Avaliação de uma Abordagem Hierárquica para Predição de Faixa Etária de Autores de Textos Escritos em Português

Alice Rezende Ribeiro[1], **Luiz Henrique de Campos Merschmann**[2]

[1] Departamento de Ciência da Computação
Universidade Federal de Lavras (UFLA) – Lavras – MG – Brasil

[2]Departamento de Computação Aplicada
Universidade Federal de Lavras (UFLA) – Lavras – MG – Brasil

alice.ribeiro2@estudante.ufla.br,luiz.hcm@ufla.br

Abstract. *Author profiling, a field of study that uses computational models to infer demographic characteristics of authors from their texts, is becoming increasingly important for various applications. Despite the growing interest in research in this area, the number of techniques and computational resources proposed in the literature for the Portuguese language is still minimal compared to those available for other languages. Moreover, most studies approach the prediction of characteristics as a flat classification problem. Therefore, this work contributes by proposing and evaluating a hierarchical classification approach involving more than one demographic characteristic for predicting the age group of an author based on his text.*

Resumo. *A caracterização autoral, área de estudo que usa modelos computacionais para inferir características demográficas de autores de textos, torna-se cada dia mais importante para diversas aplicações. Apesar do crescente interesse por pesquisas nessa área, a quantidade de técnicas e recursos computacionais propostos na literatura para a língua portuguesa ainda é muito pequena quando comparada àquela disponível para outros idiomas. Além disso, a maioria dos trabalhos aborda a predição das características como um problema de classificação plana. Desse modo, este trabalho contribui propondo e avaliando uma abordagem de classificação hierárquica envolvendo mais de uma característica demográfica para a predição da faixa etária do autor de um texto.*

1. Introdução

A Internet continua transformando a forma como as pessoas se conectam, disponibilizam informações e compartilham opiniões. Em 2024, a quantidade de usuários da Internet no mundo alcançou a marca de 5,44 bilhões, o que representa cerca de dois terços da população mundial [Statista 2025]. Com isso, a quantidade de dados criados, capturados e consumidos pelas pessoas continua a crescer ano após ano, estimando-se que, em 2025, o uso da Internet alcance aproximadamente 181 zettabytes de dados [DOMO 2023].

Essa massa de dados enorme tem sido utilizada para impulsionar a indústria de inteligência de negócios e análise de dados. Embora diversos formatos de dados estejam presentes na Internet contemporânea, grande parte deles é textual, de forma que a mineração de dados textuais é uma das ferramentas mais importantes da atualidade.

Conhecer as características dos usuários da Internet é de grande valor para empresas de *e-commerce* e aquelas que realizam recomendação de conteúdos [Goldenberg et al. 2021] pois, desse modo, elas conseguem oferecer produtos e serviços personalizados para esses usuários. Ademais, o conhecimento das características desses usuários também tem grande importância para a área forense digital [Dias 2019], auxiliando, por exemplo, em investigações de crimes cibernéticos.

No entanto, nem sempre as informações referentes aos usuários da Internet acompanham as publicações postadas pelos mesmos. Para essas situações, é possível utilizar técnicas computacionais que vêm sendo propostas por pesquisadores que atuam numa área denominada Caracterização Autoral, cujo objetivo é inferir as características dos autores a partir dos seus textos. Ainda que os autores não escolham conscientemente revelar suas características nos seus textos, elas podem ser obtidas a partir dos padrões linguísticos existentes nos mesmos, uma vez que esses padrões costumam ser compartilhados por autores com características demográficas semelhantes, tais como faixa etária, gênero e nível de escolaridade.

Dada a relevância dessa área de estudo, a caracterização autoral passou a fazer parte das tarefas propostas pela PAN-CLEF (*Plagiarism, Authorship, and Near-Duplicate Detection - at the CLEF conference*), uma competição de referência internacional que propõe desafios que estimulam a pesquisa em diferentes áreas de análise de texto, como a detecção de plágio, identificação de autoria e a caracterização autoral. Impulsionados pelas bases de dados disponibilizadas por essa competição, diversos trabalhos foram apresentados na literatura propondo diferentes abordagens e algoritmos para resolução de tarefas como a identificação de gênero e faixa etária de autores a partir de seus textos [Pizarro 2019, Takahashi et al. 2018, Basile et al. 2017, Vollenbroek et al. 2016, Álvarez-Carmona et al. 2015, López-Monroy 2014, López-Monroy et al. 2013].

Apesar da crescente contribuição de trabalhos nessa área de caracterização autoral, os avanços alcançados não ocorrem de forma homogênea para os diferentes idiomas. O que se observa é uma concentração de trabalhos focados em poucos idiomas amplamente utilizados [Hsieh et al. 2018], como é o caso do inglês, e uma quantidade muito menor de pesquisas, recursos e ferramentas computacionais para idiomas como o português.

Grande parte das propostas apresentadas na literatura aborda o problema de caracterização autoral como uma tarefa de classificação, ou seja, a partir de um corpus textual rotulado, treinam-se modelos de classificação para realizar a predição das características de um autor (gênero, faixa etária, nível de escolaridade etc.) a partir do seu texto. No entanto, a maioria dos trabalhos da literatura realiza essa predição das características de forma isolada, ou seja, com um problema de classificação plana monorrótulo.

