JOÃO PEDRO MOREIRA DE MORAIS

UMA ABORDAGEM EM CASCATA PARA PREDIÇÃO DE GÊNERO A PARTIR DE TEXTOS EM PORTUGUÊS

LAVRAS – MG

2021

JOÃO PEDRO MOREIRA DE MORAIS

UMA ABORDAGEM EM CASCATA PARA PREDIÇÃO DE GÊNERO A PARTIR DE TEXTOS EM PORTUGUÊS

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

Prof. Dr. Luiz Henrique de Campos Merschmann
Orientador

LAVRAS – MG
2021

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

Morais, João Pedro Moreira de.
Uma abordagem em cascata para predição de gênero a partir de textos em Português. / João Pedro Moreira de Morais. - 2021.
48 p.

Orientador(a): Luiz Henrique de Campos Merschmann.

Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2021.
Bibliografia.

1. Caracterização Autoral. 2. Mineração de Texto. 3. Língua Portuguesa. I. Merschmann, Luiz Henrique de Campos. II. Título.

JOÃO PEDRO MOREIRA DE MORAIS

UMA ABORDAGEM EM CASCATA PARA PREDIÇÃO DE GÊNERO A PARTIR DE TEXTOS EM PORTUGUÊS

A CASCADING APPROACH TO GENDER PREDICTION FROM PORTUGUESE TEXTS

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

APROVADA em 10 de Dezembro de 2021.

Prof. Dr. Denilson Alves Pereira UFLA
Profa. Dra. Elaine Ribeiro de Faria Paiva UFU

Prof. Dr. Luiz Henrique de Campos Merschmann
Orientador

LAVRAS – MG
2021

Dedico essa trabalho à minha mãe, pelo apoio incondicional e pelo exemplo de pessoa batalhadora e correta.

AGRADECIMENTOS

Gostaria de agradecer a todos que me acompanharam nesse anos de curso. Primeiramente aos meus pais, pelo apoio da vida toda e por me acompanharem em mais essa jornada. Sem eles eu não conseguiria. À minha namorada Camila, por me apoiar e ser minha companheira mesmo nos momentos difíceis e por não me deixar desistir. À Bhianka, que me ajudou a entender e aceitar quando era necessário. Ao meu amigo Ricardo pelos anos de amizade e carinho. Ao meu orientador Professor Luiz Henrique de Campos Merschmann, pelo profissionalismo, paciência, e por sempre acreditar em mim e no nosso trabalho. Por fim, agradeço a todos os meus amigos e familiares que contribuíram direta e indiretamente para essa conquista.

RESUMO

A área de estudo e pesquisa denominada Caracterização Autoral, cujo objetivo é analisar um texto para inferir informações a respeito do seu autor, vem sendo cada vez mais útil para diferentes setores, tais como o forense, marketing e comércio eletrônico. Apesar do crescente interesse em pesquisas nessa área, a quantidade de técnicas e ferramentas apresentadas na literatura com foco na língua portuguesa é relativamente escassa quando comparada àquela disponível para outros idiomas. Desse modo, este trabalho contribui nessa área de estudo propondo e avaliando uma abordagem em cascata, que combina um módulo que utiliza um dicionário, uma heurística de gênero e um classificador, para a predição do gênero do autor de um texto escrito em português utilizando somente o conteúdo textual. A abordagem proposta leva em consideração tanto especificidades da língua portuguesa como características de domínio dos textos. Os resultados obtidos a partir da abordagem proposta mostraram que explorar as especificidades da língua portuguesa e características de domínio dos textos pode contribuir positivamente no desempenho da tarefa de predição de gênero.

Palavras-chave: Caracterização Autoral. Gênero. Língua Portuguesa. Mineração de Texto.

ABSTRACT

Author Profiling, whose objective is the analysis of a text to uncover characteristics (e.g., gender and age) of its author, has become an important task in different areas such as forensics, marketing, and e-commerce. Although a lot of research has been conducted on this task for some widely used languages (e.g., English), there is still a lot of room for improvement in studies involving the Portuguese language. Thus, this work contributes by proposing and evaluating a cascading approach, which combines a weighted lexical approach, a heuristic and a classifier, for the gender prediction problem using only textual content written in the Portuguese language. The proposed approach takes into account both specificities of the Portuguese language and domain characteristics of the texts. The results obtained from the proposed approach showed that exploring the specificities of the Portuguese language and domain characteristics of the texts can positively contribute to the performance of the gender prediction task.

Keywords: Author Profiling. Gender. Portuguese Language. Text Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com a popularização do uso da Web, em especial das redes sociais, a quantidade de dados *online* disponível cresce a cada dia. Atualmente, textos de *blogs*, de *posts* de redes sociais, de mensagens de *chats* e outros tipos de dados não estruturados representam aproximadamente 80% de todos os dados disponíveis na internet (GUO et al., 2021). Como geralmente esses textos podem ser publicados pelas pessoas de forma anônima, a utilização de técnicas computacionais para descobrir as características dos seus autores é o foco de uma área de estudo e pesquisa denominada Caracterização Autoral (*Author Profiling*). Apesar de as características mais comumente abordadas na literatura serem gênero e idade, alguns trabalhos buscam por outras características, tais como grau de escolaridade, ocupação e até mesmo traços de personalidade (NGUYEN et al., 2013).

A caracterização autoral tem-se mostrado cada vez mais útil para diversos setores, com aplicações reais na área forense, em marketing, comércio eletrônico e outras. Por exemplo, a inferência das características dos autores a partir dos seus textos pode ser utilizada em investigações criminais para ajudar a identificar o autor de um crime, pode ajudar a melhorar as estratégias de marketing adotadas por uma empresa ou até mesmo a personalizar a oferta de produtos e serviços para um grupo de pessoas com determinadas características.

Apesar do crescente interesse pela área de Caracterização Autoral, os avanços nessa área não ocorrem de forma homogênea para os diferentes idiomas. Segundo Hsieh, Dias e Paraboni (2018), a maioria dos trabalhos da literatura se concentra em alguns poucos idiomas amplamente utilizados, como por exemplo o inglês, e ainda há um déficit de pesquisas e, consequentemente, recursos e ferramentas computacionais para idiomas como o português.

Dentre as várias características que podem ser abordadas pela caracterização autoral, o foco deste trabalho é a predição de gênero a partir de textos escritos na língua portuguesa.

A grande maioria dos trabalhos recentes presentes na literatura se baseiam na abordagem tradicional de classificação de texto para resolver o problema de predição de gênero para textos em português. Basicamente, a abordagem tradicional consiste em uma etapa de pré-processamento do texto, seguido de uma etapa de representação dos mesmos para serem processados por um algoritmo de classificação. O PAN[1], competição anual que tem como foco a estilometria textual, abordou a predição de gênero a partir de textos em português na sua edição de 2017 e, como mencionado por Rangel et al. (2017), essa abordagem tradicional foi adotada por todos os trabalhos. Além dos trabalhos publicados no PAN, essa mesma abordagem tradicional foi utilizada no trabalho de Hsieh, Dias e Paraboni (2018) e Vicente, Ba-

[1] https://pan.webis.de/

tista e Carvalho (2019), sendo que este último também incorpora alguns elementos específicos das redes sociais, como o nome do usuário e a foto de perfil. Por fim, apesar de Dias e Paraboni (2020) citarem a importância do domínio textual na classificação de gênero a partir de textos em português, eles adotam a mesma abordagem tradicional dos outros trabalhos da literatura. Resumindo, nenhum desses trabalhos chegou a explorar as especificidades da língua portuguesa ou tirar proveito do domínio textual para realizar a tarefa de predição de gênero.

A hipótese na qual se baseia este trabalho é de que explorando as características particulares da língua portuguesa e características específicas do domínio do texto é possível melhorar o desempenho preditivo da tarefa de predição de gênero a partir do conteúdo textual publicado por seu autor.

Considerando a hipótese supramencionada, as questões de pesquisa a serem respondidas neste trabalho são: 1) explorar características inerentes à língua portuguesa e de domínio do texto pode contribuir para melhoria do desempenho preditivo da tarefa de predição de gênero? e 2) como combinar essas estratégias para obter uma abordagem adequada para o problema em questão?

Visando responder as questões de pesquisa apresentadas anteriormente, a Seção 1.1 apresenta os objetivos deste trabalho.

1.1 Objetivos

O objetivo deste trabalho é propor uma abordagem que melhore os resultados presentes na literatura para predição de gênero do autor de um texto escrito na língua portuguesa utilizando somente o conteúdo textual. Esse objetivo geral pode ser subdividido nos seguintes objetivos específicos:

a) Explorar as especificidades da língua portuguesa a fim de melhorar o desempenho obtido nos trabalhos apresentados na literatura;

b) Explorar as características dos diferentes domínios textuais para auxiliar na tarefa de predição de gênero;

c) Definir como combinar diferentes estratégias de predição de gênero para obter uma abordagem eficiente para o problema em questão.

1.2 Contribuições

Desse modo, visando contribuir nessa área de estudo, o presente trabalho apresenta uma abordagem em cascata, que combina um dicionário, uma heurística e um classificador, para a predição de gênero a partir de textos escritos na língua portuguesa. Vale ressaltar que a abordagem aqui proposta faz uso somente do conteúdo textual publicado para realizar a inferência de gênero.

A abordagem proposta neste trabalho foi comparada com outras apresentadas na literatura que utilizaram somente classificadores na etapa de predição de gênero do autor de um texto. Os experimentos foram realizados com objetivo de verificar se a abordagem em cascata aqui proposta poderia contribuir com a melhoria do desempenho preditivo alcançado pelos trabalhos da literatura que foram utilizados como referência. Os resultados experimentais mostraram que a abordagem em cascata proposta neste trabalho contribuiu para o aumento do desempenho preditivo dos classificadores avaliados, comprovando a eficácia da mesma para o problema em questão.

1.3 Estrutura do Documento

Este capítulo apresentou uma introdução ao tema abordado neste trabalho, bem como a motivação e justificativas para a realização desta pesquisa. Além disso, a hipótese e as questões de pesquisa que definiram os objetivos deste trabalho também foram apresentados. O restante deste documento encontra-se organizado da forma descrita a seguir. O Capítulo 2 apresenta a fundamentação teórica com os principais conceitos relacionados a este trabalho. No Capítulo 3, alguns trabalhos relacionados ao tema desta dissertação são apresentados. Em seguida, no Capítulo 4, é apresentada a descrição da abordagem proposta. Na sequência, os experimentos computacionais são reportados no Capítulo 5, com a avaliação do método proposto, resultados e discussão. Para finalizar, no Capítulo 6 são apresentadas as considerações finais englobando uma visão geral do método proposto, as contribuições da pesquisa e possíveis ideias para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo tem como objetivo apresentar os conceitos principais para a fundamentação teórica do trabalho. A Seção 2.1 apresenta os principais conceitos e técnicas do processamento de linguagem natural. Já a Seção 2.2 apresenta as principais técnicas de representação de características textuais para modelos computacionais, sendo seguida pelas técnicas de classificação na Seção 2.3 e técnicas de avaliação dos modelos de classificação na Seção 2.4. Por fim, a Seção 2.5 aborda os principais conceitos relacionados com a caracterização autoral.

2.1 Pré-Processamento de Texto

O Processamento de Linguagem Natural (PLN) é uma subárea da Ciência da Computação que tem o seu foco em estudos voltados para a geração e compreensão automática da linguagem natural. Quando o objetivo é a compreensão da linguagem natural, faz-se necessário converter a linguagem humana em uma representação formal que possa ser manipulada por sistemas computacionais. Quando a linguagem natural está expressa na forma textual, a primeira iniciativa para se obter uma representação formal que possa ser processada por um computador é a de pré-processamento do texto. A seguir são apresentadas algumas das possíveis etapas de pré-processamento de um texto.

Tokenização: A tokenização tem como objetivo a segmentação de um texto, ou seja, separa uma sequência de caracteres utilizando os espaços, quebras de linha e/ou pontuações como delimitadores. Geralmente essa é a primeira etapa executada durante o pré-processamento de um texto, servindo como base para as etapas seguintes. A tokenização pode ser efetuada com o objetivo de segmentar os caracteres, as palavras ou até mesmo as sentenças, dependendendo da necessidade específica. A Tabela 2.1 mostra um exemplo de frase tokenizada levando em consideração as palavras e os caracteres.

Tabela 2.1 – Exemplo de tokenização de palavras e caracteres.

Estou na sua casa																	
Estou					na			sua				casa					
E	s	t	o	u		n	a		s	u	a		c	a	s	a	

Fonte: Elaborado pelo autor (2021).

N-gramas: Um n-grama pode ser considerado uma sequência contígua de n itens obtidos a partir de um texto. Esses itens podem ser palavras, caracteres ou outros símbolos presentes no texto. O valor de n é definido pelo usuário, sendo os mais comuns $n = 1$ (unigramas), $n = 2$ (bigramas) e $n = 3$ (trigramas). A Tabela 2.2 mostra os unigramas e bigramas de caracteres obtidos a partir da frase: "Olá,

como vc está?". É importante notar que no exemplo da Tabela 2.2, o caractere "_" representa os espaços em branco presentes na frase.

Tabela 2.2 – Exemplo de unigrama e bigrama de caracteres.

Olá, como vc está?																	
O	l	á	,	_	c	o	m	o	_	v	c	_	e	s	t	á	?
Ol	lá	á,	,_	_c	co	om	mo	o_	_v	vc	c_	_e	es	st	tá	á?	

Fonte: Elaborado pelo autor (2021).

Remoção de *Stopwords*: *Stopwords* são palavras que ocorrem muito frequentemente em um texto, ou seja, palavras comuns em uma determinada língua. Alguns exemplos de *stopwords* são pronomes, preposições, artigos etc. Em alguns casos, esse tipo de palavra não contribui no processamento de linguagem natural e, por isso, podem ser removidas. Isso pode diminuir consideravelmente a quantidade de palavras no texto, o que pode ser uma vantagem do ponto de vista computacional, pois a remoção de palavras comuns tende a reduzir a dimensionalidade dos dados.

Lematização : O processo de lematização consiste em transformar a palavra na sua forma mais básica, ou seja, em seu lema. A palavra *"gostaria"*, por exemplo, tem como seu lema a palavra *"gostar"*.

Etiquetagem Morfológica: Segundo Eisenstein (2019), a sintaxe de um idioma é o conjunto de regras que tem como objetivo dar sentido gramatical a um conjunto de palavras. A etiquetagem morfológica (*part of speech* – POS) consiste na análise da função gramatical de cada palavra numa frase. A língua portuguesa, por exemplo, é composta por 10 classes de palavras (adjetivos, advérbios, artigos, conjunções, interjeições, numerais, preposições, pronomes, substantivos e verbos), porém, mais classes podem ser utilizadas durante a etiquetagem morfológica. A Figura 2.1 apresenta um exemplo de etiquetagem morfológica.

