Uma Abordagem Híbrida para Predição de Faixa Etária de Autores de Textos Escritos na Língua Portuguesa

Alice Rezende Ribeiro
Universidade Federal de Lavras (UFLA) / Departamento de Ciência da Computação
ciceribeiroo@gmail.com

Luiz Henrique de Campos Merschmann
Universidade Federal de Lavras (UFLA) / Departamento de Computação Aplicada
luiz.hcm@ufla.br

Resumo

A crescente quantidade de textos disponíveis na Web torna as ferramentas de mineração de texto essenciais para a extração de informações valiosas para diversas aplicações. No entanto, além dos próprios textos, conhecer as características de seus autores é crucial para algumas organizações. Como os textos podem ser publicados anonimamente, é crescente o interesse em pesquisas voltadas para a criação de técnicas computacionais para inferir as características demográficas de seus autores. Apesar disso, para o problema da predição da faixa etária de autores de textos escritos na língua portuguesa, a quantidade limitada de recursos e o baixo desempenho preditivo evidenciam a necessidade de mais pesquisas focadas nessa tarefa. Assim, este trabalho propõe e avalia uma abordagem que, além de um classificador tradicional, utiliza dicionários de palavras para capturar as especificidades do domínio textual e aprimorar o desempenho preditivo da tarefa de predição da faixa etária. Os resultados experimentais obtidos com a abordagem proposta mostram que explorar as características do domínio dos textos pode contribuir positivamente para o desempenho dessa tarefa.

1 Introdução

Em 2025, a quantidade de usuários da Internet alcançou cerca de dois terços da população mundial, o que equivale a aproximadamente 5,56 bilhões de pessoas (Petrosyan, 2025). Essa popularização do uso da Internet vem modificando a maneira com a qual as pessoas interagem, disponibilizam dados e compartilham opiniões. Impulsionados pelas redes sociais e pelos diferentes tipos de serviços *online* existentes, os textos correspondem a grande parte dos dados veiculados na Web diariamente. Segundo Duarte (2025), dados de 2022 mostram que 24,04 bilhões de textos são gerados pelos usuários da Internet por dia. Nesse cenário, a área de Mineração de Texto torna-se essencial para agregar valor a esses dados, permitindo o processamento dos mesmos para a extração de informações valiosas para diversas aplicações.

Além dos textos publicados, conhecer características como o gênero e a faixa etária de seus autores é de fundamental importância para algumas organizações e setores da indústria e do comércio. Isso porque essa informação pode ser útil para a construção de modelos de recomendação e personalização de serviços e produtos (Riegger et al., 2021), pode colaborar na melhoria do desempenho de tarefas como a análise de sentimentos (Guimarães et al., 2017) e, até mesmo, auxiliar a área de Computação Forense, a qual faz uso de dados sobre autores de publicações postadas na Web em investigações envolvendo crimes cibernéticos (Yu, 2023).

No entanto, como em diversas situações os textos são publicados na Web de forma anônima, a utilização de técnicas computacionais para inferir as características dos seus autores torna-se necessária. Para esse cenário, técnicas vêm sendo propostas por pesquisadores que atuam em uma área de pesquisa denominada Caracterização Autoral, a qual faz uso de recursos da linguística, da estatística e da computação para inferir características como gênero, faixa etária, religiosidade, grau de escolaridade, ocupação e outras.

Vários estudos na área de Caracterização Autoral tiveram origem na PAN-CLEF (*Plagiarism, Authorship, and Near-Duplicate Detection - at the CLEF conference*), uma competição internacional que propõe desafios que incentivam a pesquisa em diferentes áreas de análise de texto. Esse evento é conhecido por resultar em trabalhos que reúnem as técnicas estado-da-arte para os problemas da área e por gerar diversos conjuntos de dados textuais que se tornam referências (Bevendorff et al., 2022).

Embora diversos estudos tenham sido realizados na área de Caracterização Autoral, os progressos alcançados não são homogêneos com relação à característica predita e ao idioma dos textos analisados.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 311–320

Por exemplo, enquanto abordagens para a tarefa de predição de gênero alcançaram desempenhos superiores a 0,9 para a métrica F1 (Dias, 2019; Silva, 2020; Hsieh et al., 2018; Flores et al., 2022; Morais e Merschmann, 2022), estudos voltados para a predição da faixa etária, considerando-a como um problema de classificação multiclasse, obtiveram resultados muito inferiores, variando de 0,41 a 0,67 de F1 (Hsieh et al., 2018; Dias, 2019; Flores et al., 2022; Delmondes Neto, 2021; Sulochana et al., 2024). Além disso, existe uma grande concentração de trabalhos para o idioma inglês (Hsieh et al., 2018), enquanto uma quantidade muito menor de pesquisas, recursos e ferramentas computacionais foi proposta para a língua portuguesa.

Sendo assim, o foco deste trabalho é a predição da faixa etária de autores a partir de seus textos escritos na língua portuguesa. No entanto, diferentemente da abordagem tradicional de classificação de texto utilizada na maioria dos trabalhos encontrados na literatura, este trabalho propõe e avalia uma abordagem alternativa que, além de um classificador tradicional, faz uso de dicionários de palavras com o objetivo de melhorar o desempenho preditivo da tarefa de predição de faixa etária.

