desempenho do classificador tradicional (macro F1 e acurácia) utilizado no último módulo da abordagem em cascata para essas instâncias que foram classificadas pelo modelo baseado em dicionários.

Tabela 5.9 – Resultados do modelo baseado em dicionários em comparação com a abordagem de classificação tradicional

Base de Dados	Cobertura	Macro F1		Acurácia	
		Dicionários	Classificador Tradicional	Dicionários	Classificador Tradicional
BlogSet BR	83,33%	0,50	0,52	0,54	0,52
BR Moral	78,98%	0,47	0,38	0,52	0,49
B5 Corpus	10,58%	0,86	0,78	0,83	0,72
e-SIC	46,46%	0,61	0,61	0,61	0,62

Fonte: Criado pela autora

Como se pode ver nos resultados apresentados na Tabela 5.9, o modelo baseado em dicionário alcança desempenho preditivo (macro F1 e acurácia), maior ou igual àquele obtido pela abordagem de classificação tradicionalmente adotada na literatura para três das quatro bases de dados avaliadas. Mais especificamente, para as bases de dados BR Moral e B5 Corpus, ele possui desempenho superior ao do classificador tradicional para as duas medidas reportadas (macro F1 e acurácia).

Vale observar também na Tabela 5.9 a correlação existente entre os limiares α_{age_X} adotados na experimentação da abordagem em cascata (ver Tabela 13) e a cobertura do modelo, ou seja, o percentual das instâncias de teste que o modelo baseado em dicionários conseguiu classificar. Para as duas bases de dados (BlogSet BR e BR Moral) em que não houve restrição ($\alpha_{age_X} = 0$) quanto aos valores de age_{lex_X} na classificação das instâncias, as coberturas foram maiores (83,33% e 78,98%).

Por fim, o desempenho da abordagem em cascata, considerando-se os dois módulos que a compõem, é apresentado na Tabela 5.10. Nessa tabela, o desempenho preditivo (macro F1 e acurácia) da abordagem proposta (colunas ‘Abordagem em Cascata’) é comparado com o desempenho das abordagens tradicionalmente adotadas na literatura (colunas ‘Baseline’). No-

vamente, como a base de dados BR Moral é avaliada utilizando-se a técnica de 10-validação cruzada, os resultados reportados para essa base correspondem a valores médios e, por isso, a comparação entre os resultados foi realizada utilizando-se o teste estatístico de Wilcoxon (com nível de significância igual a 0,05). As demais bases de dados são avaliadas a partir de uma única partição de teste (técnica *holdout*) e, portanto, sem a aplicação de um teste de significância estatística. Para o caso da base avaliada utilizando o teste estatístico de Wilcoxon, o símbolo * indica que existe diferença com significância estatística entre o resultado alcançado pela abordagem proposta e o do *baseline*.

Tabela 5.10 – Resultados da Abordagem em Cascata e do *Baseline*

Base de dados	Macro F1		Acurácia	
	<i>Baseline</i>	Abordagem em Cascata	<i>Baseline</i>	Abordagem em Cascata
BlogSet BR	0,48	0,49	0,48	0,49
BR Moral	0,43	0,47	0,42	0,49*
B5 Corpus	0,61	0,62	0,62	0,63
e-SIC	0,57	0,57	0,57	0,57

Fonte: Criado pela autora

Os detalhes do teste estatístico de Wilcoxon (com nível de significância igual a 0,05) estão na Tabela 5.11. Essa escolha se justifica pelo fato de os resultados de ambas as estratégias serem obtidos sobre os mesmos subconjuntos de dados (partições), caracterizando dados dependentes, e pela ausência de garantia de normalidade na distribuição das métricas, especialmente considerando o número reduzido de partições ($n = 10$).