Figura 2.1 – Exemplo de Etiquetagem morfológica

Texto original:

Olá	!	Será	que	poderíamos	conversar	hoje	?

Etiquetagem morfológica:

DM	PNT	V	CJ	V	V	ADV	PNT

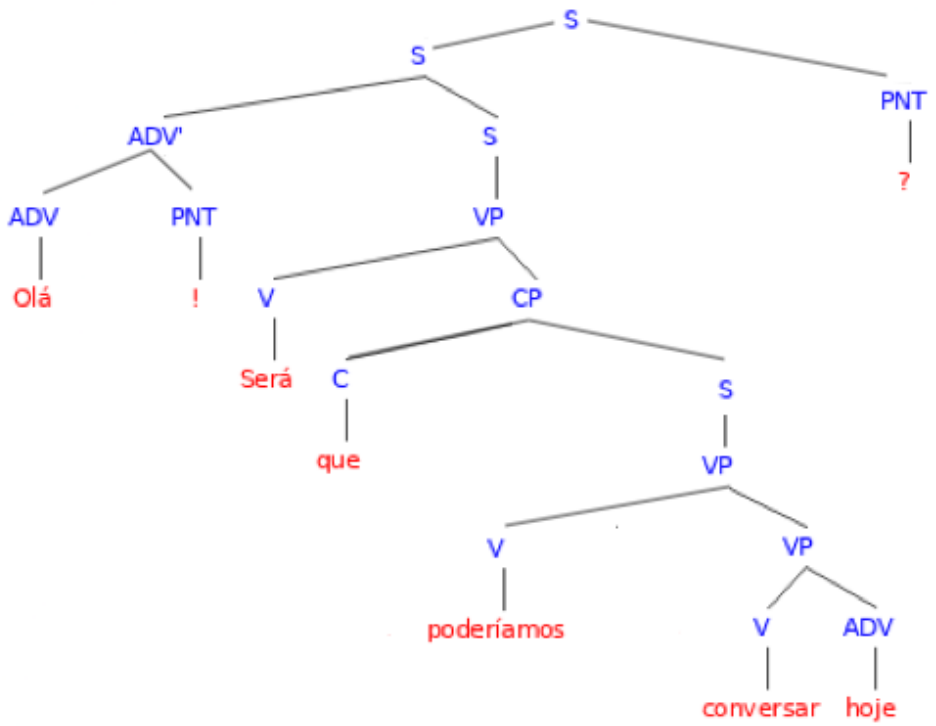
- DM: Marcador Discursivo
- PNT: Pontuação
- V: Verbo
- CJ: Conjunção
- ADV: Advérbio

Fonte: Silva et al. (2010).

***Parsing*:** Segundo Irfan et al. (2015), a técnica chamada *Parsing* tem como objetivo analisar a estrutura sintática de um texto, geralmente através da construção de uma estrutura de árvore simbolizando

a conexão entre as palavras de uma frase. A Figura 2.2 mostra o exemplo de uma árvore criada para a frase *"Olá! Será que poderíamos conversar hoje?"*.

Figura 2.2 – Exemplo de parsing



- S: Sentença
- PNT: Pontuação
- V: Verbo
- ADV': Frase Adverbial
- ADV: Advérbio
- VP: Frase Verbal
- CP: Complemento
- C: Conjunção

Fonte: Silva et al. (2010).

Treebanks: *Treebanks* são conjuntos de textos criados com anotações a respeito de cada uma das frases. Essas anotações podem ser de natureza sintática, utilizando técnicas como etiquetagem morfológica ou *parsing* ou até mesmo semântica. O projeto chamado *Universal Dependencies* (UD) (NIVRE et al., 2020) tem como objetivo criar um padrão internacional para etiquetagem morfológica e sintática, utilizando um padrão de *tags* para as diferentes características das palavras e suas ligações nas frases nos mais diversos idiomas, através da construção de um *framework* contendo *treebanks* de mais de 90 idiomas. A anotação morfológica do idioma Português no UD contém basicamente de três informações: o lema da palavra, sua etiquetagem morfológica e informações gramaticais como tempo e gênero da palavra.

2.2 Representação dos Textos

Após o pré-processamento dos textos, precisamos transformá-los para que fiquem adequados para o processamento computacional. Para ser usado por métodos de classificação, a transformação

de um texto corresponde à criação de um vetor contendo um conjunto de valores de atributos que o representa. Tradicionalmente, as palavras que compõem o texto são utilizadas como atributos. A seguir são apresentados dois tipos de representação de um texto.

Representação *Bag of Words*: Consiste em vetorizar as sentenças, transformando todo o conjunto de documentos D em uma matriz, onde cada coluna representa uma palavra do conjunto K, ou seja, o conjunto total de palavras dos textos, e cada linha, um texto. Cada valor da matriz representa a relação de cada uma das palavras com seu respectivo texto.

Várias são as métricas utilizadas para atribuir um valor a cada palavra do texto. Uma delas pode ser simplesmente o número de ocorrências da palavra k_i no texto d_j. Essa métrica representa uma relação entre os mesmos e pode ser quantificada como a frequência da palavra no texto. A Figura 2.3 ilustra esse processo.

Figura 2.3 – Exemplo de representação *bag of words*

Texto 1: Subi no muro e achei a bola.

Texto 2: E então as crianças jogaram a bola no muro.

	a	subi	no	muro	e	achei	bola	então	as	crianças	jogaram
Texto 1	1	1	1	1	1	1	1	0	0	0	0
Texto 2	1	0	1	1	1	0	1	1	1	1	1

Fonte: Elaborado pelo autor (2021).

Outra medida amplamente utilizada em tipos de representação *Bag of words* é a TF-IDF. Segundo Baeza-Yates e Ribeiro-Neto (2013), palavras ou termos que ocorrem muitas vezes num documento podem ser considerados importantes para o mesmo. Dessa maneira, Jones (1972) propõe um método que tem como objetivo calcular a importância de uma dada palavra k_i num texto d_j em relação a um conjunto de textos. A frequência do termo (TF) pode ser calculada da seguinte forma:

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \text{se } f_{i,j} > 0 \\ 0, & \text{caso contrário,} \end{cases} \tag{2.1}$$

sendo $f_{\mathrm{i,j}}$ a quantidade de vezes que uma dada palavra k_i ocorre no texto d_j. A frequencia inversa do documento (IDF), por sua vez, trata de diminuir o peso de palavras que são muito comuns nos textos, aumentando o peso das que são mais raras:

$$idf_i = \log(N/n_i) \tag{2.2}$$

sendo N o número de textos na coleção e n_i o número de textos que a palavra k_i aparece. Por fim, o TF-IDF, ou seja, o peso $w_{\mathrm{i,j}}$ da palavra k_i no texto d_{j} pode ser definido como:

$$w_{i,j} = \begin{cases} (1+\log f_{i,j}) * \log(N/n_i), & \text{se } f_{i,j} > 0 \\ 0, & \text{caso contrário} \end{cases} \tag{2.3}$$

Representação vetorial distribuída: Segundo Eisenstein (2019), um dos desafios do processamento de linguagem natural é mapear o significado das palavras. Em idiomas ricos morfologicamente, como a língua portuguesa, uma palavra pode assumir vários significados distintos em contextos diferentes. A representação vetorial distribuída tem como objetivo analisar o significado e a importância das palavras num determinado contexto a partir da geração de vetores. Um exemplo simples dessa representação seria uma matriz com a contagem de ocorrências de uma determinada palavra dentro de uma janela de tamanho específico a partir de uma palavra de interesse. A Figura 2.4 mostra um exemplo de uma matriz de ocorrências de três sentenças e janela com tamanho igual a 1. No exemplo é possível perceber que cada palavra é representada por um vetor, sendo o conjunto de palavras formando uma matriz esparsa.

Figura 2.4 – Matriz de co-ocorrências.

- Eu gosto de música
- Não gosto de esportes
- Gosto não se discute

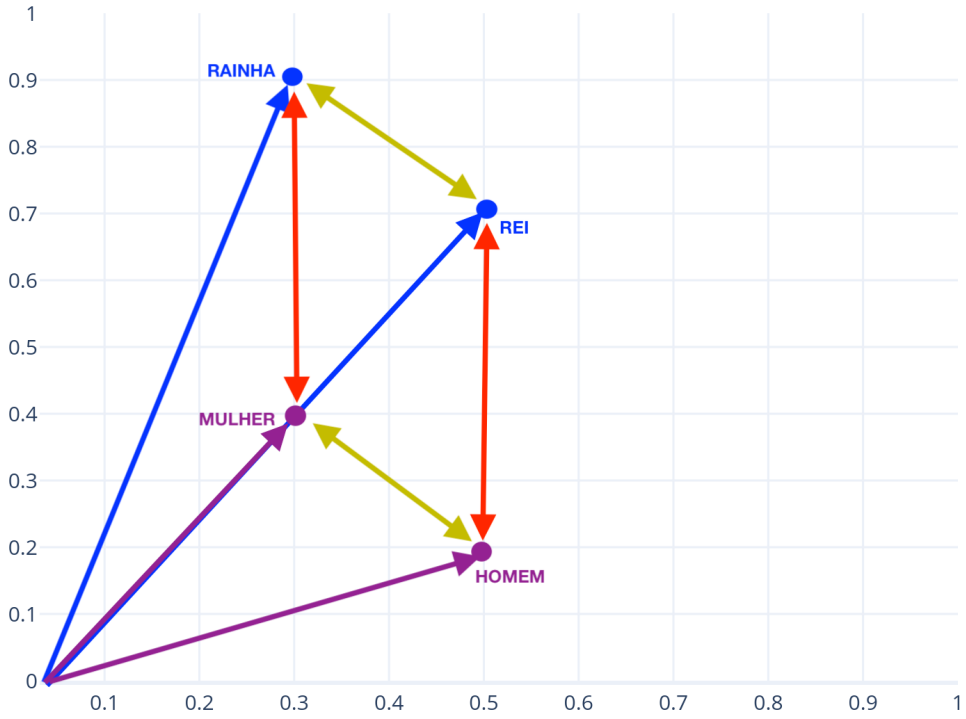
	Eu	gosto	de	música	não	esportes	se	discute
Eu	0	1	0	0	0	0	0	0
gosto	1	0	2	0	1	0	0	0
de	0	2	0	1	0	1	0	0
música	0	0	1	0	0	0	0	0
não	0	1	0	0	0	0	1	0
esportes	0	0	1	0	0	0	0	0
se	0	0	0	0	1	0	0	1
discute	0	0	0	0	0	0	1	0

Fonte: Elaborado pelo autor (2021).

A técnica proposta por Mikolov et al. (2013) conhecida como *word embeddings* cria uma representação vetorial distribuída a partir de uma abordagem mais sofisticada, com o objetivo de analisar as relações semânticas entre as palavras do texto. A técnica de *word embeddings* utiliza redes neurais artificiais para construir vetores de palavras com pesos que maximizam a probabilidade do contexto em que a palavra é observada no conjunto de textos. Com esses vetores é possível, por exemplo, medir a similaridades de palavras de acordo com suas posições no espaço multidimensional. A Figura 2.5 ilustra

um vetor 2D com a localização nos vetores das palavras *"Rainha"*, *"Rei"*, *"Mulher"* e *"Homem"*. Nesse caso, é possível analisar que a distância entre *"Rainha"* e *"Mulher"* é aproximadamente a mesma de *"Rei"* e *"Homem"* (em vermelho), bem como a distância entre *"Rainha"* e *"Rei"* é aproximadamente a mesma de *"Mulher"* e *"Homem"* (em amarelo), sendo possível medir a correlação das palavras, como por exemplo: $(\text{"}Rei\text{"} + \text{"}Mulher - \text{"}Homem\text{"} \approx \text{"}Rainha\text{"})$, ou $(\text{"}Rei - \text{"}Homem\text{"} \approx \text{"}Rainha - \text{"}Mulher\text{"})$.

Figura 2.5 – Exemplo de *Word Embeddings*



Fonte: Elaborado pelo autor (2021).

2.3 Classificadores

Segundo Weiss, Indurkhya e Zhang (2015), o objetivo dos classificadores é aprender padrões a partir de dados de exemplo fornecidos como entrada. Os algoritmos de classificação podem ser supervisionados, não-supervisionados ou semi-supervisionados.

Segundo Baeza-Yates e Ribeiro-Neto (2013), quando o aprendizado é supervisionado, a construção do modelo necessita de uma etapa de treinamento, com textos específicos para esse fim fornecidos como entrada. Essas instâncias de textos são rotuladas, ou seja, contém a informação desejada como resposta, que permite o treinamento do modelo de classificação. Os textos de treinamento são usados para construção do modelo de classificação que posteriormente é usado para predizer novas instâncias com rótulos desconhecidos. A classificação de texto supervisionada pode ser definida formalmente como: Dado uma coleção de documentos D e um conjunto $C = \{ c_1, c_2, ..., c_L \}$ com L rótulos, o classificador

textual pode ser definido pela função binária $F : D \times C \rightarrow \{0,1\}$. Se o valor for 1 para a atribuição do par $[d_j, c_p]$, com $d_j \in D$ e $c_p \in C$ então o documento d_j tem o rótulo c_p, caso contrário (se a resposta for 0), o documento d_j não tem o rótulo c_p.

Pode-se dizer que um algoritmo é não supervisionado quando os textos de treinamento fornecidos para a construção do modelo não possuem classes predefinidas, isto é, não se sabe previamente qual o rótulo de cada um dos textos. A clusterização, por exemplo, é um tipo de algoritmo que consiste em separar os documentos em clusters, ou seja, grupos. Segundo Baeza-Yates e Ribeiro-Neto (2013), uma definição formal desse algoritmo pode ser definida como: Com base numa coleção D de documentos, um algoritmo de agrupamento textual separa os documentos de acordo com uma métrica pré definida.