A hipótese deste trabalho é que explorar características específicas de um domínio textual (vocabulário associado ao mesmo) pode melhorar o desempenho da classificação da faixa etária de autores de textos escritos na língua portuguesa. Para verificar essa hipótese, foram realizados experimentos comparando a abordagem proposta com as melhores da literatura para as bases de dados avaliadas.

A seguir, a Seção 2 apresenta os trabalhos relacionados à predição da faixa etária, a Seção 3 descreve a abordagem proposta neste trabalho, a Seção 4 detalha os experimentos e os resultados obtidos e, por fim, as conclusões e sugestões de trabalhos futuros são apresentados na Seção 5.

2 Trabalhos Relacionados

Dos trabalhos da literatura que exploraram o problema de predição de faixa etária de autores a partir de seus textos escritos na língua portuguesa, alguns trataram-no como um problema de classificação binária, com as classes adulto e adolescente, outros como um problema de classificação multiclasse, onde cada classe representou uma faixa etária diferente, e um como um problema de classificação multirrótulo, onde os rótulos representavam o gênero e a faixa etária dos autores. Segue uma breve descrição de cada um desses trabalhos.

O trabalho de Guimarães et al. (2017) investigou quais características extraídas de textos publicados na plataforma Twitter são mais relevantes para, em conjunto com os dados do perfil do usuário, predizer corretamente a sua faixa etária. Nesse estudo, diversos algoritmos de classificação foram avaliados para a predição de duas faixas etárias: adolescente e adulto. Os melhores resultados foram obtidos com uma Rede Neural Convolucional Profunda, que alcançou uma medida-F de 0,94.

Já o estudo realizado por Hsieh et al. (2018) comparou diversas representações textuais para a classificação de gênero, faixa etária, grau de religiosidade e formação em Tecnologia da Informação (TI) a partir da base de dados textual B5 Corpus (Ramos et al., 2018). Utilizando regressão logística, foram analisadas as representações textuais *Bag of Words*, n-gramas de caracteres, TF-IDF, LIWC+P e *Word2Vec*. Os melhores resultados para a tarefa de predição de faixa etária foram obtidos com a representação TF-IDF, cujo valor de medida-F variou de 0,56 a 0,61 para as três faixas etárias preditas.

Em (Dias, 2019), diversas tarefas de caracterização autoral, incluindo a predição de faixas etárias como um problema de classificação multiclasse, foram exploradas a partir de bases de dados textuais em diferentes idiomas (português, inglês e espanhol) e pertencentes a diversos domínios (redes sociais, questionários, SMS e blogs). Seis modelos baseados em redes neurais e *Word Embeddings* foram comparados com um *baseline* composto por uma regressão logística com a vetorização de texto usando TF-IDF. Considerando apenas os conjuntos de dados em português utilizados na avaliação da tarefa de predição de faixa etária, a rede neural convolucional associada à vetorização TF-IDF apresentou o melhor desempenho (medida F1) para as bases B5 Corpus (0,62) e BlogSet BR (0,45), enquanto, para a base BR Moral, nenhuma das abordagens ultrapassou o desempenho do *baseline*, que obteve F1 igual a 0,41.

Do mesmo modo que em (Guimarães et al., 2017), utilizando uma base de dados formada por características extraídas dos textos e dos perfis dos seus autores na rede social Twitter, o trabalho (Silva, 2020) realizou a predição de faixa etária considerando apenas as classes adolescente e adulto. No entanto, nesse trabalho, o autor aplicou as estratégias de classificação multirrótulo *Classifier Chains* e *Label Powerset* em conjunto com diferentes classificadores para predição de gênero

e faixa etária. Como resultado, a abordagem proposta alcançou 0,92 de micro F1 na predição de faixa etária, não conseguindo superar os resultados reportados em (Guimarães et al., 2017).

Delmondes Neto (2021) avaliou modelos de classificação baseados em redes neurais artificiais para as tarefas de classificação de gênero e faixa etária interdomínio, ou seja, onde os modelos treinados foram testados em dados de diferentes domínios (avaliações de produtos, *posts* do Facebook, conteúdos de *blogs*, solicitações ao governo e *tweets*). Os experimentos envolvendo quatro bases de dados textuais mostraram que, para a tarefa de classificação de faixa etária, mesmo para o modelo interdomínio que apresentou o melhor resultado (BERT), houve uma perda na medida macro F1 em relação aos modelos de domínio único.

No trabalho apresentado em (Flores et al., 2022), a partir de uma coleção de textos referentes a solicitações feitas por cidadãos ao governo brasileiro por meio de um sistema denominado e-SIC, os autores avaliaram quatro tarefas de caracterização autoral: a predição de gênero, de faixa etária, de nível de escolaridade e de ocupação. Dentre os diversos modelos de classificação avaliados, a LSTM (*Long Short-Term Memory*) apresentou o melhor desempenho para as quatro tarefas de caracterização autoral, sendo que, na predição de faixa etária, obteve uma F1 ponderada igual a 0,67.