Tabela 5.11 – Resultados do Teste Estatístico entre a Abordagem em Cascata e do *Baseline*

Base de dados	Métrica	Estatística	p-value	Resultado
BR Moral	Acurácia	2	0,0249	H0 é rejeitado
BR Moral	Macro F1	13,5	0,2859	H0 é aceito

Fonte: Criado pela autora

Os resultados da Tabela 5.10 mostram que a abordagem proposta neste trabalho alcançou desempenho preditivo sempre melhor ou igual ao das abordagens dos trabalhos de referência (*baseline*). Mais especificamente, para as bases de dados BlogSet BR, BR Moral e B5 Corpus, o desempenho da abordagem em cascata foi superior ao do *baseline*. Um ponto importante a ser observado nesses resultados é que o maior ganho de desempenho com relação ao *baseline* ocorreu para a base de dados BR Moral, base em que a cobertura do modelo baseado em

dicionário foi alta (78,98%) e, obviamente, o ganho de desempenho do modelo baseado em dicionário em relação ao classificador tradicional ficou entre os maiores para as bases avaliadas (ver Tabela 5.9). Isso comprova o potencial de contribuição do módulo baseado em dicionários na melhoria do desempenho da tarefa de classificação de faixas etárias. Vale observar que, para a base B5 Corpus, apesar de o ganho de desempenho do modelo baseado em dicionário em relação ao classificador tradicional ter sido semelhante àquele obtido na base BR Moral, como a cobertura do dicionário foi baixa (10,58%), obteve-se uma pequena melhora em comparação com o *baseline*, o que indica que a cobertura do dicionário também é um fator relevante no resultado final da abordagem em cascata.

6 CONCLUSÃO

À medida que o uso da Internet se torna mais frequente e, portanto, a quantidade de dados disponível na *web* aumenta, a importância das ferramentas de mineração de texto para as estratégias de negócio também cresce. Contudo, o incremento do volume de conteúdos publicados anonimamente representa um desafio para a aquisição de informações úteis para a geração de modelos de personalização para o comércio e o marketing digital e para aplicações em setores como o forense digital. Assim, a área de estudo em Caracterização Autoral tem como foco a proposição de abordagens para inferir características de autores com base nos seus textos.

Embora diferentes estudos existam na área de Caracterização Autoral, o problema da predição da faixa etária (como um problema de classificação multiclasse) ainda possui um desempenho preditivo inferior ao de outras características, como, por exemplo, gênero. Além disso, há pouca variação na literatura nas abordagens propostas para a predição da faixa etária, sendo a maioria delas limitada à abordagem tradicional de classificação de textos.

Dessa forma, este trabalho contribui nessa área de estudo propondo e avaliando uma abordagem em cascata para a predição da faixa etária de autores de textos escritos na língua portuguesa. A abordagem proposta faz uso de dicionários de palavras para capturar características específicas do domínio textual com o objetivo de melhorar o desempenho preditivo da tarefa de predição de faixa etária.

Experimentos computacionais foram realizados com bases de dados textuais de diferentes domínios para avaliar a abordagem aqui proposta. Uma avaliação comparativa foi realizada entre a abordagem proposta e abordagens existentes em trabalhos da literatura. Os resultados encontrados demonstram que explorar o vocabulário específico de cada domínio pode contribuir para melhorar o desempenho preditivo da tarefa de predição da faixa etária. Para as bases avaliadas, a estratégia proposta neste trabalho alcançou sempre desempenhos melhores ou iguais àqueles obtidos por abordagens encontradas na literatura.

Como trabalhos futuros, pode-se avaliar a proposta aqui apresentada com bases de dados contendo textos em outros idiomas, visto que a abordagem aqui apresentada não é específica a língua portuguesa. Além disso, formas alternativas de construção dos dicionários podem ser avaliadas com o objetivo de alcançar desempenhos preditivos ainda melhores.