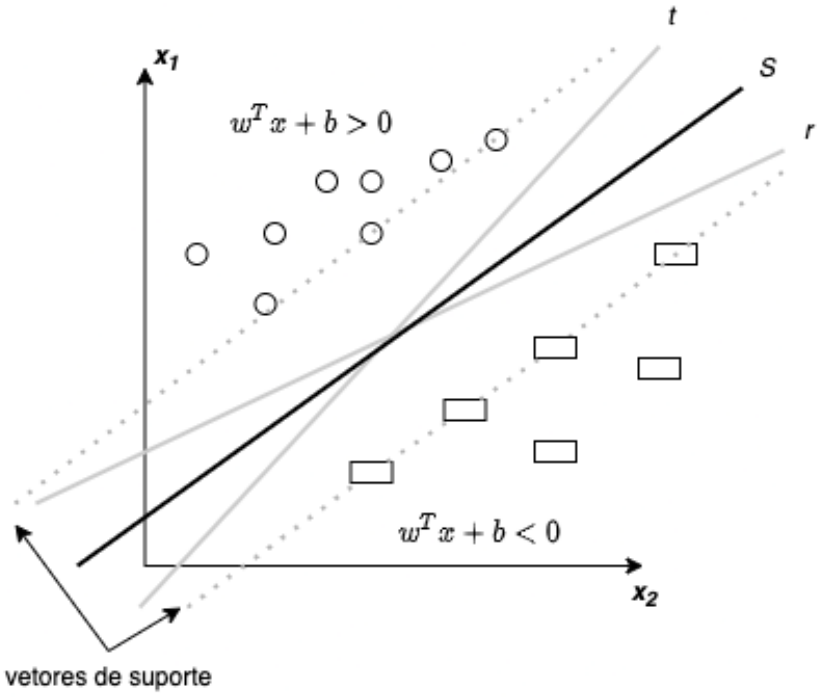
Por fim, o aprendizado semi-supervisionado trabalha mesclando modelos de classificação supervisionados e não-supervisionados. O modelo de aprendizado de máquina supervisionado muitas vezes necessita de um grande número de dados rotulados para a etapa de treinamento, no entanto, essas amostras nem sempre estão disponíveis facilmente, e em alguns casos as bases de dados precisam ser construídas manualmente com ajuda de especialistas.

Os algoritmos de classificação utilizados nesse trabalho foram essencialmente supervisionados. A seção a seguir apresenta alguns deles.

Máquina de Vetor de Suporte: Sendo amplamente utilizado na caracterização autoral, a máquina de vetor de suporte (SVM - *Support Vector Machines*), é uma técnica de classificação baseada na teoria de aprendizado estatístico proposta por Boser, Guyon e Vapnik (1992). Esse classificador tem como objetivo construir um espaço t-dimensional a partir de um conjunto de atributos disponibilizados como treinamento. A partir dessa representação, o objetivo é então a criação de um hiperplano, ou seja, uma superfície de decisão que maximize a distância entre vetores de ambos os lados do hiperplano. A partir da construção do hiperplano, novos dados podem ser submetidos ao espaço t-dimensional para serem classificados a partir de sua posição em referência ao hiperplano. A Figura 2.6 ilustra um espaço bidimensional criado a partir de dois atributos distintos x_1 e x_2. Os pontos sob a linha pontilhada representam os vetores de suporte, isto é, vetores que influenciam diretamente na superfície de decisão, e que são utilizados para maximizar a separação entre classes (no caso, S). t e r representam outras possibilidades de separação em relação aos vetores de suporte, mas que não maximizam as distâncias entre os pontos. Têm-se ainda a equação $w^T x + b$, que define qual a posição do vetor no espaço dimensional, sendo w^T o vetor de coeficientes, ou seja, o peso de cada um dos atributos no espaço dimensional, e x o vetor de atributos. A constante b, chamado de *bias* ou $intercept$, tem como objetivo mover o hiperplano a partir da origem. A posição final do vetor em relação ao hiperplano definirá o seu respectivo rótulo.

Redes Neurais Artificiais: Segundo Baeza-Yates e Ribeiro-Neto (2013), Redes Neurais Artificiais (ou RNAs) podem ser consideradas uma versão simplificada dos neurônios interconectados do

Figura 2.6 – Exemplo de máquina de vetor de suporte bidimensional com três linhas possíveis separando os pontos no espaço vetorial



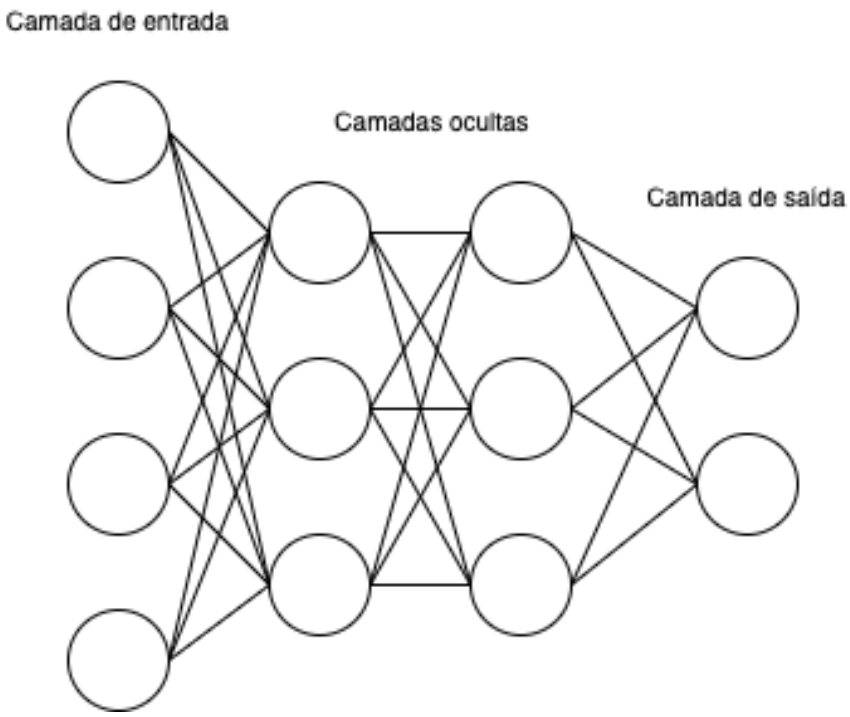
Fonte: Adaptado de Baeza-Yates e Ribeiro-Neto (2013)

cérebro humano. Numa estrutura de RNA, vários nós são conectados entre si, cada um deles contendo um peso numérico associado, indicando o sinal e a força da conexão entre eles. O perceptron é considerada a parte central de uma RNA e é onde ocorre o processamento dos dados de entrada por uma função, até a saída. Esses perceptrons podem conter uma camada para problemas linearmente separáveis, ou várias camadas ocultas para problemas complexos e linearmente inseparáveis, os chamados perceptron multicamadas (*multilayer perceptron*). A Figura 2.7 mostra um perceptron multicamadas com duas camadas ocultas. A camada de entrada é responsável por receber os dados que serão utilizados no treinamento dos pesos das camadas ocultas. Após o processamento e ponderação pelas camadas ocultas, a camada de saída tem como responsabilidade computar o resultado do processamento (FERREIRA, 2014).

Regressão Logística: Segundo Barros (2019), o classificador binário baseado no modelo linear generalizado chamado regressão logística tem como objetivo estimar a probabilidade de um evento ocorrer em função de um conjunto de variáveis independentes. O classificador parte do princípio da probabilidade da ocorrência, ou não, de um evento num intervalo de zero a um. Pode-se, por exemplo, estabelecer duas classes distintas *a* e *b*, e avaliar uma variável *x* quanto a sua probabilidade de pertencer a cada uma das classes a partir de um limiar 0.5. Caso a probabilidade condicional da variável *x* seja maior que o limiar, ela pertence a classe *a*, pertencendo a classe *b* caso contrário.

Multinomial Naive Bayes: O método probabilístico chamado *Naive Bayes* (NB) é baseado no teorema de Thomas Bayes e tem como princípio determinar a probabilidade de um evento *x* ocorrer

Figura 2.7 – Representação de um *perceptron* com camadas ocultas.



Fonte: Elaborado pelo autor (2021).

baseado em y já ter ocorrido. Na Equação 2.4, $P(B|A)$ é a probabilidade condicional de um evento B ocorrer dado que A ocorreu. $P(B)$ e $P(A)$ são as respectivas probabilidade de A e B, por fim, $P(A|B)$ é a probabilidade de A ocorrer dado que B ocorreu (BARROS, 2019).

$$P(B|A) = \frac{P(A|B)(P(B)}{P(A)} \tag{2.4}$$

Sendo amplamente utilizado na classificação de textos, o classificador *Multinomial Naive Bayes* (MNB) é considerado uma evolução do *NaiveBayes* e considera cada palavra do texto como uma variável booleana, levando em conta informações como o número de vezes que cada palavra ocorre em um documento para o cálculo da probabilidade condicional (EISENSTEIN, 2019).

Comitê de Classificadores: comitê de classificadores é a técnica de classificação focada em combinar os resultados de vários classificadores individuais num único resultado final. Esse processo de classificação pode ser dividido em três etapas, sendo a primeira delas, chamada geração, o treinamento de cada um dos classificadores individualmente, sendo possível combinações de diferentes conjuntos de atributos. Na etapa conhecida por seleção, é feita a escolha de um dos subconjuntos dos modelos criados na fase anterior. A etapa final é chamada integração, onde é feita a combinação das predições dos classificadores individuais e a apresentação do resultado, sendo essa combinação podendo ser feita por técnicas baseadas em regras pré estabelecidas ou por algoritmos de classificação (CRUZ et al., 2015).

Após a construção do modelo de classificação, isto é, escolha das técnicas de pré-processamento, representação dos textos e escolha do classificador, a etapa de avaliação tem como objetivo estimar a efetividade dos modelos. A seguir serão apresentadas algumas técnicas de avaliação.

2.4 Medidas de Avaliação

A etapa de avaliação dos modelos de classificação é crucial para a validação dos modelos propostos. Atualmente, várias são as métricas utilizadas para analisar o desempenho dos classificadores, sendo as mais comumente utilizadas a Precisão, Revocação, F1, Macro F1 e Micro F1 (BAEZA-YATES; RIBEIRO-NETO, 2013).

As métricas chamadas Precisão e Revocação estão representadas nas equações 2.5 e 2.6. Sendo VP_S (Verdadeiro Positivo) o número de instâncias em que um dado rótulo S foi classificado corretamente pelo modelo de classificação, FP_S (Falso Positivo) a quantidade de vezes que o modelo de classificação erroneamente classificou a instância como sendo da classe S e FN_S (Falso Negativo) a quantidade de instâncias que teriam o rótulo S mas o modelo de classificação as atribuiu incorretamente.

$$P_S = \frac{VP_S}{VP_S + FP_S} \tag{2.5}$$

$$R_S = \frac{VP_S}{VP_S + FN_S} \tag{2.6}$$

A partir da combinação da Precisão e a Revocação calcula-se a métrica F1. A Equação 2.7 mostra a medida-F$_S$ (F1$_S$), que corresponde à média harmônica entre a Precisão e a Revocação para o rótulo S.

$$F1_S = \frac{2 \times P_S \times R_S}{P_S + R_S} \tag{2.7}$$

Segundo Baeza-Yates e Ribeiro-Neto (2013), as métricas Macro e Micro F1 são derivações da F1. A primeira, Macro F1, leva em consideração a importância de cada rótulo no conjunto de dados, calculando a razão entre o somatório da medida F1 de cada rótulo pelo número total de rótulos, como visto na Equação 2.8. Já a Micro F1 tem como objetivo atribuir a cada instância do conjunto de dados a mesma importância, utilizando o cálculo da precisão e revocação sobre todos rótulos, como apresentado nas equações 2.9, 2.10 e 2.11.

$$F1\text{-}Macro = \frac{\sum_{i=1}^{|S|} F1_{s_i}}{|S|} \tag{2.8}$$

$$F1\text{-}Micro = \frac{2 \times P \times R}{P + R} \tag{2.9}$$

$$P = \frac{\sum_{s_i \in S} VP_s}{\sum_{s_i \in S} (VP_s + FP_s)} \tag{2.10}$$

$$R = \frac{\sum_{s_i \in S} VP_s}{\sum_{s_i \in S} (VP_s + FN_s)} \tag{2.11}$$

2.5 Caracterização Autoral

Segundo Neal et al. (2017), a caracterização autoral é uma subárea de estudo da área de pesquisa chamada estilometria textual. Essa área de pesquisa parte do princípio de que o texto e a forma de escrita de uma pessoa é quantificável e mensurável. Sendo uma área de pesquisa com mais de 100 anos de existência, tem suas raízes nos estudos da frequência do comprimento das palavras realizados por Augustus de Morgan em 1851. Apesar de outras contribuições relevantes, como a chamada Lei de Zip, descoberto por George Kingsley Zipf em 1932, os trabalhos de Mosteller e Wallace com os textos publicados entre 1787 e 1788 conhecidos como *The Federalist Papers* tem sido considerado o marco zero para a estilometria baseada em modelos computacionais (MOSTELLER; WALLACE, 1964).

A caracterização autoral é a subárea da estilometria textual que tem como objetivo inferir informações como idade e gênero do autor de um texto. Kocher e Savoy (2016) cita que mulheres têm uma tendência maior a usar pronomes pessoais (principalmente *"eu"* e *"nós"*) do que os homens. Já Peersman, Daelemans e Vaerenbergh (2011) mostram que pessoas mais velhas usam menos pronomes do que os jovens, enquanto preposições se tornam mais comuns. Essa área de estudo tem como objetivo extrair características do texto a fim de encontrar indícios de informações como gênero do autor do texto.

Segundo Reddy, Vardhan e Reddy (2016), muitas são as características presentes na escrita que podem revelar informações sobre o autor. Entre elas destacam-se as características léxicas, sintáticas, estruturais e específicas de domínio.

As chamadas características léxicas consideram o texto como um conjunto de caracteres ou palavras. No caso da análise léxica no âmbito de caracteres, algumas métricas utilizadas são o número total de caracteres utilizado no texto, número de caracteres maiúsculos, frequência do uso de uma determinada letra etc. Esse tipo de análise tem como vantagem o fato de que é independente de linguagem, além de não ser afetada por ruídos nos textos, como por exemplo, erros de escrita. Goot et al. (2018) utiliza características presentes nos caracteres como números de vogais e letras maiúsculas para predizer o gênero. Já em relação a palavras, características como erros de escrita e abreviações podem ser impor-

tantes para a análise textual, pois podem revelar informações contextuais, como visto em Hsieh, Dias e Paraboni (2018). Algumas análises efetuadas em características léxicas no que se refere a palavras são, por exemplo, a riqueza do vocabulário, estimando a diversidade das sentenças, ou a quantidade do uso de palavras abreviadas.