3 Abordagem Proposta

A abordagem proposta neste trabalho é composta por dois módulos que são utilizados de maneira sequencial com o objetivo de predizer a faixa etária do autor de um texto. O primeiro deles, denominado modelo baseado em dicionários, faz uso de múltiplos dicionários de palavras associadas a pesos para, a partir da incidência dessas palavras num determinado texto cuja faixa etária do seu autor é desconhecida, tentar predizer a faixa etária do mesmo. Nesses dicionários, o peso associado a cada palavra indica o grau de aderência da mesma a textos escritos por autores de uma determinada faixa etária. Já o segundo módulo da abordagem corresponde a um modelo de classificação multiclasse tradicional.

A Figura 1 ilustra a abordagem híbrida proposta. Dado um texto para o qual se deseja predizer a faixa etária do seu autor, o processo começa com o pré-processamento desse texto, onde são removidas as *stopwords*, as pontuações e os caracteres específicos do domínio (por exemplo, tags HTML para *blogs* e *hashtags* para textos extraídos de redes sociais). Além disso, todos os caracteres do texto são convertidos para caixa baixa antes de ele ser vetorizado. Em seguida, utilizando-se o modelo baseado em dicionários, tenta-se definir a faixa etária do autor do texto. Quando o modelo baseado em dicionários não é capaz de predizer a faixa etária do autor, o texto é submetido ao segundo módulo da abordagem, um modelo de classificação tradicional, para a predição da faixa etária.

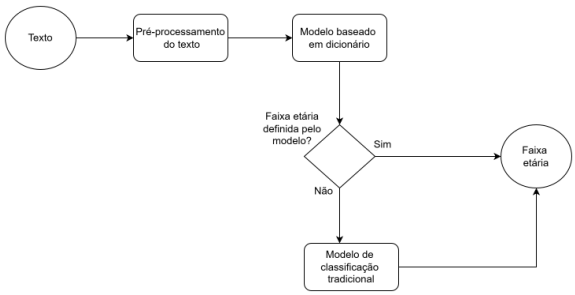

Figura 1: Abordagem híbrida proposta

As subseções a seguir detalham os dois módulos que compõem a abordagem proposta.

3.1 Modelo Baseado em Dicionários

A utilização de um modelo baseado em dicionários para definição da faixa etária do autor de um texto parte do pressuposto de que pessoas de faixas etárias distintas utilizam vocabulários parcialmente diferentes para se expressar textualmente. Portanto, a ideia central deste modelo é usar as palavras existentes no texto para tentar identificar a faixa etária do seu autor. O modelo aqui proposto é baseado na abordagem apresentada em (Sap et al., 2014), que fez uso de um classificador SVM com *kernel* linear para criar um dicionário de palavras que é utilizado para a definição do gênero do autor de um texto.

Enquanto na proposta apresentada em (Sap et al., 2014) um único dicionário é criado a partir do treinamento de um classificador SVM Linear para resolver um problema de classificação binária (predição de gênero), neste trabalho, uma adaptação foi realizada para lidar com um problema de classificação multiclasse (predição de faixas etárias), fazendo com que múltiplos dicionários sejam utilizados para a definição da classe de uma instância. Além disso, diferentemente de (Sap et al., 2014), que treina um classificador SVM utilizando a frequência relativa das palavras contidas num conjunto de textos, os valores TF-IDF (*Term*

Frequency-Inverse Document Frequency) das palavras são utilizados como atributos preditivos.

O treinamento do classificador SVM Linear resulta na criação de dicionários de palavras associadas a pesos, onde as palavras correspondem aos atributos utilizados no seu treinamento e os pesos são os coeficientes que definem (juntamente com a constante denominada *intercept* – w_0) cada hiperplano de separação criado pelo classificador SVM no espaço n-dimensional (onde n é o número de atributos da base de dados de treinamento). O raciocínio é que, quanto maior for o valor absoluto do peso (coeficiente na equação do hiperplano) associado a uma determinada palavra, mais relevante é essa palavra para determinada(s) classe(s).

No problema de classificação de faixas etárias abordado neste estudo, os textos podem estar associados a três faixas etárias diferentes. Portanto, um modelo SVM Linear (usando a estratégia "*one-versus-rest*") é treinado para a criação de três dicionários de palavras que possuem pesos associados, onde esses pesos indicam a relevância das palavras para determinada(s) faixa(s) etária(s). Mais especificamente, cada dicionário contém palavras associadas a pesos positivos e negativos, sendo que o sinal e a magnitude dos pesos são os indicadores utilizados para definir que a palavra é predominantemente relacionada com textos escritos por autores de determinada(s) faixa(s) etária(s).

Como o peso associado a cada palavra nos dicionários indica a sua relevância para alguma(s) faixa(s) etária(s), diferentemente do que foi realizado em (Sap et al., 2014), neste trabalho, utiliza-se um limiar (definido empiricamente) para os pesos associados às palavras de modo a manter nos dicionários somente as palavras que possuem maior poder discriminativo para a predição das faixas etárias. Para isso, a estratégia utilizada foi manter nos dicionários originalmente gerados pelo classificador SVM Linear somente um percentual das palavras com os maiores pesos em valor absoluto.