REFERÊNCIAS

- ADNAN, K.; AKBAR, R. An analytical study of information extraction from unstructured and multidimensional big data. **Journal of Big Data**, v. 6, 10 2019.
- BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval: The Concepts and Technology Behind Search**. [S.l.]: Addison-Wesley, 2011. ISBN 978-0-321-41691-9.
- BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação - 2ed: Conceitos e Tecnologia das Máquinas de Busca**. Bookman Editora, 2013. ISBN 9788582600498. Disponível em: <https://books.google.com.br/books?id=YWk3AgAAQBAJ>.
- BEVENDORFF, J. *et al.* Overview of pan 2022: Authorship verification, profiling irony and stereotype spreaders, and style change detection. In: _____. [S.l.]: Springer International Publishing, 2022. p. 382–394.
- CLEF, P. **Author Profiling 2015**. 2015. Acessado em: 24 jan. 2024. Disponível em: <https://pan.webis.de/clef15/pan15-web/author-profiling.html#synopsis>.
- CLEF, P. **Author Profiling 2016**. 2016. Acessado em: 24 jan. 2024. Disponível em: <https://pan.webis.de/clef16/pan16-web/author-profiling.html>.
- CLEF, P. **PAN - Shared Tasks**. 2024. Acessado em: 24 jan. 2024. Disponível em: <https://pan.webis.de/shared-tasks.html>.
- DIAS, R. F. S. **Caracterização autoral a partir de textos utilizando redes neurais artificiais**. Dissertação (Dissertação de Mestrado) — Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2019. Acessado em: 05 mai. 2023. Disponível em: <https://teses.usp.br/teses/disponiveis/100/100139/tde-24012020-202805/>.
- DOMO. **Data Never Sleeps 12.0**. 2024. <https://www.domo.com/learn/infographic/data-never-sleeps-12>. Acessado em: 15 abril de 2025.
- DONGARE, A.; KHARDE, R.; KACHARE, A. Introduction to artificial neural network. **International Journal of Engineering and Innovative Technology**, 2012.
- EISENSTEIN, J. **Introduction to Natural Language Processing**. Cambridge, MA, USA: MIT Press, 2019.
- FLORES, A. M.; PAVAN, M. C.; PARABONI, I. User profiling and satisfaction inference in public information access services. **Journal of Intelligent Information Systems**, Kluwer Academic Publishers, USA, v. 58, n. 1, p. 67–89, feb 2022. ISSN 0925-9902. Disponível em: <https://doi.org/10.1007/s10844-021-00661-w>.
- GUIMARÃES, R. G. *et al.* Age groups classification in social network using deep learning. **IEEE Access**, v. 5, p. 10805–10816, 2017.
- HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. San Francisco, CA, USA: Elsevier Science, 2011.
- HILBE, J. **Practical Guide to Logistic Regression**. New York: Taylor & Francis, 2015.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, p. 1735–1780, 11 1997.

HSIEH, F.; DIAS, R.; PARABONI, I. Author profiling from facebook corpora. In: CALZOLARI, N. *et al.* (Ed.). **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1407>.

INSIGHTS, F. B. **Business Intelligence [BI] Market Size Share | Growth, 2032**. 2025. <https://www.fortunebusinessinsights.com/business-intelligence-bi-market-103742>. Acessado em: 30 mar. 2024.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

MASTERCARD. **Age Personalization**. 2020. Acessado em: 16 jan. 2024. Disponível em: <https://www.mastercardservices.com/en/expert-insights/age-personalization>.

MEKALA, S.; BULUSU, V. V.; REDDY, R. A survey on authorship attribution approaches - volume 2. **International Journal of Computational Engineering Research (IJCER)**, v. 8, n. 5, p. 35–40, 2018.

MIKHEEV, M.; GUSYNINA, Y.; SHORNIKOVA, T. A. Building neural network for pattern recognition. In: **Proceedings of the 2020 International Russian Automation Conference (RusAutoCon)**. [S.l.]: IEEE, 2020.

MIKOLOV, T. *et al.* Efficient estimation of word representations in vector space. **Proceedings of Workshop at ICLR**, 2013.

MORAIS, J.; MERSCHMANN, L. A cascade approach for gender prediction from texts in portuguese language. In: **Proceedings of the 28th Brazilian Symposium on Multimedia and the Web**. Porto Alegre, RS, Brasil: SBC, 2022. p. 151–158. Disponível em: <https://sol.sbc.org.br/index.php/webmedia/article/view/22079>.