Com foco na análise dos padrões estruturais das frases do texto, as características sintáticas levam em consideração que, mesmo que inconscientemente, autores utilizam padrões sintáticos semelhantes mesmo em textos diferentes. Isto é, padrões sintáticos podem ser considerados uma marca do autor. Além disso, no caso de idiomas onde o gênero gramatical é bem definido, como no caso do português, o uso de expressões pode revelar, por exemplo, o gênero do autor. Mechti, Jaoua e Belguith (2013) analisa o uso de pronomes, preposições, verbos entre outros para predição de gênero em espanhol. Já Hsieh, Dias e Paraboni (2018) utiliza as características sintáticas para predizer o gênero em português, alcançando bons resultados. Diferentemente das características léxicas, a análise sintática é totalmente dependente da linguagem e necessita de ferramentas poderosas para se analisar padrões sintáticos com boa precisão.

As características estruturais, por sua vez, tem como foco estatísticas a respeito do texto em si, como tamanho dos parágrafos, palavras por frase, entre outros. Reddy, Vardhan e Reddy (2016) cita que o tamanho do texto pode ser uma característica importante para a detecção de *spam*. Vollenbroek et al. (2016) mostra ainda que o tamanho das frases pode ser um indicativo do nível de escrita do autor do texto, e utiliza, além da média do tamanho das frases, a média do tamanho das palavras para a predição de gênero.

Por fim, as características específicas de domínio têm se mostrado de grande importância na inferência de características como gênero e idade. Esse tipo de característica pode incorporar características léxicas para determinar o uso de elementos específicos de domínio, como por exemplo o uso de menções, no caso de posts de redes sociais, ou até mesmo palavras-chave que podem indicar a inclinação do autor do texto para um determinado tópico. Segundo Reddy, Vardhan e Reddy (2016), mulheres têm uma tendência em escrever mais sobre casamentos e moda, enquanto homens preferem assuntos como tecnologia e política. Sap et al. (2014) criam um dicionário com palavras que ajudam na discriminação do gênero. Dias e Paraboni (2020) realizam um estudo avaliando a importância do domínio na tarefa de predição de gênero em português.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentadas abordagens recentes propostas e avaliadas em trabalhos que focaram na predição de gênero a partir de textos escritos no idoma português.

O PAN[1], competição anual que tem como foco a estilometria textual, tem sido essencial na contribuição da área de estudo da caracterização autoral. A edição de 2017 disponibilizou uma base de dados com *tweets* em português, onde vários competidores apresentaram abordagens distintas para a tarefa de predição de gênero.

A abordagem de Basile et al. (2017), vencedores da edição de 2017 do PAN, utilizou n-gramas de caracteres e n-gramas de palavras como atributos. Para representação dos atributos foi utilizado o esquema de pesos TF-IDF. O classificador utilizado foi o SVM com kernel linear. Para validação dos resultados, utilizou-se a técnica de validação cruzada, com $k = 5$. Os autores atingiram o resultado de 84.5% de F1 na predição de gênero.

Markov, Gómez-Adorno e Sidorov (2017) utilizaram abordagem similar, utilizando além dos n-gramas convencionais e outros tipos de atributos, n-gramas tipados propostos por Sapkota et al. (2015). Para representação dos documentos, os autores utilizaram várias técnicas distintas, como *bag of words* e TF-IDF. Sobre os classificadores, foram utilizados, entre outros, o Multinomial Naive Bayes e o SVM. Testando várias combinações diferentes, os autores utilizaram validação cruzada com $k = 10$ para avaliação dos resultados, alcançando 84% de acurácia na predição de gênero em português.

Martinc et al. (2017) focaram na remoção de tags HTML, Hashtags, menções etc, além da aplicação da técnica de POS-Tagging e remoção de stopwords. O esquema de pesos utilizado foi o TF-IDF. Os autores então compararam várias técnicas diferentes. Na etapa de classificação vários foram os classificadores utilizados, entre eles SVM, Regressão Logística, Random Forest e o Extreme Gradient Boosting. Utilizando a validação cruzada com $k = 10$, os autores atingiram uma acurácia de 84% na predição de gênero.

Miura et al. (2017) focou numa abordagem diferente das dos demais competidores do PAN 2017, criando modelos baseados em combinações de redes neurais artificiais para a predição de gênero e classificação de variedade de idioma. A etapa de pré-processamento consistiu na limpeza e organização dos tweets. A criação dos modelos de redes neurais se baseou em trabalhos anteriores, combinando informações disponíveis nas palavras com informações de caracteres na construção de uma rede neural complexa. Na criação das camadas do modelo, as palavras e caracteres são consideradas representações diferentes do mesmo tweet, que são então concatenadas e processadas por diferentes redes neurais. Com

[1] https://pan.webis.de/

essa abordagem, os autores atingiram 98% na classificação da variedade do idioma português e 87% na predição de gênero.

Utilizando uma abordagem com comitês de classificadores, Ciobanu et al. (2017) avaliaram diferentes modelos na predição de gênero. Os autores criaram então várias configurações para cada modelo, alternando, por exemplo, em números de n-gramas de caracteres ou de palavras. Para os resultados, foi utilizada a técnica de Validação cruzada com $k = 3$. Os autores chegaram em 77% de acurácia na predição de gênero.

Poulston, Waseem e Stevenson (2017) também optaram pela abordagem com comitê de classificadores, combinando dois classificadores probabilísticos. O primeiro, um classificador regressão logística treinado a partir das representações de texto com n-gramas a níveis de palavras e representação de documentos com TF-IDF. Além disso, termos que apareciam em mais de 90% dos documentos foram removidos. No segundo modelo, utilizando um classificador baseado no processo Gaussiano. Para representação dos documentos foi utilizado *word embeddings*. Os autores atingiram 83,88% de acurácia na predição de gênero, e 97,63% de acurácia na predição de variedade de idioma.

Arcia et al. (2017) adotaram duas estratégias distintas para representação e classificação dos textos. A primeira, chamada de representação baseada em instâncias, agrega todos os textos de um mesmo autor em um único documento. A segunda, chamada de representação baseada em protótipo, cria um único documento para cada classe (masculino/feminino), com todos os textos de todos os autores. Após a construção das duas estratégias e pré-processamento dos textos, com remoção de *hashtags*, url's entre outros. Na etapa de classificação, o autor então calcula a similaridade dos documentos, utilizando o algoritmo de vizinhos mais próximos. A acurácia foi medida utilizando a técnica de validação cruzada. O modelo baseado em instâncias foi o que apresentou melhores resultados tanto na classificação de gênero quanto na variedade de idiomas. Para o idioma espanhol, a predição de gênero alcançou um resultado de 68% de acurácia.

Com foco em textos mais longos, Hsieh, Dias e Paraboni (2018) utiliza um corpus de postagens do Facebook na tarefa de predição de gênero. Os autores utilizaram vários modelos de representação de dados distintos, a fim de comparar seus resultados. Entre eles estão o *bag of words*, além de um modelo utilizando o esquema de pesos TF-IDF, ambos utilizando os 3 mil termos mais frequentes, entre outros. O classificador utilizado no experimento foi o regressão logística. Utilizando validação cruzada com $k = 10$, os autores reportaram que o modelo utilizando TF-IDF conseguiu melhores resultados, atingindo 88% de F1 na predição de gênero.

Vários autores utilizam mais do que somente o conteúdo textual de redes sociais para a predição de gênero em português. Vicente, Batista e Carvalho (2019), por exemplo, utilizaram, além dos textos de usuários do twitter, informações como o nome do usuário, a foto de perfil e o conteúdo da página

de perfil. Nos experimentos, os autores criaram um modelo de classificação para cada tipo de atributo (foto,nome,conteúdo textual etc), e compararam o resultado de cada classificador. Além disso, utilizaram o resultado de cada classificador como atributo para um novo classificador combinado. Os autores reportaram que o melhor resultado entre os modelos criados foi o com conteúdo textual, num modelo que consistia em n-gramas e um classificador SVM, atingindo 93.5% de acurácia. Já com o classificador combinado, isto é, um modelo criado a partir dos resultados dos modelos anteriores, os autores alcançaram 96.9% de acurácia, também utilizando um classificador SVM.

O domínio dos textos é um fator determinante na tarefa de predição de gênero. O trabalho de Dias e Paraboni (2020) se propõe a avaliar os resultados da tarefa de predição de gênero com modelos criados com textos de um único domínio e com textos de domínios distintos. Os autores utilizaram várias bases de dados textuais em português de domínios distintos, como textos de blogs pessoais (BlogSetBR), textos de opiniões pessoais (BRMoral) e solicitações governamentais (E-SIC). O impacto do domínio textual na tarefa de predição de gênero foi evidenciado pelos autores, já que quando o treinamento dos classificadores foi realizado a partir de instâncias de múltiplos domínios, o desempenho foi inferior àquele obtido quando um único domínio foi utilizado. A representação dos textos foi obtida a partir de uma média ponderada dos vetores das palavras (*word embeddings*) que compõem o texto. A medida *tf-idf* de cada palavra foi utilizada como peso para o cálculo da média ponderada dos vetores. Com a técnica de avaliação *holdout*, os melhores resultados para a medida F1 foram obtidos utilizando-se o classificador *Multilayer Perceptron*, a saber, 78% para a base BlogSetBR, 74% para a base BRMoral e 79% para a E-SIC.

Além do domínio, segundo Goot et al. (2018), o idioma do texto também é um desafio na tarefa de predição de gênero. Os autores propuseram um modelo de classificação abstrato e independente de idioma, baseado na transformação das características léxicas, como quantidade de letras maiúsculas e quantidades de vogais por palavra, em características abstratas. Utilizando bases de dados do Twitter em Alemão, Francês, Espanhol e Português, os autores extraíram características léxicas abstratas e combinaram todas as bases de dados. Para classificação, utilizaram o classificador SVM. Os autores alcançaram um resultado de 68,7% de acurácia usando validação cruzada com $k = 10$. Segundo os autores esse resultado revela que características abstratas do texto têm um potencial preditivo a ser explorado.

Por fim, Krüger e Hermann (2019) se propuseram a fazer uma revisão da literatura disponível sobre identificação de gênero e comparar as abordagens utilizadas até o momento. Para identificar os trabalhos nessa área, foi feito uma pesquisa limitada a resultados publicados entre janeiro de 2017 e janeiro de 2019. Segundo os autores, nenhuma das abordagens levou em consideração gêneros não binários, tratando o problema da classificação de gênero como um problema binário, o que pode ser um problema no uso prático dos modelos de classificação de gênero. Sobre os resultados, a melhor

abordagem para textos em português foi de Miura et al. (2017), alcançando 85,75% de acurácia na base de textos do Twitter disponibilizada pelo PAN 2017. Em textos em inglês, Markov et al. (2017) alcançou a marca de 93,4% de acurácia. Segundo os autores, apesar de serem bons números, a precisão ainda é baixa, visto que no caso da classificação de gênero em português, por exemplo, aproximadamente 14 em cada 100 pessoas seriam classificadas erroneamente, o que pode ser um problema na utilização em casos reais.

Como visto anteriormente, alguns trabalhos já abordaram o problema de predição de gênero a partir de textos em português. Apesar disso, nenhum deles levou em consideração as especificidades da língua portuguesa, que é considerada morfologicamente rica, ou o domínio textual. Portanto, este trabalho se diferencia dos demais por criar uma abordagem que leva em consideração tanto as especificidades do idioma português quanto as características de domínio do texto para, em conjunto com as técnicas já disponíveis na literatura, realizar a tarefa de predição de gênero.

O objetivo desse capítulo foi apresentar os trabalhos mais recentes disponíveis na literatura para a predição de gênero a partir de textos em português. O próximo capítulo apresenta em detalhes a abordagem proposta nesse trabalho.

4 ABORDAGEM PROPOSTA

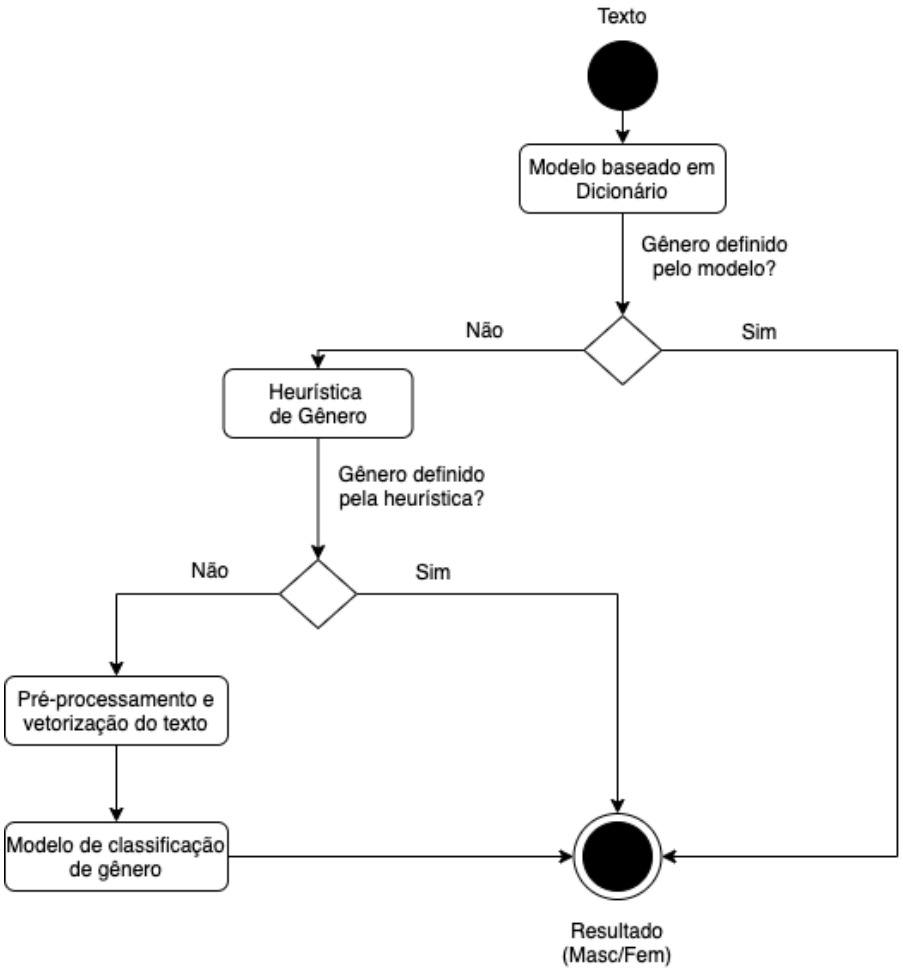
Este capítulo descreve a abordagem proposta neste trabalho. Assim como apresentado no Capítulo 3, a caracterização autoral com foco na predição de gênero já foi realizada utilizando-se diferentes abordagens para diferentes idiomas. No entanto, no contexto de aprendizagem de máquina, a abordagem que tipicamente é utilizada envolve as etapas de pré-processamento do texto, de representação do mesmo por meio de um vetor numérico e de treinamento de um modelo de classificação, com os trabalhos encontrados na literatura variando entre a escolha das técnicas utilizadas em cada uma dessas etapas. A abordagem proposta neste trabalho, ao invés de utilizar apenas um classificador para predizer o gênero do autor de um texto, propõe a utilização de uma abordagem híbrida na qual, além de um classificador, um dicionário e uma heurística são usados para auxiliar na predição do gênero.