A definição da faixa etária do autor de um texto é obtida a partir do conjunto de valores dos índices age_lex_A, age_lex_B e age_lex_C, calculados a partir dos dicionários construídos para as faixas etárias A, B e C, respectivamente. Observe na Equação 1 que o cálculo desses índices leva em consideração o peso associado a cada palavra do texto presente no dicionário da faixa etária X e o valor TF-IDF das mesmas no texto em questão.

$$age\_lex_X = \left(\sum_{p \in lex_X} w_p * x_p \right) + w_0, \quad (1)$$

onde w_p é o peso associado à palavra p presente no dicionário da faixa etária $X \in \{A, B, C\}$, x_p é o valor TF-IDF da palavra p no texto e w_0 é a constante denominada *intercept*.

O valor do age_lex_X para uma determinada instância define a sua posição com relação ao hiperplano de separação usado na criação do dicionário da faixa etária X e, portanto, pode ser usado para definir a faixa etária da mesma. A Figura 2 ilustra, num espaço bidimensional, nove instâncias (i_1 até i_9) pertencentes a três faixas etárias distintas (A, B e C) e três hiperplanos de separação obtidos com o treinamento de um classificador SVM Linear a partir dessas instâncias. Nesse exemplo, as instâncias i_1 até i_3, i_4 até i_6 e i_7 até i_9 estão associadas a autores nas faixas etárias A, B e C, respectivamente. Além disso, tem-se outras duas instâncias, denominadas i_k e i_y, para as quais se deseja definir a faixa etária do seu autor.

Considerando o exemplo da Figura 2, observa-se que as instâncias da faixa etária A encontram-se acima do hiperplano A e, portanto, possuem $age\_lex_A > 0$. Isso significa que uma nova instância só será atribuída à faixa etária A se ela tiver $age\_lex_A > 0$. No entanto, essa não é a única condição necessária pois, como mostra a Figura 2, instâncias como a i_6 e a i_9 apresentam $age\_lex_A > 0$ e não pertencem à faixa etária A. Portanto, a atribuição de uma faixa etária X ($X = A, B$ ou C) para uma nova instância depende de uma análise conjunta dos valores de

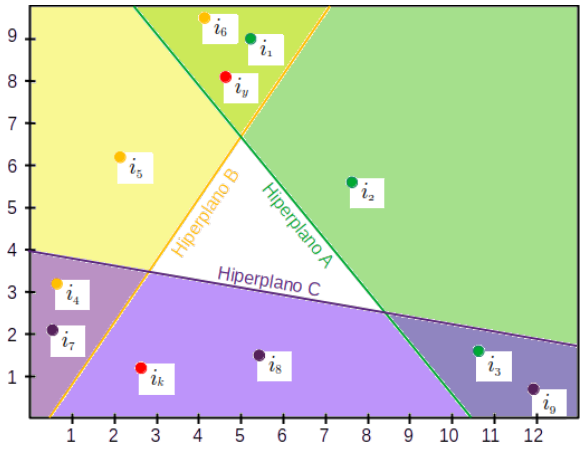

Figura 2: Exemplo de definição da faixa etária a partir dos dicionários

age_lex_A, age_lex_B e age_lex_C calculados para essa instância. Sendo assim, neste exemplo, têm-se as seguintes condições para a atribuição de cada uma das faixas etárias a uma nova instância:

$$\text{Faixa etária} = \begin{cases} A, & \text{se } age\_lex_A > 0 \text{ e } age\_lex_B < 0 \text{ e } age\_lex_C > 0 \\ B, & \text{se } age\_lex_A < 0 \text{ e } age\_lex_B > 0 \text{ e } age\_lex_C > 0 \\ C, & \text{se } age\_lex_A < 0 \text{ e } age\_lex_B < 0 \text{ e } age\_lex_C < 0 \end{cases}$$

As condições, como as exemplificadas anteriormente, são formuladas para cada base de dados a partir da posição das instâncias de treinamento de cada classe (faixa etária) em relação aos hiperplanos de separação definidos pelo SVM Linear. Vale observar que, dada uma nova instância para a qual se deseja definir a faixa etária do seu autor, pode acontecer de ela não atender a nenhuma das condições necessárias para definição da sua classe e, nesse caso, ela não é classificada pelo modelo baseado em dicionários. Continuando com o exemplo da Figura 2, enquanto a instância i_k atende à condição para ser classificada como faixa etária C, a instância i_y não atende as condições para nenhuma das faixas etárias e, portanto, não é classificada por esse módulo da abordagem.

Em resumo, dada uma instância para ser classificada, os cálculos dos índices age_lex_X, para $X = A, B$ e C, devem ser realizados para se tentar definir a sua classe. No entanto, ao invés de se utilizar apenas o sinal (positivo ou negativo) do valor do age_lex_X para definir a classe da instância, na abordagem proposta neste trabalho, adotam-se valores limiares α_age_A, α_age_B e α_age_C para os índices age_lex_A, age_lex_B e age_lex_C, respectivamente, de modo que um texto só é considerado pertencente a uma determinada faixa etária se atender às seguintes condições para aquela faixa etária:

$$(age\_lex_X - \alpha\_age_X) * \alpha\_age_X \geq 0, \; \forall X = \{A, B, C\} \quad (2)$$

Esses limiares têm como objetivo aumentar o grau de certeza da classificação, evitando que instâncias (textos) que não contêm um vocabulário típico de uma única faixa etária sejam atribuídas a alguma faixa etária. Ou seja, o modelo baseado em dicionários classifica somente as instâncias que respeitam determinadas distâncias (definidas pelos α_age_X) para os hiperplanos de separação que definem os dicionários.