NEAL, T. *et al.* Surveying stylometry techniques and applications. **ACM Computing Surveys (ACM Comput. Surv.)**, v. 50, n. 1, p. 1–35, 2017.

NETO, J. P. D. **Caracterização autoral interdomínio a partir de textos**. Dissertação (Dissertação) — Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2021. Disponível em: <https://teses.usp.br/teses/disponiveis/100/100131/tde-06092021-142746/>.

NGUYEN, D. *et al.* Computational sociolinguistics: A survey. **Computational Linguistics**, v. 41, n. 4, p. 801–865, 2015.

RAMOS, R. *et al.* Building a corpus for personality-dependent natural language understanding and generation. In: **Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)**. [S.l.]: European Language Resources Association (ELRA), 2018.

SANTOS, H. d.; WOLOSZYN, V.; VIEIRA, R. Blogset-br: A brazilian portuguese blog corpus. In: **Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

SANTOS, W.; PARABONI, I. Moral stance recognition and polarity classification from twitter and elicited text. In: **Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)**. Varna, Bulgaria: INCOMA Ltd., 2019.

SAP, M. *et al.* Developing age and gender predictive lexica over social media. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1146–1151.

SILVA, D. H. **Classificação de Gêneros e Faixas Etárias em Redes Sociais Online por Meio de Técnicas de Aprendizagem Multidimensional**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Lavras, 2020.

STATISTA. **Internet usage worldwide - statistics facts**. 2025. <https://www.statista.com/topics/1145/internet-usage-worldwide/#topicOverview>. Acessado em: 11 março de 2025.

YU, S. Cyber profiling: Predicting political orientation with socmint. **Telematics and Informatics Reports**, v. 10, p. 100058, 2023. ISSN 2772-5030. Disponível em: <https://www.sciencedirect.com/science/article/pii/S277250302300018X>.

ZHANG, Q. Convolutional neural networks. In: **Proceedings of the 3rd International Conference on Electromechanical Control Technology and Transportation (ICECTT 2018)**. [S.l.: s.n.], 2018.

APÊNDICES

APÊNDICE A – ABORDAGEM TRADICIONAL - HIPERPARÂMETROS

Tabela 1 – Hiperparâmetros utilizados no classificador CNN

Hiperparâmetro	B5 Corpus	BlogSet BR
Filters		60
Kernel Size		1
Strides		1
Dropout Rate	0,1	0,6
Pool Size		4
Dense Units		512
Batch Size		32
Epochs		10
Early Stopping - Patience		5
Early Stopping - Monitor	'val_accuracy'	
Early Stopping - Min Delta		0
Early Stopping - Mode	'max'	

Fonte: Criado pela autora

Tabela 2 – Hiperparâmetros utilizados no classificador LSTM

Hiperparâmetro	e-SIC
Input Length	200
Embedding Size	256
Lstm Units	16
Dropout Rate	0,5
Learning Rate	0,001
Batch Size	128
Epochs	20
Early Stopping - Patience	5
Early Stopping - Monitor	'val_accuracy'
Early Stopping - Min Delta	0
Early Stopping - Mode	'auto'

Tabela 3 – Hiperparâmetros utilizados no classificador Regressão Logística - BR Moral

Partição	C	Penalty
1	10	'l1'
2	1	'l2'
3	1	'l2'
4	10	'l2'
5	1.000	'l1'
6	0,001	'l1'
7	1	'l2'
8	0,001	'l1'
9	1.000	'l2'
10	100	'l1'

Fonte: Criado pela autora

APÊNDICE B – EXECUÇÕES INDIVIDUAIS - BR MORAL

Tabela 4 – Resultado das Execuções Individuais do Modelo *Baseline*

Partição	1	2	3	4	5	6	7	8	9	10
Acurácia	0,47	0,45	0,43	0,41	0,33	0,37	0,45	0,37	0,45	0,42
Média	0,42									
Desvio Padrão	0,05									

Fonte: Criado pela autora

Tabela 5 – Resultado das Execuções Individuais do Modelo *Baseline*

Partição	1	2	3	4	5	6	7	8	9	10
Macro F1	0,45	0,42	0,48	0,41	0,35	0,31	0,5	0,45	0,46	0,45
Média	0,43									
Desvio Padrão	0,06									