4.1 Abordagem em Cascata

A Figura 4.1 ilustra o funcionamento da abordagem em cascata proposta neste trabalho. A abordagem proposta é constituída por três módulos que são utilizados de forma sequencial para realizar a predição do gênero do autor de um texto. O primeiro deles, chamado modelo baseado em dicionário, foi projetado a partir da abordagem proposta em Sap et al. (2014), onde um dicionário formado por palavras associadas a pesos é obtido a partir de um conjunto de treinamento (textos com o gênero do seu autor conhecido). Nesse dicionário, o peso vinculado a cada palavra indica o grau de aderência da mesma a textos escritos por autores de um determinado gênero. Nesse módulo, a partir da verificação da incidência das palavras do dicionário num texto cujo gênero do seu autor é desconhecido, pode ser possível predizer o gênero do autor do mesmo. Quando a utilização do modelo baseado em dicionário não é suficiente para realizar a predição de gênero, o texto em análise é processado pelo segundo módulo da abordagem proposta, denominado heurística de gênero. Essa heurística realiza a análise morfossintática do texto a fim de encontrar trechos que indiquem o gênero do seu autor. Caso a heurística consiga definir o gênero do autor com um certo grau de confiança, o gênero definido pela heurística é retornado como resultado final. Caso contrário, ou seja, se a heurística não conseguir definir o gênero, então o texto é processado pelo terceiro e último módulo da abordagem proposta, um modelo de classificação. Nesse caso, o texto passa por um processo de pré-processamento e é vetorizado antes de ser submetido ao modelo de classificação para a realização da predição.

Cada um dos módulos da abordagem proposta têm suas especificidades e trabalham separadamente a fim de obter o resultado final. As seções a seguir apresentam cada um deles em detalhes.

Figura 4.1 – Abordagem proposta



Fonte: Elaborado pelo autor (2021).

4.2 Modelo baseado em dicionário

O módulo inicial da abordagem proposta, chamado modelo baseado em dicionário, utiliza o método SVM Linear (FAN et al., 2008) para construir um dicionário de palavras que possuem pesos associados e, em seguida, verifica a incidência das palavras desse dicionário no texto cujo gênero do seu autor é desconhecido para realizar a predição de gênero.

Os pesos associados às palavras representam a intensidade da relação das mesmas com o gênero do autor do texto, sendo os pesos negativos indicadores de palavras predominantemente vinculadas a textos escritos por pessoas do gênero masculino e os pesos positivos indicadores de palavras predominantemente relacionadas com textos escritos por pessoas do gênero feminino.

A definição do gênero do autor de um texto a partir desse modelo baseado em dicionário ocorre a partir do cálculo de um índice denominado g_{lex}. Na abordagem proposta em (SAP et al., 2014) o critério

mostrado na Equação 4.1 é utilizado para a definição do gênero do autor de um texto.

$$\text{gênero} = \begin{cases} feminino, & \text{se } g_{lex} >= 0 \\ masculino, & \text{caso contrário.} \end{cases} \tag{4.1}$$

A abordagem aqui proposta altera esse critério adotando valores limiares diferentes para o g_{lex}, sendo α_{masc} o limiar para o gênero masculino e α_{fem} o limiar para o gênero feminino. Assim como mostra a Equação 4.2, agora caso o valor de g_{lex} seja positivo e maior ou igual a α_{fem}, o gênero retornado será o feminino. Por outro lado, se o valor de g_{lex} for negativo e menor ou igual a α_{masc}, o gênero retornado será o masculino. No entanto, se o valor de g_{lex} não se enquadrar em um desses dois casos, nenhum gênero será retornado pelo modelo baseado em dicionário.

$$\text{gênero} = \begin{cases} feminino, & \text{se } g_{lex} >= \alpha_{fem} \quad (\alpha_{fem} > 0) \\ masculino, & \text{se } g_{lex} <= \alpha_{masc} \quad (\alpha_{masc} < 0) \\ \varnothing, & \text{para os demais casos.} \end{cases} \tag{4.2}$$

A Equação 4.3 mostra a forma utilizada para se calcular o índice g_{lex} para um determinado texto. Observe que o cálculo desse índice leva em consideração a frequência relativa das palavras do dicionário no texto em análise e o peso associado a cada uma dessas palavras.

$$g_{lex} = \left(\sum_{p \in lex} w_{lex}(p) * \frac{freq(p, doc)}{freq(*, doc)} \right) + w_0, \tag{4.3}$$

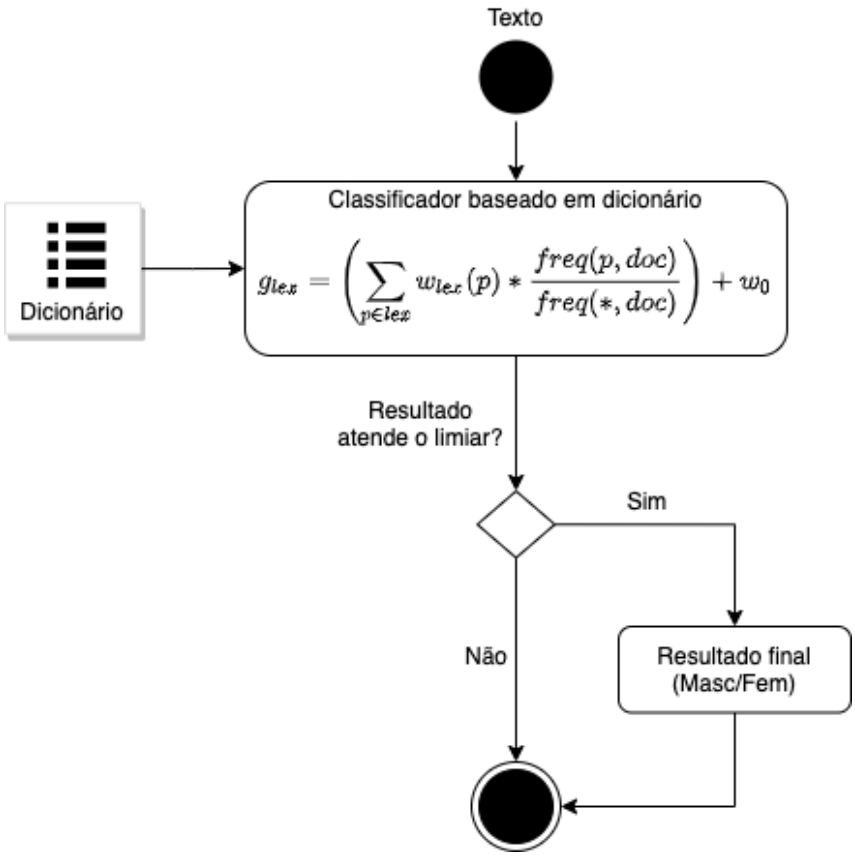
onde $w_{lex}(p)$ representa o peso associado a uma palavra p presente no dicionário (lex), $freq(p, doc)$ a frequência da palavra p no texto, $freq(*, doc)$ o total de palavras do texto e w_0 é uma constante denominada *intercept*.

A Figura 4.2 apresenta a visão geral do processamento de um texto realizado pelo módulo baseado em dicionário.

Os detalhes envolvidos na criação do dicionário bem como no cálculo dos limiares utilizados na Equação 4.2 são apresentados a seguir.

Criação do dicionário: A primeira etapa deste módulo é a criação de um dicionário com palavras que caracterizem os gêneros masculino e feminino. Assim como na abordagem proposta em Sap et al. (2014), um classificador SVM com *kernel* linear é treinado utilizando como atributos preditivos a frequência relativa das palavras contidas no conjunto de textos de treinamento e como atributo classe o gênero do autor do texto. A partir do treinamento do classificador SVM Linear, obtém-se o peso de cada um dos atributos, ou seja, das palavras existentes nos textos utilizados como entrada para o classificador. Esses pesos nada mais são do que os coeficientes que definem (juntamente com a constante w_0) o

Figura 4.2 – Visão geral do módulo baseado em dicionário



Fonte: Elaborado pelo autor (2021).

hiperplano de separação (ver Seção 2.3) obtido pelo classificador SVM num espaço t-dimensional (onde *t* é a quantidade de atributos). A lógica é que quanto maior é o valor (em módulo) do peso (coeficiente) associado a uma determinada palavra (atributo), mais relevante é essa palavra (atributo) para a definição do gênero do autor do texto (atributo classe).

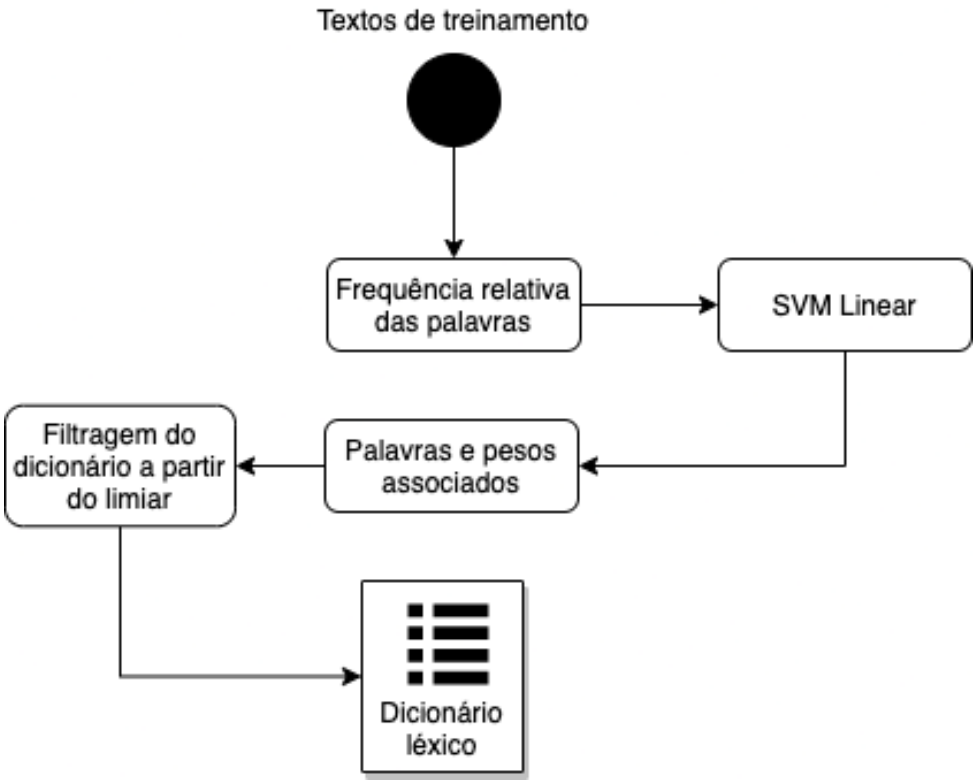
No dicionário construído a partir da estratégia aqui apresentada, as palavras associadas aos pesos negativos indicam a tendência do autor do texto ser do gênero masculino, enquanto que as palavras associadas aos pesos positivos (ou nulos) indicam a tendência do autor do texto ser do gênero feminino. Vale ressaltar que, quanto maior o valor desses pesos em módulo, mais forte é essa tendência com relação a um dos gêneros.

É importante mencionar que, diferentemente da estratégia apresentada em Sap et al. (2014), a abordagem deste trabalho adotou um limiar em relação ao peso (definido empiricamente) para definir quais palavras podem compor o dicionário. O objetivo desse limiar é selecionar somente as palavras que possuem maior poder de discriminação para a predição do gênero, eliminando, portanto, palavras que possuem pesos mais próximos do valor zero e, desse modo, são menos importantes no processo de definição do gênero do autor de um texto. Com essa estratégia, apesar de construirmos dicionários potencialmente menores do que os da proposta original (SAP et al., 2014) e, com isso, existir a possibi-

lidade de não conseguirmos classificar um determinado texto (quando o mesmo não contiver pelo menos uma palavra do dicionário), o objetivo é alcançar um alto desempenho preditivo para aqueles textos que puderem ser classificados a partir da estratégia aqui proposta.

A Figura 4.3 ilustra esse processo de construção do dicionário. A partir da extração das frequências relativas das palavras dos textos do conjunto de treinamento, um classificador SVM Linear é treinado. Após isso, as palavras com seus respectivos pesos são filtradas a partir de um valor limiar para eliminação das palavras que menos contribuem para a discriminação do gênero.

Figura 4.3 – Construção do dicionário



Fonte: Elaborado pelo autor (2021).

Assim como mencionado anteriormente, após a construção do dicionário, calcula-se o índice g_{lex} para um texto que se deseja predizer o gênero de seu autor. Se esse índice respeitar determinados valores limite, então o modelo baseado em dicionário retornará o gênero obtido como resultado. Portanto, a seguir é apresentada a estratégia utilizada para o cálculo desses limiares definidos para o índice g_{lex} (ver Equação 4.2).

Definição do limiares para o índice g_{lex}: A justificativa para a adoção desses limiares é semelhante àquela apresentada para a filtragem de palavras que compõem o dicionário. Ou seja, o objetivo é maximizar o desempenho preditivo do modelo baseado em dicionário, ainda que para isso tenhamos uma diminuição no percentual de instâncias (textos) para as quais esse módulo consegue realizar a predição de gênero. Para maximizar o desempenho, a ideia é retornar a predição de gênero somente para textos

cujo índice g_{lex} corresponda a um valor absoluto maior do que a média absoluta obtida para textos de um determinado gênero.

Portanto, na estratégia aqui proposta, dois limiares são definidos: α_{fem} (limiar para o gênero feminino) e α_{masc} (limiar para o gênero masculino). Esses limiares são definidos a partir do índice g_{lex} calculado para as instâncias de treinamento (as mesmas que foram utilizadas na construção do dicionário). Mais especificamente, α_{fem} corresponde ao valor médio de g_{lex} das instâncias de treinamento (cujo autor é do gênero feminino) que são corretamente classificadas pelo modelo baseado em dicionário considerando-se o critério apresentado na Equação 4.1. Da mesma forma o cálculo é realizado para obtenção do limiar α_{masc}, sendo que, neste caso, obviamente, apenas as instâncias de treinamento associadas ao gênero masculino são consideradas no cálculo.