Esses limiares α_age_X são definidos a partir dos índices age_lex_X calculados para as instâncias de treinamento da classe X, ou seja, a partir das mesmas instâncias que foram utilizadas no treinamento do modelo SVM Linear para a construção dos dicionários. Mais especificamente, a partir do conjunto de valores age_lex_X calculados para as instâncias de treinamento da faixa etária X, tanto a média quanto os quartis (separatrizes que dividem o conjunto em quatro partes iguais) são considerados como possíveis valores para α_age_X. Vale lembrar que esses limiares visam maximizar o desempenho preditivo do modelo baseado em dicionários, ainda que, para isso, se tenha uma redução na quantidade de instâncias para as quais esse modelo consegue realizar a predição da faixa etária. Os melhores valores de α_age_X para cada base de dados são definidos empiricamente.

3.2 Modelo de Classificação Tradicional

Seguindo o que a literatura apresenta para a predição da faixa etária de autores de textos na língua portuguesa como um problema de classificação multiclasse, o segundo módulo da abordagem híbrida proposta faz uso de um modelo de classificação tradicional. Esse modelo de classificação é treinado a partir de um conjunto de dados de treinamento formado por textos pré-processados e vetorizados. Desse modo, todos os textos que não tiveram a faixa etária do seu autor definida pelo primeiro módulo da abordagem (modelo baseado em dicionários) são classificados por esse modelo de classificação tradicional.

4 Experimentos Computacionais

Esta seção apresenta as bases de dados textuais utilizadas nos experimentos (Seção 4.1), as configurações utilizadas nas implementações realizadas (Seção 4.2) e as análises dos resultados obtidos a partir da abordagem proposta em comparação àqueles alcançados pelas abordagens encontradas na literatura (Seção 4.3).

4.1 Bases de Dados

Os experimentos computacionais foram conduzidos a partir de quatro bases de dados utilizadas por outros trabalhos da literatura que contêm textos em português. Essas bases apresentam características diversas com relação à origem dos dados (*websites*, *blogs* e redes sociais), ao conteúdo e ao tamanho dos textos. A seguir, tem-se uma breve descrição de cada uma das bases de dados.

BlogSet BR: base com 2604 textos de *blogs* publicados por diferentes usuários. Cada texto reúne uma ou mais publicações de um autor. Criada por Santos et al. (2018), essa base foi extraída

da plataforma Blogspot a partir de mais de 7 milhões de textos que abordam temas diversos, indo desde cuidados pessoais até política internacional. Essa base, já utilizada em (Dias, 2019) e (Delmondes Neto, 2021), possui a distribuição das instâncias nas faixas etárias mostrada na Tabela 1.

Tabela 1: Distribuição das instâncias na base BlogSet BR

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

BR Moral: base com 510 textos opinativos gerados por autores distintos. Essa base de dados, disponibilizada por Santos e Paraboni (2019) e também utilizada em (Dias, 2019), contém textos que abordam temas como religião, política, legalização de drogas, pena de morte e outros. A Tabela 2 apresenta a distribuição das instâncias nas diferentes faixas etárias.

Tabela 2: Distribuição das instâncias na base BR Moral

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

B5 Corpus: base contendo 516 textos extraídos de postagens de usuários distintos no Facebook, cada um reunindo até 1000 postagens de um usuário. Neste trabalho, somente o subconjunto dos textos da base original que continham as informações de faixa etária de seus autores foi utilizado. Criada por Ramos et al. (2018), essa base faz parte de um corpus com textos informais sobre diversos temas e foi utilizada em (Hsieh et al., 2018) e em (Dias, 2019). A Tabela 3 apresenta a distribuição das instâncias de cada faixa etária.

Tabela 3: Distribuição das instâncias na base B5 Corpus

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

e-SIC: base com 47.762 textos obtidos no e-SIC (Sistema Eletrônico do Serviço de Informações ao Cidadão) disponibilizado pelo governo brasileiro. Criada por Flores et al. (2022), essa base possui textos que correspondem a requisições feitas pelos cidadãos ao governo, englobando tópicos como impostos, organizações e políticas públicas. A distribuição das instâncias nas faixas etárias é mostrada na Tabela 4.

Tabela 4: Distribuição das instâncias na base e-SIC

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

A Tabela 5 resume as principais características das bases de dados utilizadas neste trabalho. A coluna 'Domínio' especifica a natureza dos textos de cada base. Em seguida, a coluna 'Instâncias' mostra a quantidade de textos disponíveis nas bases. Por fim, a coluna 'Palavras/Texto' apresenta a quantidade média de palavras por texto.

Tabela 5: Características da bases de dados

Base	Domínio	Instâncias	Palavras/Texto
BlogSet BR	Blogs	2.604	5.390,18
BR Moral	Opiniões	510	427,81
B5 Corpus	*Facebook*	516	2.846
e-SIC	e-Gov	47.762	73,64

4.2 Configuração Experimental

Toda a implementação de código para o pré-processamento dos dados e construção dos dois módulos da abordagem proposta foi realizada na linguagem *Python* e fez uso de módulos e bibliotecas da linguagem, tais como *NLTK* (Bird et al., 2009), *Scikit-learn* (Pedregosa et al., 2011) e *Keras* (Chollet, 2015).