Fonte: Criado pela autora

Tabela 6 – Resultado das Execuções Individuais do Modelo Baseado em Dicionários

Partição	1	2	3	4	5	6	7	8	9	10
Acurácia	0,5	0,56	0,63	0,53	0,56	0,5	0,41	0,46	0,61	0,48
Média	0,52									
Desvio Padrão	0,07									

Fonte: Criado pela autora

Tabela 7 – Resultado das Execuções Individuais do Modelo Baseado em Dicionários

Partição	1	2	3	4	5	6	7	8	9	10
Macro F1	0,51	0,52	0,57	0,39	0,48	0,47	0,35	0,47	0,46	0,46
Média	0,47									
Desvio Padrão	0,06									

Fonte: Criado pela autora

Tabela 8 – Resultado das Execuções Individuais do Modelo Tradicional (Instâncias classificadas pelo dicionário)

Partição	1	2	3	4	5	6	7	8	9	10
Acurácia	0,38	0,5	0,6	0,38	0,27	0,27	0,43	0,1	0,6	0,44
Média	0,4									
Desvio Padrão	0,16									

Fonte: Criado pela autora

Tabela 9 – Resultado das Execuções Individuais do Modelo Tradicional (Instâncias classificadas pelo dicionário)

Partição	1	2	3	4	5	6	7	8	9	10
Macro F1	0,47	0,52	0,49	0,41	0,45	0,19	0,38	0,20	0,38	0,32
Média	0,38									
Desvio Padrão	0,16									

Fonte: Criado pela autora

Tabela 10 – Resultado das Execuções Individuais da Cascata

Partição	1	2	3	4	5	6	7	8	9	10
Acurácia	0,47	0,55	0,63	0,49	0,47	0,45	0,42	0,39	0,6	0,47
Média	0,49									
Desvio Padrão	0,08									

Fonte: Criado pela autora

Tabela 11 – Resultado das Execuções Individuais da Cascata

Partição	1	2	3	4	5	6	7	8	9	10
Macro F1	0,48	0,51	0,57	0,46	0,46	0,41	0,38	0,36	0,57	0,45
Média	0,47									
Desvio Padrão	0,07									

Fonte: Criado pela autora

APÊNDICE C – RESULTADOS - MATRIZ DE CONFUSÃO

Quadro .1 – BlogSet BR - Matriz de confusão para a faixa etária 1 (10-25), faixa etária 2 (26-40) e faixa etária 3 (>40) para a abordagem tradicional (somente instâncias classificadas pelo modelo baseado em dicionário).

Tradicional		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	18	14	3
	Faixa Etária 2	7	16	8
	Faixa Etária 3	0	8	9

Fonte: Criado pela autora

Quadro .2 – BlogSet BR - Matriz de confusão para a faixa etária 1 (10-25), faixa etária 2 (26-40) e faixa etária 3 (>40) para modelo baseado em dicionários.

Dicionário		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	27	4	4
	Faixa Etária 2	10	6	15
	Faixa Etária 3	3	2	12

Fonte: Criado pela autora

Quadro .3 – BlogSet BR - Matriz de confusão para a faixa etária 1 (10-25), faixa etária 2 (26-40) e faixa etária 3 (>40) para o *baseline*.

<i>Baseline</i>		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	55	52	25
	Faixa Etária 2	33	98	68
	Faixa Etária 3	17	97	93

Fonte: Criado pela autora

Quadro .4 – BlogSet BR - Matriz de confusão para a faixa etária 1 (10-25), faixa etária 2 (26-40) e faixa etária 3 (>40) para a abordagem cascata.

Cascata		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	64	42	26
	Faixa Etária 2	36	88	75
	Faixa Etária 3	20	61	96

Fonte: Criado pela autora

Quadro .5 – BR Moral - Matriz de confusão para a faixa etária 1 (0-23), faixa etária 2 (24-30) e faixa etária 3 (31-99) para a abordagem tradicional (somente instâncias classificadas pelo modelo baseado em dicionário). Somatório das 10 partições.