Com a utilização desses limiares, aqueles textos que possuem palavras do dicionário associadas tanto ao gênero masculino quanto ao gênero feminino ou com poucas palavras associadas a somente um gênero e, por causa disso, atingem um índice g_{lex} tal que $\alpha_{masc} < g_{lex} < \alpha_{fem}$, acabam não sendo classificados pelo modelo baseado em dicionário.

Caso o gênero do autor do texto não consiga ser definido pelo módulo baseado em dicionário, o processamento do texto é efetuado pelo segundo módulo, chamado heurística de gênero. A heurística de gênero tem como objetivo analisar o texto morfossintaticamente a fim de encontrar expressões que indiquem o gênero do autor do texto. A seção a seguir apresenta em detalhes essa heurística.

4.3 Heurística de Gênero

Segundo Bechara (2009), a língua portuguesa é considerada morfologicamente rica, sendo a flexão de gênero muito presente em adjetivos, substantivos, pronomes etc. Desse modo, a heurística proposta nesse trabalho tem como objetivo explorar essa riqueza morfológica a fim de encontrar expressões que possam indicar o gênero do autor de um texto. O Algoritmo 1 mostra uma visão geral do funcionamento da heurística. Cada uma das etapas serão detalhadas a seguir.

As linhas 4 e 5 correspondem às primeiras etapas do processamento da heurística. A primeira etapa consiste na tokenização sentencial de um dado texto. Após isso, caso a frase seja iniciada com aspas, ela é ignorada. Essa validação tem como objetivo evitar a análise de frases que indiquem citações a terceiros. Em seguida, a heurística processa cada frase do texto (linhas 6 a 20) com objetivo de capturar expressões que indiquem o gênero do autor do texto. Os detalhes desse processamento são apresentados a seguir.

Inicialmente realiza-se a construção de uma estrutura em árvore que representa as relações sintáticas entre as palavras de uma frase. Além disso, uma análise morfológica das palavras que compõem

Figura 4.4 – Heurística de gênero

Algorithm 1 Heurística de gênero

```
1:  function GENERO(texto)
2:      f ← 0
3:      m ← 0
4:      frases ← separar_frases(texto)
5:      frases ← remover_citações(frases)
6:      for frase in frases do
7:          analise_morfossintatica ← cria_analise_morfossintatica(frase)
8:          for palavra in analise_morfossintatica do
9:              if palavra is verbo_ligacao_singular then
10:                 genero ← analisa_ligacoes(palavra)
11:                 if genero = masculino then
12:                     m += 1
13:                 else
14:                     if genero = feminino then
15:                         f += 1
16:                     end if
17:                 end if
18:             end if
19:         end for
20:     end for
21:     genero_final ← ponderacao_limiar(m,f)
22:     return genero_final
23: end function
```

Fonte: Elaborado pelo autor (2021).

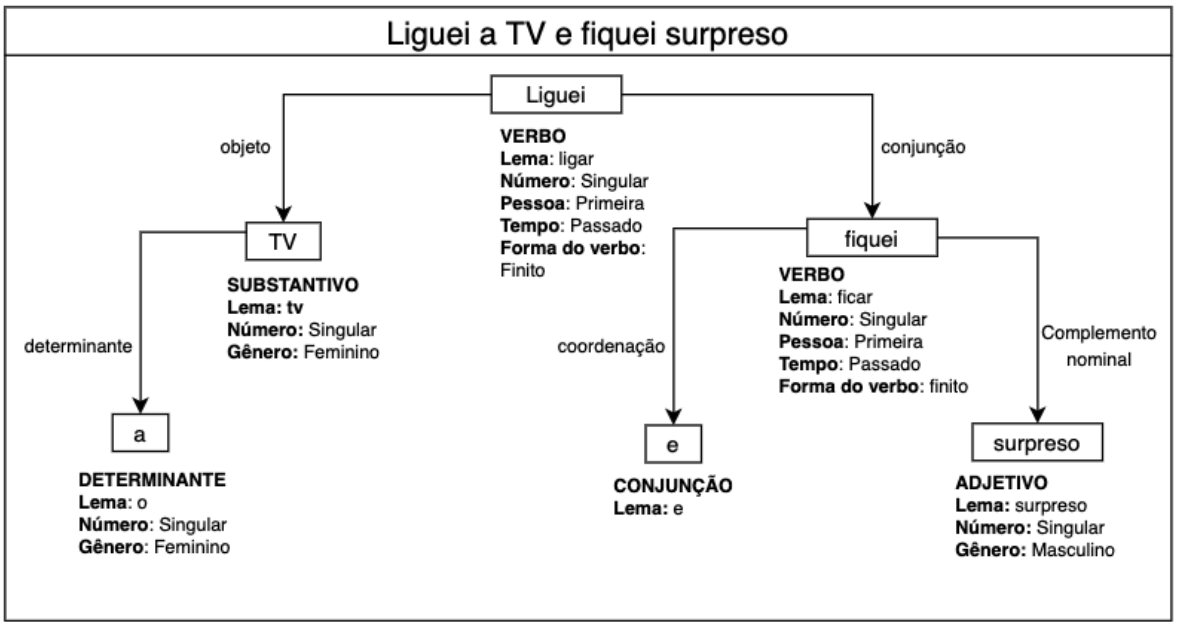
a frase nos permite capturar informações como a classe gramatical e o gênero das palavras (linha 7). A Figura 4.5 ilustra esse processo.

Segundo Faraco e Moura (2010), verbos de ligação como ser, estar, permanecer etc. têm como principal função estabelecer relações entre termos da frase, como a relação entre o sujeito da frase e um termo que expressa características do mesmo. A partir disso, a proposta da heurística de gênero é buscar verbos de ligação nas frases que estejam conjugados na primeira pessoa do singular (linha 9), possibilitando, dessa forma, a definição do gênero do autor do texto. Os seguintes verbos de ligação foram considerados na heurística proposta: ser, estar, permanecer, ficar, parecer, andar (no sentido de "encontrar-se"), viver (no sentido de "estar sempre"), virar (no sentido de "tornar-se") e continuar.

Uma vez encontrado um verbo de ligação em uma frase, visando a obtenção de palavras que possam revelar o gênero do autor do texto, a terceira etapa do processo (linha 10) consiste em analisar as palavras associadas tanto aos nós filhos como ao nó pai daquele que representa esse verbo de ligação na árvore de dependência sintática obtida na etapa anterior. No entanto, dentre as palavras analisadas, somente serão consideradas aquelas que atenderem as seguintes condições:

1. A palavra encontra-se após o verbo de ligação na frase.

Figura 4.5 – Exemplo de análise morfossintática



Fonte: Elaborado pelo autor (2021).

2. A palavra é um adjetivo ou um verbo.
3. Se a palavra for um adjetivo, ela deve estar no singular. Essa condição garante que a expressão em análise esteja na primeira pessoa do singular. Essas restrições fazem com que a heurística consiga identificar expressões como “estou bonito” e “sou bondosa”.
4. Se a palavra for um verbo, ela também deve estar no singular. Além disso, para analisar somente expressões que estejam na voz passiva, ou seja, quando o sujeito recebe a ação do verbo, são analisados apenas verbos no particípio passado (BECHARA, 2009). Com essas restrições a heurística é capaz de capturar expressões como “eu fui reprovado” e “eu fiquei ofendido”.

Por fim, a quarta e última etapa da heurística (linhas 11 a 16) é realizada caso alguma palavra seja identificada na etapa anterior. Por exemplo, na frase utilizada na Figura 4.5, o adjetivo “surpreso” define o gênero do autor do texto como masculino. No entanto, mais de uma palavra pode ser obtida como resultado da terceira etapa. Isso pode ocorrer, por exemplo, quando um texto é formado por várias frases. Nesse caso, o processamento para a definição de gênero é realizado da maneira descrita a seguir.

A linha 21 do Algoritmo 1 representa a última etapa da heurística, para o caso de termos múltiplas palavras para a definição do gênero do autor do texto. Nesses casos a heurística adota um limiar $c \in [0,5;1)$ (parâmetro definido pelo usuário) para definir o resultado final. A ideia é que esse limiar represente o grau de confiança da heurística com relação ao gênero definido pela mesma. Considere que na terceira etapa foram identificadas f palavras do gênero *Feminino* e m palavras do gênero *Masculino*. Nesse caso, a heurística apresentará como resultado o gênero do autor do texto de acordo com as condições definidas na Equação 4.4.

$$\text{Gênero} = \begin{cases} \textit{Feminino} & \text{se } (f > m) \text{ e } (\frac{f}{f+m} > c) \\ \textit{Masculino} & \text{se } (m > f) \text{ e } (\frac{m}{m+f} > c) \end{cases} \tag{4.4}$$

Vale observar que a heurística aqui proposta não será capaz de atribuir um gênero para o autor de um texto quando o mesmo não possuir frases com as características capturadas pela mesma ou quando a quarta etapa não atingir o limiar definido pelo usuário.

Caso a heurística de gênero não consiga definir o gênero do autor, os textos são processados pelo último módulo, que consiste em um modelo de classificação, que será discutido a seguir.

4.4 Modelo de Classificação

O último módulo da abordagem proposta é o chamado Modelo de Classificação. Esse módulo consiste em uma abordagem convencional para predição de gênero, sendo a primeira etapa de pré-processamento dos textos (conforme visto na Seção 2.1), seguido por uma etapa de representação dos textos para modelos computacionais (ver Seção 2.2) e treinamento dos classificadores com o textos da base de treinamento. Dessa forma, textos com autor de gênero desconhecido podem ser submetidos e classificados pelo modelo de classificação.

5 EXPERIMENTOS COMPUTACIONAIS

Um conjunto de experimentos computacionais foi realizado com o objetivo de avaliar a abordagem proposta neste trabalho. Vale lembrar que a hipótese é que a abordagem proposta é capaz de melhorar o desempenho preditivo da tarefa de classificação de gênero alcançado por trabalhos estado da arte apresentados na literatura para diferentes bases de dados compostas por textos escritos na língua portuguesa. Essa hipótese é fundamentada na ideia de que explorar as especificidades da língua portuguesa e características específicas do domínio do texto pode contribuir positivamente na tarefa de predição de gênero.

Os detalhes dos experimentos realizados neste trabalho serão apresentados da forma descrita a seguir. A Seção 5.1 apresenta uma descrição das bases de dados utilizadas. Em seguida, a configuração experimental é apresentada na Seção 5.2. Por fim, a Seção 5.3 mostra e discute os resultados obtidos.

5.1 Bases de Dados

Para realização dos experimentos foram selecionadas seis bases de dados textuais já utilizadas em outros trabalhos da literatura que apresentam características diversas no que diz respeito ao conteúdo, à origem dos dados (sites, *blogs*, redes sociais etc.) e tamanho dos textos. A seguir tem-se uma breve descrição de cada uma delas.

b5-corpus: base composta por 1019 textos escritos na língua portuguesa do Brasil provenientes de postagens no Facebook realizadas por 1019 usuários distintos. Cada texto corresponde à junção de até 1000 postagens (*Facebook status*) de cada usuário. Além disso, essa base de dados contém informações demográficas de cada usuário, sendo o gênero uma delas. Essa base de dados é parte de um corpus criado por Ramos et al. (2018) e os seus textos têm um teor informal relacionado a assuntos diversos.

BlogSetBR: base formada por 2602 textos de *blogs* escritos na língua portuguesa do Brasil e publicados por usuários distintos. O texto de cada usuário corresponde a uma ou mais publicações do mesmo. Essa base foi elaborada por Santos, Woloszyn e Vieira (2018) a partir de mais de 7 milhões de textos coletados da plataforma Blogspot[1]. Para cada texto tem-se a informação do gênero do seu autor. Os textos estão relacionados a temas variados, indo desde cuidados pessoais até política internacional.

PAN-17: base composta por textos de *tweets* de 2000 autores distintos, sendo que cada texto corresponde a uma agregação de pelo menos 100 *tweets* de cada autor. Desses textos que tratam de assuntos diversos, metade foi escrito em português brasileiro e metade em português europeu. Essa base,

[1] https://developers.google.com/blogger/

rotulada com o gênero de cada autor, foi utilizada na tradicional competição PAN-CLEF (RANGEL et al., 2017) promovida em 2017.

BRmoral: base formada a partir da agregação de 3400 textos opinativos curtos escritos na língua portuguesa brasileira gerados por 433 autores distintos. Esses textos versam sobre opiniões produzidas como respostas para questões sobre temas como legalização de drogas, pena de morte, aborto e outros. O gênero dos autores desses textos está disponível nessa base de dados disponibilizada em Santos e Paraboni (2019).

B2W-Reviews01: base composta por 126244 textos correspondentes a avaliações de produtos realizadas por consumidores por meio do site de uma grande empresa varejista do Brasil. Essa base de dados, disponibilizada em Real, Oshiro e Mafra (2019), contém avaliações de mais de 40 mil produtos que foram coletadas entre janeiro e maio de 2018. Vale ressaltar que os textos são caracterizados pelo uso de uma linguagem informal e possuem tamanhos muito variados.

E-SIC: base contendo 44698 textos obtidos a partir de uma coleção de solicitações feitas por cidadãos por meio do e-SIC (Sistema Eletrônico do Serviço de Informações ao Cidadão) disponibilizado pelo governo brasileiro. Esses textos são tipicamente impessoais e abordam tópicos relacionados a organizações, impostos, políticas públicas e outros.

A Tabela 5.1 apresenta as principais características das bases de dados utilizadas neste trabalho. A coluna 'Domínio' especifica o tipo de texto de cada base. Em seguida, as colunas 'Masculino' e 'Feminino' mostram a quantidade de instâncias (textos) disponível na base para cada um dos gêneros. Por fim, a coluna 'Palavras/Texto' mostra a quantidade média de palavras por texto.

Tabela 5.1 – Características da bases de dados

Base de Dados	Domínio	Masculino	Feminino	Palavras / Texto
b5-corpus	Facebook	578	441	2.389,03
BlogSetBR	Blogs	1038	1564	5.801,75
PAN-17	Twitter	1000	1000	1.076,52
BRmoral	Opinião	285	148	432,70
B2W-Reviews01	*Reviews*	65129	61115	23,81
E-SIC	E-Gov	28805	15893	76,29

Fonte: Elaborado pelo autor (2021).

5.2 Configuração Experimental

Todos os módulos que compõem a abordagem proposta nesse trabalho foram implementados utilizando a linguagem Python. Além disso, várias especificidades foram adotadas na implementação de cada desses módulos. A seguir serão apresentadas as configurações adotadas para cada um dos módulos.