As subseções a seguir detalham as implementações do modelo baseado em dicionários e dos modelos de classificação tradicionais, cujos códigos estão disponíveis em `https://github.com/luizhcm/AgePrediction`.

4.2.1 Modelo Baseado em Dicionários

Na etapa de pré-processamento dos textos, a identificação e remoção das *stopwords* foi realizada a partir da lista de *stopwords* para a língua portuguesa presente na biblioteca *NLTK*. Em seguida, a construção do modelo baseado em dicionários

começa por uma redução no tamanho do vocabulário a ser utilizado no treinamento do classificador SVM Linear para a obtenção dos dicionários. Por não contribuírem no processo de distinção das faixas etárias, foram removidas todas as palavras que apareciam em menos de 1% dos textos de cada base de dados. Após essa remoção, a vetorização dos textos foi feita utilizando-se o módulo *TfidfVectorizer* da biblioteca *Scikit-learn*.

Para a criação dos dicionários, o treinamento do classificador SVM fez uso do módulo SVC da biblioteca *Scikit-learn*, com o parâmetro *kernel* definido como 'linear' e *decision_function_shape* como 'ovr' (estratégia '*one versus rest*'). Já os valores dos hiperparâmetros C, max_iter e tol foram definidos utilizando-se a estratégia de calibração de hiperparâmetros denominada *Grid Search*, a qual otimizou a medida macro F1 a partir de uma validação cruzada com 10 partições para a base BR Moral e 5 partições para as demais. A Tabela 6 apresenta, para cada base de dados, os melhores hiperparâmetros encontrados pela estratégia *Grid Search* para cada base de dados.

Tabela 6: Hiperparâmetros do SVM

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

Visando manter nos dicionários somente as palavras com maior poder discriminatório das faixas etárias (ver Seção 3.1), experimentos foram realizados para avaliar as reduções de 0%, 50% e 75% da quantidade de palavras dos dicionários. Exceto para a base de dados e-SIC, onde a redução de 50% do dicionário originalmente obtido foi a que apresentou o melhor resultado, para as demais bases de dados, a utilização do dicionário sem qualquer redução na quantidade de palavras foi a configuração que alcançou o melhor desempenho.

Por fim, na abordagem proposta neste trabalho, a definição da faixa etária do autor de um texto a partir dos dicionários é feita somente se os valores dos índices age_lex_X respeitarem os limiares α_age_X (ver Equação 2). Experimentos foram realizados para avaliar o desempenho preditivo desse módulo para os seguintes valores de α_age_X: zero, a média, o primeiro quartil (Q1), o segundo quartil (Q2) e o terceiro quartil (Q3) do conjunto de valores age_lex_X das instâncias de treinamento da faixa etária X. Para as bases BlogSet BR e BR Moral, em todas as faixas etárias, os melhores desempenhos foram alcançados com $\alpha\_age_X = 0$. Já para as bases B5 Corpus e e-SIC, os melhores desempenhos para cada faixa etária foram obtidos com os valores de α_age_X apresentados na Tabela 7.

Tabela 7: α_age_X adotado por base de dados

Base	Faixa Etária	α_age_A	α_age_B	α_age_C
B5 Corpus	A	Q3	Q1	Q1
	B	Q1	Q3	Q1
	C	Q1	Q1	Q3
e-SIC	A	Q2	0	0
	B	0	Q2	0
	C	0	0	Q2

4.2.2 Modelos de Classificação Tradicionais

Uma vez que a avaliação da abordagem proposta é feita comparando-a com a abordagem tradicional de classificação utilizada nos trabalhos da literatura, ao invés de se realizar a comparação tomando como base os resultados reportados nesses trabalhos, para se ter uma comparação precisa, as abordagens da literatura foram reimplementadas de modo que as mesmas partições de dados utilizadas no treinamento e no teste dos modelos fossem empregadas na comparação das abordagens. Essa reimplementação seguiu, para cada base de dados, exatamente a mesma configuração experimental (métodos de vetorização do texto, algoritmos de classificação e as técnicas de avaliação dos classificadores) utilizada nos trabalhos de referência, cujos detalhes encontram-se descritos na Seção 2. Além disso, esses classificadores reimplementados de acordo com os trabalhos de referência foram usados no segundo módulo da abordagem proposta neste trabalho.

A Tabela 8 mostra, para cada base de dados, o trabalho da literatura que foi utilizado como referência (coluna 'Ref.') e um resumo da configuração experimental utilizada no mesmo, a saber: a técnica utilizada na avaliação da tarefa de classificação (coluna 'Avaliação'), o método de vetorização do texto ('Vetorização') e o método adotado para a classificação da faixa etária (coluna 'Classificador').

Tabela 8: Resumo da configuração experimental

Base	Ref.	Avaliação	Vetorização	Classificador
BlogSet BR	(Dias, 2019)	*Holdout* 80/20	TF-IDF	CNN
BR Moral	(Dias, 2019)	10-Validação Cruzada	TF-IDF	Regressão Logística
B5 Corpus	(Dias, 2019)	*Holdout* 80/20	TF-IDF	CNN
e-SIC	(Flores et al., 2022)	*Holdout* 80/20	*Tokenizer keras*	LSTM

Para a reimplementação dos classificadores utilizados nos trabalhos de referência, a biblioteca *sklearn* foi utilizada para a regressão logística e as bibliotecas *keras*, *tensorflow* e *scikeras* para as redes neurais convolucionais (CNNs) e a *Long Short Term Memory* (LSTM).