Tradicional		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	96	56	11
	Faixa Etária 2	69	66	21
	Faixa Etária 3	28	32	27

Fonte: Criado pela autora

Quadro .6 – BR Moral - Matriz de confusão para a faixa etária 1 (0-23), faixa etária 2 (24-30) e faixa etária 3 (31-99) para modelo baseado em dicionários. Somatório das 10 partições.

Dicionário		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	113	42	8
	Faixa Etária 2	87	55	14
	Faixa Etária 3	42	27	18

Fonte: Criado pela autora

Quadro .7 – BR Moral - Matriz de confusão para a faixa etária 1 (0-23), faixa etária 2 (24-30) e faixa etária 3 (31-99) para o *baseline*. Somatório das 10 partições.

Baseline		Classe Predit		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	98	66	23
	Faixa Etária 2	74	68	40
	Faixa Etária 3	31	55	55

Fonte: Criado pela autora

Quadro .8 – BR Moral - Matriz de confusão para a faixa etária 1 (0-23), faixa etária 2 (24-30) e faixa etária 3 (31-99) para a abordagem cascata. Somatório das 10 partições.

Cascata		Classe Predit		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	108	72	7
	Faixa Etária 2	58	109	15
	Faixa Etária 3	54	52	35

Fonte: Criado pela autora

Quadro .9 – B5 Corpus - Matriz de confusão para a faixa etária 1 (18-20), faixa etária 2 (23-25) e faixa etária 3 (28-61) para a abordagem tradicional (somente instâncias classificadas pelo modelo baseado em dicionário).

Tradicional		Classe Predit		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	5	0	0
	Faixa Etária 2	2	3	0
	Faixa Etária 3	0	0	1

Fonte: Criado pela autora

Quadro .10 – B5 Corpus - Matriz de confusão para a faixa etária 1 (18-20), faixa etária 2 (23-25) e faixa etária 3 (28-61) para modelo baseado em dicionários.

Dicionário		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	5	0	0
	Faixa Etária 2	3	2	0
	Faixa Etária 3	0	0	1

Fonte: Criado pela autora

Quadro .11 – B5 Corpus - Matriz de confusão para a faixa etária 1 (18-20), faixa etária 2 (23-25) e faixa etária 3 (28-61) para o *baseline*.

Baseline		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	29	3	5
	Faixa Etária 2	16	15	7
	Faixa Etária 3	7	2	20

Fonte: Criado pela autora

Quadro .12 – B5 Corpus - Matriz de confusão para a faixa etária 1 (18-20), faixa etária 2 (23-25) e faixa etária 3 (28-61) para a abordagem cascata.

Cascata		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	29	3	5
	Faixa Etária 2	15	16	7
	Faixa Etária 3	7	2	20

Fonte: Criado pela autora

Quadro .13 – e-SIC - Matriz de confusão para a faixa etária 1 (17-30), faixa etária 2 (31-42) e faixa etária 3 (>42) para a abordagem tradicional (somente instâncias classificadas pelo modelo baseado em dicionário).

Tradicional		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	1105	345	202
	Faixa Etária 2	371	927	287
	Faixa Etária 3	218	277	706

Fonte: Criado pela autora

Quadro .14 – e-SIC - Matriz de confusão para a faixa etária 1 (17-30), faixa etária 2 (31-42) e faixa etária 3 (>42) para modelo baseado em dicionários.

Dicionário		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	1142	344	166
	Faixa Etária 2	456	790	339
	Faixa Etária 3	238	168	795

Fonte: Criado pela autora

Quadro .15 – e-SIC - Matriz de confusão para a faixa etária 1 (17-30), faixa etária 2 (31-42) e faixa etária 3 (>42) para o *baseline*.

<i>Baseline</i>		Classe Predita		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	2059	779	516
	Faixa Etária 2	832	1877	687
	Faixa Etária 3	619	668	1516

Fonte: Criado pela autora

Quadro .16 – e-SIC - Matriz de confusão para a faixa etária 1 (17-30), faixa etária 2 (31-42) e faixa etária 3 (>42) para a abordagem cascata.