Modelo baseado em dicionário: A implementação do modelo baseado em dicionário utiliza a biblioteca *Scikit Learn* (PEDREGOSA et al., 2011) como principal ferramenta. O único pré-processamento do texto utilizado neste módulo foi a tokenização lexical, feito com a ferramenta NLTK – *Natural Language Toolkit* (BIRD; KLEIN; LOPER, 2009). Os atributos utilizados para treinamento do classificador utilizado neste módulo, assim como em Sap et al. (2014), foram as frequências relativas das palavras que aparecem em pelo menos 1% dos textos. Além disso, é importante mencionar que na configuração do classificador SVM com *kernel* linear, o parâmetro C foi definido utilizando a estratégia de calibração de parâmetros denominada *Grid Search*. A calibração de parâmetros também foi responsável pela definição do valor limiar para a filtragem do dicionário (ver Seção 4.2). A Tabela 5.2 mostra, para cada uma das bases de dados, os parâmetros definidos a partir do melhor resultado de F1 obtido pelo *Grid-Search* utilizando a validação cruzada com 10 partições ($k = 10$). Após a coluna que apresenta o valor do parâmetro C do classificador SVM Linear, a coluna "Limiar" contém valores entre 0 e 1 que representam o percentual de palavras que são mantidas no dicionário. Desse modo, um valor igual a 0,10, por exemplo, significa que somente 10% das palavras originalmente contidas no dicionário construído são mantidas após a etapa de filtragem. Obviamente, sempre são mantidas as palavras com maior poder de discriminação do gênero, ou seja, somente aquelas com maior peso em valor absoluto.

Tabela 5.2 – Parâmetros C e Limiar para cada uma das bases de dados

Base de dados	**C**	**Limiar**
b5-corpus	1,0	0,10
PAN-17	1,0	0,40
B2W-Reviews01	1,0	0,01
Blogset-BR	5,0	0,40
BRMoral	5,0	0,40
E-SIC	1,0	0,01

Fonte: Elaborado pelo autor (2021).

Heurística de gênero: No caso da heurística de gênero, o único pré-processamento realizado no texto foi a tokenização (utilizando-se a biblioteca NLTK – *Natural Language Toolkit*) para a separação do mesmo em sentenças. Para definição do valor do parâmetro c da heurística, experimentos foram realizados com $c = \{0,6; 0,75; 0,85\}$. Os resultados reportados neste trabalho foram obtidos com o valor de parâmetro $c = 0,75$, que foi o que propiciou o melhor desempenho médio da heurística. Por fim, para capturar as relações sintáticas entre as palavras e realizar a análise morfológica das mesmas utilizou-se a ferramenta Stanza[2].

Modelo de classificação de gênero: Uma vez que o objetivo é comparar a abordagem proposta com outras já apresentadas na literatura, para cada base de dados utilizada, a configuração experimental

[2] https://stanfordnlp.github.io/stanza/index.html

(técnicas e algoritmos com seus respectivos parâmetros) foi exatamente a mesma adotada nos trabalhos utilizados como referência, cujos detalhes foram apresentados no Capítulo 3.

A Tabela 5.3 mostra, para cada base de dados, qual trabalho da literatura foi utilizado como referência (coluna 'Referência') e as principais características da configuração experimental utilizada no mesmo, a saber: a forma utilizada para representar o texto em um vetor numérico (coluna 'Vetorização do Texto'), método utilizado na classificação de gênero (coluna 'Classificador') e a técnica utilizada na avaliação dos classificadores (coluna 'Técnica de Avaliação'). Vale observar que apenas para a base *B2W-Reviews01* não foram encontrados na literatura trabalhos que realizassem a tarefa de predição de gênero a partir da mesma. Desse modo, a mesma configuração experimental utilizada para a base *b5-corpus* foi escolhida para o processamento dessa base.

Tabela 5.3 – Resumo da configuração experimental

Base de Dados	Referência	Vetorização do Texto	Classificador	Técnica de Avaliação
b5-corpus	(HSIEH; DIAS; PARABONI, 2018)	*Bag of Words* com TF-IDF	Regressão Logística	Validação Cruzada (10-*fold*)
PAN-17	(BASILE et al., 2017)	*Bag of Words* com TF-IDF	SVM	Validação Cruzada (5-*fold*)
B2W-Reviews01	–	*Bag of Words* com TF-IDF	Regressão Logística	Validação Cruzada (10-*fold*)
BlogSet-BR	(DIAS; PARABONI, 2020)	Word2Vec	Multilayer Perceptron	Holdout 80%/20%
BRmoral	(DIAS; PARABONI, 2020)	Word2Vec	Multilayer Perceptron	Holdout 80%/20%
E-SIC	(DIAS; PARABONI, 2020)	Word2Vec	Multilayer Perceptron	Holdout 80%/20%

Fonte: Elaborado pelo autor (2021).

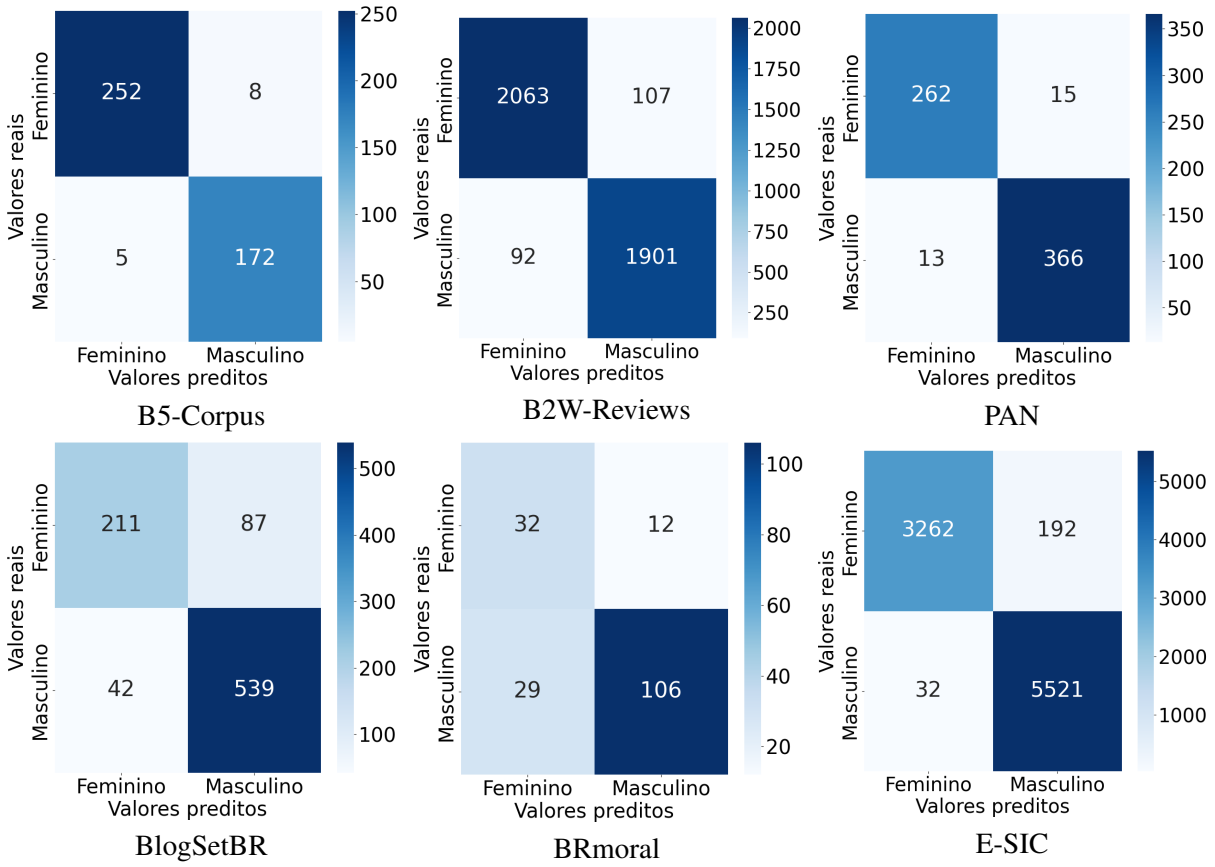
5.3 Resultados e Discussões

Nesta seção, inicialmente serão apresentados os desempenhos preditivos obtidos individualmente por cada um dos dois primeiros módulos da abordagem proposta, a saber, do modelo baseado em dicionário e da heurística de gênero. Em seguida, serão mostrados e discutidos os resultados da análise comparativa realizada entre os trabalhos da literatura utilizados como referência e a abordagem em cascata aqui proposta.

As Figuras 5.1 e 5.2 mostram as matrizes de confusão do modelo baseado em dicionário e da heurística de gênero, respectivamente. Em seguida, a Tabela 5.4 mostra a precisão, revocação e medida F1 de ambos os módulos da abordagem proposta para cada uma das bases de dados. Além disso, as colunas 'Cobertura do Módulo' mostram o percentual das instâncias de teste que cada um dos módulos consegue classificar.

Como pode ser observado nos resultados apresentados na Tabela 5.4, para a maioria das bases de dados avaliadas, tanto o modelo baseado em dicionário quanto a heurística de gênero alcançaram um valor de F1 superior a 0,9, que corresponde a um desempenho satisfatório se tomarmos com referência os resultados obtidos por trabalhos da literatura para essas mesmas bases de dados. É importante observar que não podemos fazer uma comparação direta com os resultados da literatura, uma vez que o cálculo do F1 apresentado na Tabela 5.4 levou em consideração apenas as instâncias capturadas pelo modelo

Figura 5.1 – Matrizes de confusão do modelo baseado em dicionário para cada uma das bases de dados



Fonte: Elaborado pelo autor (2021).

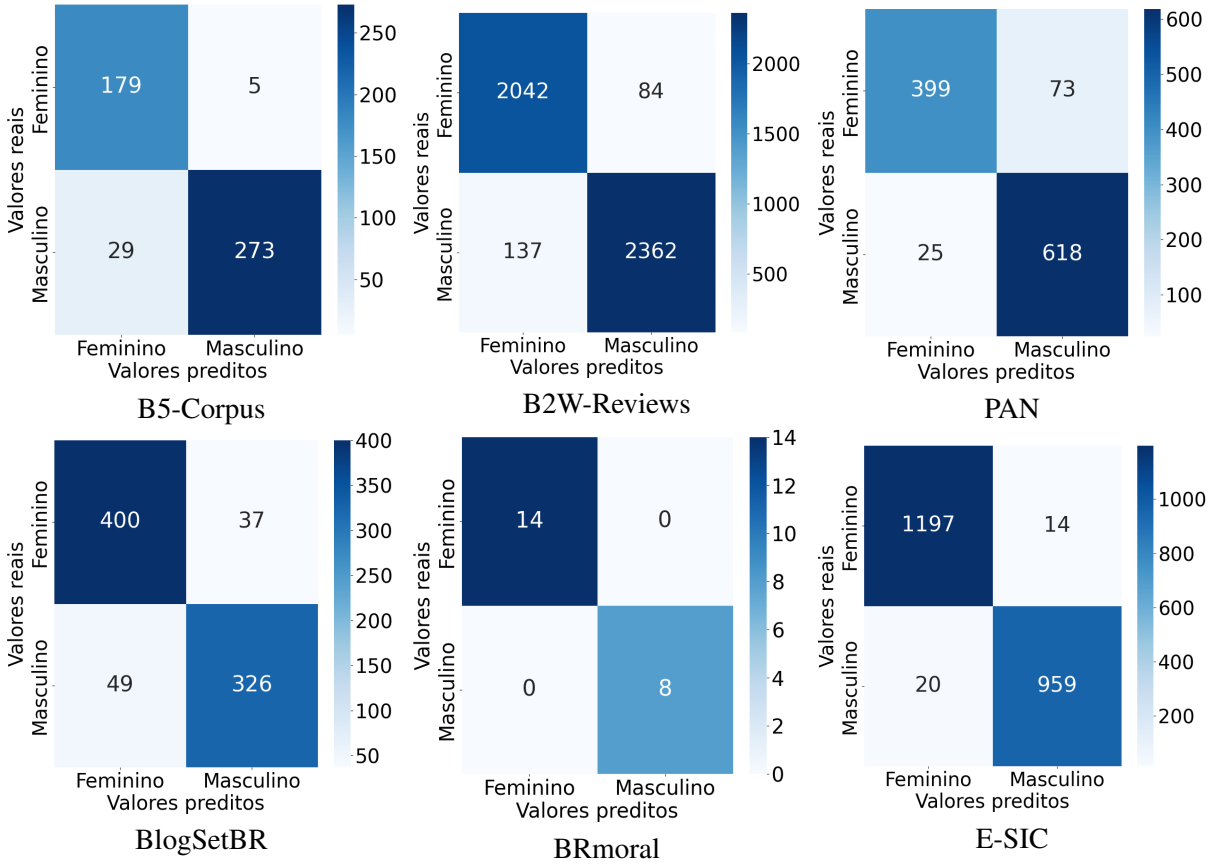
baseado em dicionário e pela heurística de gênero. Vale ressaltar que mesmo para as duas bases de dados em que isso não ocorreu (BlogSetBR e BRMoral), o desempenho alcançado por esses dois módulos também foi satisfatório se tomarmos como referência os resultados obtidos por trabalhos da literatura para essas bases. Desse modo, esses resultados indicam o potencial que esses dois primeiros módulos da abordagem em cascata possuem para melhorar o desempenho preditivo da tarefa de predição de gênero.

A seguir serão apresentados os resultados obtidos a partir da análise comparativa entre os trabalhos da literatura utilizados como referência e aqueles alcançados com a abordagem em cascata proposta neste trabalho.

Para garantir uma comparação justa entre a abordagem aqui proposta e aquelas apresentadas na literatura que foram utilizadas como referência neste trabalho, ao invés de simplesmente utilizarmos os resultados reportados nos trabalhos de referência, executamos os mesmos experimentos descritos nesses trabalhos para utilizarmos como *baseline* da nossa avaliação comparativa. Desse modo, as mesmas partições de dados utilizadas no treinamento e teste dos classificadores foram empregadas para comparar as abordagens.

Uma vez que o protocolo experimental foi exatamente o mesmo dos trabalhos de referência, as pequenas diferenças encontradas entre os resultados reportados na literatura e os que obtivemos após a

Figura 5.2 – Matrizes de confusão da heurística de gênero para cada uma das bases de dados



Fonte: Elaborado pelo autor (2021).