4.3 Análise dos Resultados

Nesta seção, inicialmente são apresentados os resultados alcançados com a reimplementação das abordagens propostas na literatura, os quais servirão como base de comparação com a abordagem híbrida proposta neste trabalho. Em seguida, os resultados obtidos com a abordagem híbrida são apresentados e discutidos.

A Tabela 9 mostra os resultados obtidos a partir da reimplementação dos trabalhos da literatura. Nessa tabela, a métrica adotada em cada trabalho de referência é apresentada na coluna 'Métrica', os resultados reportados nesses trabalhos são mostrados na coluna 'Reportado' e, por fim, a coluna '*Baseline*' apresenta os resultados alcançados a partir da reimplementação das abordagens desses trabalhos. É importante destacar que, como a base BR Moral é avaliada a partir de uma validação cruzada com 10 partições, os resultados apresentados correspondem às médias dos desempenhos nas 10 partições de teste. Uma vez que o protocolo experimental adotado na reimplementação foi o mesmo dos trabalhos de referência, as diferenças existentes entre os resultados reportados na literatura e os obtidos a partir das reimplementações devem-se à aleatoriedade inerente ao processo de particionamento dos dados para geração dos conjuntos de treinamento e teste utilizados na construção e avaliação dos modelos preditivos e ao processo de calibração de hiperparâmetros dos algoritmos.

Tabela 9: Resultados da reimplementação dos trabalhos de referência

Base de Dados	Métrica	Reportado	*Baseline*
BlogSet BR	Macro F1	0,45	0,48
BR Moral	Macro F1	0,41	0,43
B5 Corpus	Macro F1	0,62	0,61
e-SIC	F1 ponderada	0,67	0,57

A Tabela 10 apresenta os desempenhos preditivos (macro F1) da abordagem proposta (coluna 'Abordagem Híbrida') e das abordagens adotadas nos trabalhos da literatura (coluna '*Baseline*'). Nessa tabela, os resultados destacados em negrito correspondem ao maior valor de desempenho preditivo para cada base de dados. Vale lembrar que, para a base BR Moral, os resultados correspondem a valores médios das 10 partições de teste e, por isso, é apresentado também o desvio padrão. Já as demais bases são avaliadas a partir de uma única partição de teste obtida a partir da técnica *holdout*.

Tabela 10: Comparação entre a abordagem híbrida e o *baseline*

Base de dados	*Baseline*	Abordagem Híbrida
BlogSet BR	0,48	**0,49**
BR Moral	0,43±0,06	**0,47±0,07**
B5 Corpus	0,61	**0,62**
e-SIC	**0,57**	**0,57**

Os resultados apresentados na Tabela 10 mostram que a abordagem híbrida proposta alcançou um resultado sempre melhor ou igual ao das abordagens dos trabalhos de referência (*baseline*).

Para entender a contribuição do modelo baseado em dicionários no desempenho da abordagem híbrida, a Tabela 11 traz um comparativo entre o desempenho (macro F1) do modelo baseado em dicionários (coluna 'Dicionários') e aquele que seria obtido pelo segundo módulo da abordagem proposta (coluna 'Classificador Tradicional') para o conjunto de instâncias classificadas pelo modelo baseado em dicionários. Além disso, essa tabela apresenta, na coluna 'Cobertura', o percentual das instâncias de teste que o modelo baseado em dicionários conseguiu classificar.

Tabela 11: Comparação entre o modelo baseado em dicionários e a abordagem de classificação tradicional

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

A partir dos resultados das Tabelas 10 e 11, pode-se observar que o ganho de desempenho da abordagem híbrida em relação ao *baseline* não depende somente do bom desempenho do modelo baseado em dicionário, mas também da cobertura alcançada pelo mesmo. Como se pode ver na Tabela 11, comparativamente ao classificador tradicional, o modelo baseado em dicionários apresentou os maiores ganhos de desempenho para as bases BR Moral e B5 Corpus. No entanto, esse ganho contribuiu de maneira mais acentuada no desempenho final da abordagem híbrida somente para a base BR Moral,

que foi exatamente a base para a qual o modelo baseado em dicionários teve uma maior cobertura (78,98% contra 10,58% da base B5 Corpus).

5 Conclusão

As redes sociais e os diferentes tipos de serviços *online* existentes produziram um aumento significativo da quantidade de textos atualmente disponíveis na Web. Nesse cenário, cresce a cada dia a importância do desenvolvimento de ferramentas de mineração de texto para auxiliar na análise desse grande volume de dados.

No entanto, para algumas organizações e setores da indústria e do comércio, além dos textos, conhecer características como o gênero e a faixa etária das pessoas que os publicam na Web é de fundamental importância para suas estratégias de atuação. Como os textos podem ser publicados de forma anônima, a área de estudo denominada Caracterização Autoral tem como foco a proposição de abordagens computacionais para inferir as características dos autores a partir dos seus textos.