Cascata		Classe Preditá		
		Faixa Etária 1	Faixa Etária 2	Faixa Etária 3
Classe real	Faixa Etária 1	2096	778	480
	Faixa Etária 2	917	1740	739
	Faixa Etária 3	639	559	1605

Fonte: Criado pela autora

APÊNDICE D – RESULTADOS - MÉTRICAS DE AVALIAÇÃO POR CLASSE

Tabela 12 – Resultados do modelo baseado em dicionários em comparação com a abordagem de classificação tradicional - Por classe

Base de Dados	Faixa Etária	Tradicional			Dicionário		
		Precisão	Revogação	F1	Precisão	Revogação	F1
BlogSet BR	Faixa Etária 1	0,72	0,51	0,60	0,68	0,77	0,72
	Faixa Etária 2	0,42	0,52	0,46	0,50	0,19	0,28
	Faixa Etária 3	0,45	0,53	0,49	0,39	0,51	0,53
BR Moral	Faixa Etária 1	0,44	0,0,71	0,54	0,53	0,60	0,56
	Faixa Etária 2	0,40	0,39	0,39	0,47	0,0,62	0,55
	Faixa Etária 3	0,50	0,19	0,26	0,68	0,0,21	0,32
BS Corpus	Faixa Etária 1	0,63	1,00	0,77	0,71	0,60	0,83
	Faixa Etária 2	1,00	0,40	0,57	1,00	1,00	0,75
	Faixa Etária 3	1,00	1,00	1,00	1,00	0,69	1,00
c-SIC	Faixa Etária 1	0,65	0,67	0,66	0,62	0,69	0,65
	Faixa Etária 2	0,60	0,58	0,59	0,61	0,50	0,55
	Faixa Etária 3	0,59	0,59	0,59	0,61	0,66	0,64

Fonte: Criado pela autora

Tabela 13 – Resultados da Abordagem em Cascata e do *Baseline* - Por classe

Base de Dados	Faixa Etária	<i>Baseline</i>			Cascata		
		Precisão	Revogação	F1	Precisão	Revogação	F1
BlogSet BR	Faixa Etária 1	0,52	0,42	0,46	0,53	0,48	0,51
	Faixa Etária 2	0,45	0,49	0,47	0,46	0,44	0,45
	Faixa Etária 3	0,50	0,53	0,51	0,49	0,49	0,49
BR Moral	Faixa Etária 1	0,48	0,47	0,50	0,51	0,56	0,52
	Faixa Etária 2	0,36	0,39	0,38	0,43	0,62	0,52
	Faixa Etária 3	0,5	0,41	0,39	0,59	0,33	0,41
B5 Corpus	Faixa Etária 1	0,56	0,78	0,66	0,57	0,78	0,65
	Faixa Etária 2	0,75	0,39	0,52	0,76	0,42	0,54
	Faixa Etária 3	0,63	0,69	0,66	0,63	0,69	0,66
e-SIC	Faixa Etária 1	0,59	0,61	0,60	0,57	0,62	0,60
	Faixa Etária 2	0,56	0,55	0,56	0,57	0,51	0,54
	Faixa Etária 3	0,56	0,54	0,55	0,57	0,57	0,57

Fonte: Criado pela autora

Tabela 14 – BR Moral - Resultados do modelo baseado em dicionários em comparação com a abordagem de classificação tradicional - Por classe e por partição