Tabela 5.4 – Precisão, Revocação e F1 das respectivas abordagens

	Modelo baseado em dicionário				**Heurística de gênero**			
Base de dados	Precisão	Revocação	F1	Cobertura do Módulo	Precisão	Revocação	F1	Cobertura do Módulo
b5-corpus	0,970	0,970	0,970	42,84%	0,936	0,930	0,933	47,64%
PAN-17	0,957	0,957	0,957	32,80%	0,914	0,912	0,913	55,75%
B2W-Reviews	0,952	0,953	0,953	3,30%	0,953	0,952	0,952	3,66%
BlogSetBR	0,852	0,853	0,853	33,76%	0,894	0,894	0,894	31,18%
BRmoral	0,806	0,771	0,788	35,10%	1,00	1,00	1,00	4,31%
E-SIC	0,976	0,975	0,975	20,87%	0,984	0,984	0,984	5,07%

Fonte: Elaborado pelo autor (2021).

execução dos experimentos se devem à aleatoriedade existente no processo de subdivisão das bases para geração dos conjuntos de treinamento e teste dos classificadores.

A Tabela 5.5 apresenta os resultados obtidos a partir dos experimentos realizados para comparar os resultados da literatura com aqueles obtidos a partir da abordagem em cascata. Nessa tabela, os resultados reportados pelos artigos de referência são apresentados na coluna 'Referência'. Em seguida, a coluna 'Baseline' apresenta os resultados obtidos a partir dos experimentos que realizamos seguindo a abordagem e o protocolo experimental de cada trabalho de referência. Os resultados obtidos a partir

da abordagem em cascata proposta neste trabalho são mostrados na coluna 'Abordagem em Cascata'. A coluna 'Métrica' apresenta a métrica que foi utilizada na avaliação comparativa. Vale observar que utilizamos a mesma métrica de avaliação adotada em cada trabalho de referência.

Tabela 5.5 – Resultados dos experimentos

Base de Dados	Referência	*Baseline*	Abordagem em Cascata	Métrica
b5-corpus	88,0%	88,5%	• **91,0%**	F1
PAN-17	84,5%	84,9%	• **89,0%**	Acurácia
B2W-Reviews	–	68,3%	**68,6%**	F1
BlogSetBR	78,0%	77,2%	**80,6%**	F1
BRmoral	74,0%	74,5%	**75,4%**	F1
E-SIC	79,0%	78,0%	**80,5%**	F1

Fonte: Elaborado pelo autor (2021).

Antes de analisarmos os resultados obtidos, é importante destacar que a Tabela 5.5 foi horizontalmente dividida em duas partes devido ao fato de as três primeiras bases terem sido avaliadas utilizando-se a técnica de validação cruzada e, as três últimas, a técnica *houldout*. Desse modo, os resultados apresentados para as três primeiras bases correspondem a valores médios das k partições de teste e, por isso, a comparação entre esses resultados foi realizada utilizando-se o teste estatístico de Wilcoxon. Já as três últimas bases são avaliadas de acordo o resultado obtido a partir de uma única partição de teste e, portanto, sem a aplicação de um teste de significância estatística.

Na Tabela 5.5, os resultados destacados em negrito correspondem ao maior valor de desempenho preditivo para cada base de dados. Além disso, para o caso das bases avaliadas segundo o teste estatístico de Wilcoxon (com nível de significância $\alpha = 0,05$), o símbolo • indica que existe diferença com significância estatística entre o resultado alcançado pela abordagem proposta e o do *baseline*.

Os resultados mostram que a abordagem proposta alcança desempenho preditivo sempre melhor ao das abordagens dos trabalhos de referência (*baseline*). Mais especificamente, para o caso das bases avaliadas utilizando-se o teste estatístico, para duas das três bases o desempenho da abordagem proposta foi significativamente superior ao do *baseline*. Para as bases que foram avaliadas segundo a técnica *houldout*, a abordagem proposta obteve desempenho sempre superior ao do *baseline*.

Outro ponto importante a ser observado nesses resultados é que o ganho de desempenho com relação ao *baseline* foi mais expressivo quando a cobertura dos dois primeiros módulos da abordagem em cascata foi maior, indicando o potencial de contribuição dos mesmos na melhoria do desempenho da tarefa de classificação de gênero.

6 CONCLUSÃO

O crescente número de pessoas que utilizam a internet faz com que a quantidade de dados *online* disponível seja cada vez maior. Principalmente devido às redes sociais e aos diversos tipos de serviços *online* existentes, os textos representam grande parte dos dados atualmente disponíveis na internet. No entanto, como na maioria dos casos os textos podem ser publicados de forma anônima, o uso de técnicas computacionais para inferir as características dos seus autores é objeto de estudo da área de pesquisa denominada Caracterização Autoral.

Apesar do crescente interesse por essa área de pesquisa, a quantidade de trabalhos na literatura e de recursos e ferramentas computacionais disponíveis para a língua portuguesa ainda é relativamente pequena quando comparada àquela disponível para outros idiomas.

Desse modo, este trabalho contribui nessa área de estudo apresentando uma abordagem em cascata para predição de gênero de autores de textos escritos na língua portuguesa que leva em consideração tanto especificidades da língua como características de domínio dos textos.

Os resultados obtidos a partir da abordagem proposta mostraram que explorar as especificidades da língua portuguesa e características de domínio dos textos pode contribuir positivamente no desempenho da tarefa de predição de gênero. Essa conclusão foi obtida por meio de experimentos computacionais realizados a partir de bases de dados textuais de domínios diversos já utilizadas por outros trabalhos apresentados na literatura para a tarefa de predição de gênero de autores de textos. Nesses experimentos, a abordagem proposta foi sempre superior àquelas apresentadas na literatura.

Como trabalho futuro sugere-se avaliar configurações alternativas dos módulos aqui utilizados com objetivo de alcançar desempenhos preditivos ainda melhores.

REFERÊNCIAS

ARCIA, Y. A.; CASTRO-CASTRO, D.; BUENO, R. O.; MUÑOZ, R. Author profiling, instance-based similarity classification. In: **Notebook for PAN at CLEF**. Dublin, Ireland: CLEF, 2017. p. 40–48.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação - Conceitos e tecnologia das máquinas de busca.** 2. ed. Porto Alegre, RS: Bookman, 2013.

BARROS, C. B. **Classificadores de regressão logística, Naive Bayes e Random Forest na análise do Tropismo do HIV-1 de subtipo B**. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2019.

BASILE, A.; DWYER, G.; MEDVEDEVA, M.; RAWEE, J.; HAAGSMA, H.; NISSIM, M. N-gram: New groningen author-profiling model. **arXiv preprint arXiv:1707.03764**, 2017.

BECHARA, E. **Moderna Gramática Portuguesa**. Rio de Janeiro, RJ: Editora Nova Fronteira, 2009.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit**. [S.l.]: "O'Reilly Media, Inc.", 2009.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: . NY, USA: Association for Computing Machinery, 1992. p. 144–152.

CIOBANU, A. M.; ZAMPIERI, M.; MALMASI, S.; DINU, L. P. Including dialects and language varieties in author profiling. **arXiv preprint arXiv:1707.00621**, 2017.

CRUZ, R. M.; SABOURIN, R.; CAVALCANTI, G. D.; Ing Ren, T. Meta-des: A dynamic ensemble selection framework using meta-learning. **Pattern Recognition**, p. 1925–1935, 2015.

DIAS, R.; PARABONI, I. Cross-domain author gender classification in brazilian portuguese. In: **Proceedings of The 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 1227–1234.

EISENSTEIN, J. **Introduction to Natural Language Processing**. Cambridge, MA: The MIT Press, 2019.

FAN, R.-E.; CHANG, K.-W.; HSIEH, C.-J.; WANG, X.-R.; LIN, C.-J. Liblinear: A library for large linear classification. **the Journal of machine Learning research**, JMLR, p. 1871–1874, 2008.

FARACO, F. C. E.; MOURA, F. M. **Gramática**. São Paulo,SP: Editora Ática, 2010.

FERREIRA, T. G. **Nicesim: um simulador interativo baseado em técnicas de aprendizado de máquina para avaliação de recém-nascidos prematuros em uti neonatal**. Dissertação (Mestrado) — Universidade Federal de Viçosa, 2014.

GOOT, R. van der; LJUBEŠIĆ, N.; MATROOS, I.; NISSIM, M.; PLANK, B. Bleaching text: Abstract features for cross-lingual gender prediction. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 383–389.

GUO, Y.; LIU, J.; TANG, W.; HUANG, C. Exsense: Extract sensitive information from unstructured data. **Computers & Security**, Elsevier, 2021.

HSIEH, F.; DIAS, R.; PARABONI, I. Author Profiling from Facebook Corpora. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**. Miyazaki, Japan: European Language Resources Association, 2018. p. 1210–132.

IRFAN, R.; KING, C. K.; GRAGES, D.; EWEN, S.; KHAN, S. U.; MADANI, S. A.; KOLODZIEJ, J.; WANG, L.; CHEN, D.; RAYES, A. et al. A survey on text mining in social networks. **The Knowledge Engineering Review**, Cambridge University Press, p. 157–170, 2015.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, p. 11–21, 1972.

KOCHER, M.; SAVOY, J. Unine at clef 2016: Author profiling. In: **Notebook for PAN at CLEF**. Évora,Portugal: CLEF, 2016. p. 10–21.

KRÜGER, S.; HERMANN, B. Can an online service predict gender? on the state-of-the-art in gender identification from texts. In: **International Workshop on Gender Equality in Software Engineering**. Montreal, QC, Canada: IEEE, 2019. p. 13–16.

MARKOV, I.; GÓMEZ-ADORNO, H.; SIDOROV, G. Language-and subtask-dependent feature selection and classifier parameter tuning for author profiling. In: **Notebook for PAN at CLEF**. Dublin, Ireland: CLEF, 2017. p. 222–232.

MARKOV, I.; GÓMEZ-ADORNO, H.; SIDOROV, G.; GELBUKH, A. The winning approach to cross-genre gender identification in russian at rusprofiling 2017. **Training**, p. 1–216, 2017.

MARTINC, M.; SKRJANEC, I.; ZUPAN, K.; POLLAK, S. Pan 2017: Author profiling-gender and language variety prediction. In: **Notebook for PAN at CLEF**. Dublin, Ireland: CLEF, 2017. p. 315–322.

MECHTI, S.; JAOUA, M.; BELGUITH, L. H. Author profiling using style-based features. In: **Notebook for PAN at CLEF 2013**. Valencia, Spain: CLEF, 2013. p. 80–92.

MIKOLOV, T.; CHEN, K.; CORRADO, G. S.; DEAN, J. Efficient estimation of word representations in vector space. In: **International Conference on Learning Representations**. Arizona, EUA: ICLR, 2013.

MIURA, Y.; TANIGUCHI, T.; TANIGUCHI, M.; OHKUMA, T. Author profiling with word+character neural attention network. In: **Notebooks for PAN at CLEF**. Dublin, Ireland: CLEF, 2017. p. 100–114.

MOSTELLER, F.; WALLACE, D. L. **The federalist: Inference and disputed authorship**. Stanford, CA: Addison-Wesley, 1964.

NEAL, T.; SUNDARARAJAN, K.; FATIMA, A.; YAN, Y.; XIANG, Y.; WOODARD, D. Surveying stylometry techniques and applications. **ACM Computing Surveys (CSUR)**, ACM, New York, USA, p. 1–36, 2017.

NGUYEN, D.; GRAVEL, R.; TRIESCHNIGG, D.; MEDER, T. "how old do you think i am?"a study of language and age in twitter. In: **Seventh International AAAI Conference on Weblogs and Social Media**. Massachusetts USA: ICWSM, 2013.

NIVRE, J.; MARNEFFE, M.-C. de; GINTER, F.; HAJIč, J.; MANNING, C. D.; PYYSALO, S.; SCHUSTER, S.; TYERS, F.; ZEMAN, D. Universal dependencies v2: An evergrowing multilingual treebank collection. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 4034–4043.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. JMLR. org, p. 2825–2830, 2011.

PEERSMAN, C.; DAELEMANS, W.; VAERENBERGH, L. V. Predicting age and gender in online social networks. In: **Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents**. New York, NY, USA: Association for Computing Machinery, 2011. p. 37–44.

POULSTON, A.; WASEEM, Z.; STEVENSON, M. Using tf-idf n-gram and word embedding cluster ensembles for author profiling. In: **Notebooks for PAN at CLEF**. Dublin, Ireland: CLEF, 2017. p. 50–61.

RAMOS, R.; NETO, G.; SILVA, B.; MONTEIRO, D.; PARABONI, I.; DIAS, R. Building a corpus for personality-dependent natural language understanding and generation. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**. Miyazaki, Japan: European Language Resources Association, 2018. p. 1138–1145.

RANGEL, F.; ROSSO, P.; POTTHAST, M.; STEIN, B. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. **Working Notes Papers of the CLEF**, p. 1613–1623, 2017.

REAL, L.; OSHIRO, M.; MAFRA, A. B2w-reviews01 an open product reviews corpus. In: **XII Symposium in Information and Human Language Technology and Collocates Events**. Salvador,BA: STIL, 2019. p. 200–208.

REDDY, T. R.; VARDHAN, B. V.; REDDY, P. V. A survey on authorship profiling techniques. **International Journal of Applied Engineering Research**, Research India Publications, Mascara, Algeria, p. 3092–3102, 2016.

SANTOS, H. D. P. dos; WOLOSZYN, V.; VIEIRA, R. BlogSet-BR: A Brazilian Portuguese Blog Corpus. In: **11th International Conference on Language Resources and Evaluation**. Miyazaki, Japan: ELRA, 2018. p. 1110–1123.

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**. Varna, Bulgaria: INCOMA Ltd, 2019. p. 1148–1160.

SAP, M.; PARK, G.; EICHSTAEDT, J.; KERN, M.; STILLWELL, D.; KOSINSKI, M.; UNGAR, L.; SCHWARTZ, H. A. Developing age and gender predictive lexica over social media. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1146–1151.

SAPKOTA, U.; BETHARD, S.; MONTES, M.; SOLORIO, T. Not all character n-grams are created equal: A study in authorship attribution. In: **Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 93–102.

SILVA, J.; BRANCO, A.; CASTRO, S.; REIS, R. Out-of-the-box robust parsing of portuguese. In: **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer, 2010. p. 75–85.

VICENTE, M.; BATISTA, F.; CARVALHO, J. P. C. Gender detection of twitter users based on multiple information sources. In: ____. **Interactions Between Computational Intelligence and Mathematics Part 2**. Berna, SW: Springer International Publishing, 2019. p. 39–54.

VOLLENBROEK, M. B. O.; CARLOTTO, T.; KREUTZ, T.; MEDVEDEVA, M.; POOL, C.; BJERVA, J.; HAAGSMA, H.; NISSIM, M. Gronup: Groningen user profiling. In: **Notebook for PAN at CLEF**. Groningen, The Netherlands: CLEF, 2016. p. 846–857.

WEISS, S. M.; INDURKHYA, N.; ZHANG, T. **Fundamentals of predictive text mining**. London,UK: Springer, 2015.