Apesar de existirem diferentes estudos nessa área, o problema de predição da faixa etária, como um problema de classificação multiclasse, ainda é desafiador por apresentar um desempenho preditivo inferior ao de outras características como, por exemplo, o gênero. Além disso, as abordagens propostas na literatura para a predição da faixa etária geralmente se limitam a um processo tradicional de classificação de textos.

Sendo assim, este trabalho contribui nessa área de estudo propondo e avaliando uma abordagem que faz uso de dois módulos para predizer a faixa etária de autores de textos escritos na língua portuguesa. Mais especificamente, além de um modelo de classificação tradicional, a abordagem proposta faz uso de múltiplos dicionários de palavras para capturar características específicas do domínio textual com o objetivo de melhorar o desempenho preditivo da tarefa de predição de faixa etária.

Os experimentos computacionais realizados para avaliar a abordagem proposta mostraram que o uso de um modelo baseado em dicionários em conjunto com um classificador tradicional pode contribuir na melhoria de desempenho da tarefa de predição de faixa etária. Para as bases de dados avaliadas neste trabalho, a abordagem proposta alcançou desempenhos sempre melhores ou iguais àqueles alcançados pelas propostas encontradas na literatura.

Como trabalhos futuros, pode-se avaliar a abordagem proposta utilizando-se bases de dados com textos escritos em outros idiomas, uma vez que ela não é específica para textos na língua portuguesa. Além disso, formas alternativas de construção e utilização dos dicionários, visando aumentar a cobertura do primeiro módulo da abordagem, podem ser avaliadas com o objetivo de alcançar desempenhos preditivos ainda melhores.

Agradecimentos

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

Referências

Janek Bevendorff, Berta Chulvi, Elisabetta Fersini, Annina Heini, Mike Kestemont, Krzysztof Kredens, Maximilian Mayerl, Reynier Bueno, Piotr Pezik, Martin Potthast, Francisco Rangel Pardo, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, e Eva Zangerle. 2022. *Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection*, páginas 382–394. Springer International Publishing.

Steven Bird, Ewan Klein, e Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., Sebastopol, CA.

François Chollet. 2015. Keras. `https://keras.io`.

José Pereira Delmondes Neto. 2021. Caracterização autoral interdomínio a partir de textos. Dissertação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo.

Rafael Felipe Sandroni Dias. 2019. Caracterização autoral a partir de textos utilizando redes neurais artificiais. Dissertação de mestrado, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.

Fabio Duarte. 2025. Amount of data created daily. `https://explodingtopics.com/blog/data-generated-per-day`. Acessado em: 09 de agosto de 2025.

Arthur Marçal Flores, Matheus Camasmie Pavan, e Ivandré Paraboni. 2022. User profiling and satisfaction inference in public information access services. *Journal of Intelligent Information Systems*, 58(1):67–89.

Rita Georgina Guimarães, Renata Lopes Rosa, Denise De Gaetano, Demóstenes Zegarra Rodríguez, e Graça Bressan. 2017. Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816.

Fernando Hsieh, Rafael Dias, e Ivandré Paraboni. 2018. Author profiling from facebook corpora. Em *Proceedings of the Eleventh International Conference on*

Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

João Morais e Luiz Merschmann. 2022. A cascade approach for gender prediction from texts in portuguese language. Em *Proceedings of the 28th Brazilian Symposium on Multimedia and the Web*, páginas 151–158, Porto Alegre, RS, Brasil. SBC.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, e E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ani Petrosyan. 2025. Internet usage worldwide - statistics & facts. https://www.statista.com/topics/1145/internet-usage-worldwide/#topicOverview. Acessado em: 09 de agosto de 2025.

Ricelli Ramos, Georges Neto, Barbara Barbosa Claudino Silva, Danielle Sampaio Monteiro, Ivandré Paraboni, e Rafael Dias. 2018. Building a corpus for personality-dependent natural language understanding and generation. Em *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Anne-Sophie Riegger, Jan F. Klein, Katrin Merfeld, e Sven Henkel. 2021. Technology-enabled personalization in retail stores: Understanding drivers and barriers. *Journal of Business Research*, 123:140–155.

Henrique dos Santos, Vinicius Woloszyn, e Renata Vieira. 2018. Blogset-br: A brazilian portuguese blog corpus. Em *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wesley Santos e Ivandré Paraboni. 2019. Moral stance recognition and polarity classification from twitter and elicited text. Em *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgaria. INCOMA Ltd.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, e H. Schwartz. 2014. Developing age and gender predictive lexica over social media. Em *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1146–1151.

Douglas Henrique Silva. 2020. Classificação de gêneros e faixas etárias em redes sociais online por meio de técnicas de aprendizagem multidimensional. Dissertação de mestrado, Universidade Federal de Lavras.

Balla Charishma Sulochana, Bhavya Sri Pragada, Boya Chaitanya Kiran, Gaddam Anvith Reddy, e Manju Venugopalan. 2024. Author identity unveiled: Gender and age prediction from textual patterns using bert. Em *4th International Conference on Intelligent Technologies*, páginas 1–6.

Szde Yu. 2023. Cyber profiling: Predicting political orientation with socmint. *Telematics and Informatics Reports*, 10:100058.