Partição	Faixa Etária	Tradicional			Dicionário		
		Precisão	Revogação	F1	Precisão	Revogação	F1
1	1	0,50	0,41	0,45	0,47	0,47	0,47
	2	0,33	0,36	0,34	0,47	0,57	0,52
	3	0,56	0,71	0,63	0,75	0,43	0,55
2	1	0,44	0,42	0,43	0,59	0,84	0,70
	2	0,41	0,50	0,45	0,50	0,54	0,52
	3	0,64	0,50	0,56	0,75	0,23	0,35
3	1	0,53	0,71	0,61	0,65	0,93	0,76
	2	0,53	0,47	0,50	0,65	0,65	0,65
	3	0,57	0,40	0,47	0,50	0,20	0,29
4	1	0,68	0,89	0,77	0,71	0,59	0,65
	2	0,57	0,67	0,62	0,48	0,73	0,58
	3	0,60	0,21	0,32	0,00	0,00	0,00
5	1	0,44	0,53	0,48	0,53	0,60	0,56
	2	0,46	0,31	0,37	0,63	0,63	0,63
	3	0,43	0,60	0,50	0,33	0,20	0,25
6	1	0,40	1,00	0,57	0,53	0,56	0,55
	2	0,00	0,00	0,00	0,43	0,60	0,50
	3	0,00	0,00	0,00	1,00	0,22	0,36
7	1	0,48	0,72	0,58	0,43	0,50	0,45
	2	0,38	0,38	0,38	0,36	0,50	0,42
	3	1,00	0,10	0,18	1,00	0,10	0,18
8	1	0,41	1,00	0,59	0,45	0,59	0,51
	2	0,00	0,00	0,00	0,41	0,39	0,40
	3	0,00	0,00	0,00	1,00	0,33	0,50
9	1	0,40	0,71	0,51	0,59	0,71	0,65
	2	0,58	0,39	0,47	0,65	0,83	0,73
	3	0,25	0,11	0,15	0,00	0,00	0,00
10	1	0,38	0,38	0,38	0,53	0,50	0,52
	2	0,35	0,43	0,39	0,41	0,64	0,50
	3	0,22	0,17	0,19	0,60	0,25	0,35

Fonte: Criado pela autora

Tabela 15 – BR Moral - Resultados da Abordagem em Cascata e do *Baseline* - Por classe e por partição

Partição	Faixa Etária	<i>Baseline</i>			Cascata		
		Precisão	Revogação	F1	Precisão	Revogação	F1
1	1	0,50	0,41	0,45	0,47	0,47	0,47
	2	0,33	0,36	0,34	0,47	0,57	0,52
	3	0,56	0,71	0,63	0,75	0,43	0,55
2	1	0,73	0,84	0,78	0,63	0,63	0,63
	2	0,44	0,54	0,48	0,46	0,77	0,57
	3	0,71	0,38	0,50	0,75	0,23	0,35
3	1	0,53	0,71	0,61	0,65	0,93	0,76
	2	0,53	0,47	0,50	0,65	0,65	0,65
	3	0,57	0,40	0,47	0,50	0,20	0,29
4	1	0,67	0,59	0,63	0,63	0,59	0,61
	2	0,46	0,67	0,54	0,05	0,67	0,57
	3	0,00	0,00	0,00	0,00	0,00	0,00
5	1	0,44	0,53	0,48	0,53	0,60	0,56
	2	0,46	0,31	0,37	0,63	0,63	0,63
	3	0,43	0,60	0,50	0,33	0,20	0,25
6	1	0,39	0,44	0,41	0,53	0,56	0,55
	2	0,32	0,33	0,32	0,43	0,60	0,50
	3	0,33	0,22	0,27	1,00	0,22	0,36
7	1	0,44	0,61	0,51	0,43	0,50	0,46
	2	0,47	0,44	0,45	0,36	0,50	0,42
	3	1,00	0,40	0,57	1,00	0,10	0,18
8	1	0,56	0,82	0,67	0,46	0,59	0,51
	2	0,43	0,33	0,39	0,41	0,39	0,40
	3	0,33	0,17	0,22	1,00	0,33	0,50
9	1	0,40	0,71	0,51	0,59	0,71	0,65
	2	0,58	0,39	0,47	0,65	0,83	0,73
	3	0,25	0,11	0,15	0,00	0,00	0,00
10	1	0,38	0,38	0,38	0,53	0,50	0,52
	2	0,35	0,43	0,39	0,41	0,64	0,50
	3	0,22	0,17	0,19	0,60	0,25	0,35

Fonte: Criado pela autora