O foco deste trabalho é a predição da faixa etária de autores de textos escritos na língua portuguesa. No entanto, diferentemente da abordagem plana comumente adotada em trabalhos da literatura, o problema de classificação de faixa etária será tratado como um problema de classificação hierárquica, onde a estrutura hierárquica inclui classes que representam tanto as faixas etárias como os gêneros dos autores dos textos. A proposta dessa abordagem surgiu da hipótese de que o estilo de escrita de um autor é influenciado pela combinação de suas características demográficas e, portanto, uma abordagem que permita explorar a interdependência entre diferentes características poderia colaborar para

a melhoria do desempenho preditivo da tarefa de classificação de faixa etária.

Portanto, o objetivo deste trabalho é avaliar se a utilização da classificação hierárquica envolvendo mais de uma característica demográfica (gênero e a faixa etária) dos autores de textos contribui para melhorar o desempenho preditivo da tarefa de classificação de faixa etária alcançada pela abordagem plana adotada por trabalhos presentes na literatura.

Com o intuito de alcançar o objetivo supramencionado, três bases de dados textuais rotuladas com as características de gênero e faixa etária, previamente utilizadas por outros trabalhos da literatura, foram escolhidas para a realização de uma avaliação comparativa de desempenho preditivo entre a abordagem de classificação plana usada nesses trabalhos e a abordagem hierárquica proposta neste estudo. Os resultados dos experimentos computacionais realizados confirmam o potencial da abordagem hierárquica na resolução da tarefa de predição de faixa etária de autores de textos.

O restante deste artigo está organizado da forma descrita a seguir. A Seção 2 descreve os trabalhos da literatura que também abordaram o problema de predição de faixa etária a partir de textos escritos na língua portuguesa. Em seguida, a Seção 3 apresenta a abordagem hierárquica proposta neste trabalho. Os experimentos realizados para avaliação da abordagem proposta, bem como a discussão dos resultados obtidos, são apresentados na Seção 4. Por fim, a Seção 5 apresenta as conclusões e sugestões de trabalhos futuros.

2. Trabalhos Relacionados

A seguir, são descritos trabalhos da literatura que focaram na predição de faixa etária de autores de textos escritos na língua portuguesa.

[Guimarães et al. 2017] avaliaram diversos classificadores para predição de faixa etária (adolescente e adulto) utilizando uma base de dados construída a partir de características extraídas tanto dos textos como dos perfis dos seus autores na rede social Twitter. O melhor desempenho preditivo foi obtido por uma Rede Neural Convolucional Profunda, que alcançou 0,94 de medida-F.

[Hsieh et al. 2018] avaliaram várias representações textuais para a classificação de diversas características de autores de textos obtidos na rede social Facebook. Utilizando regressão logística, foram analisadas as representações textuais *Bag of Words*, n-gramas de caracteres, TF-IDF, LIWC+P e *Word2Vec*. Para predição de faixa etária, os melhores resultados foram obtidos com TF-IDF, cujo valor de medida-F variou de 0,56 a 0,61 para as três faixas etárias preditas.

Em [Dias 2019], foram avaliados diversos modelos de classificação, incluindo modelos baseados em aprendizado profundo, para diversas tarefas de caracterização autoral utilizando textos de diferentes domínios e idiomas. A classificação de faixas etárias a partir de textos em português foi avaliada a partir de três bases de dados. Para duas dessas bases, os melhores resultados foram obtidos a partir de uma rede neural convolucional usando TF-IDF (macro F1 iguais a 0,45 e 0,62) e, para a terceira base, o melhor desempenho (macro F1 igual a 0,41) foi obtido com um modelo de regressão logística também utilizando TF-IDF.

Assim como em [Guimarães et al. 2017], o trabalho apresentado em [Silva 2020]

utiliza uma base de dados construída a partir de características extraídas tanto dos textos como dos perfis dos seus autores na rede social Twitter para predição de faixa etária considerando apenas duas classes, adolescente e adulto. No entanto, nesse trabalho, os autores aplicaram as técnicas de classificação multirrótulo *Classifier Chains* e *Label Powerset* para predição de gênero e faixa etária. Na predição de faixa etária, o modelo de classificação multidimensional alcançou 0,92 de micro F1, não conseguindo superar os resultados apresentados em [Guimarães et al. 2017].

[Delmondes Neto 2021] desenvolveu modelos de classificação baseados em redes neurais artificiais para a tarefa de caracterização autoral interdomínio, ou seja, onde os modelos treinados são testados em dados de diferentes domínios. Os experimentos envolvendo quatro bases de dados textuais mostraram que, mesmo para o modelo interdomínio que apresentou o melhor resultado (BERT) para a tarefa de classificação de faixa etária, houve uma perda na medida macro F1 em relação aos modelos de domínio único.

Por fim, [Flores et al. 2022] avaliaram diversas tarefas de caracterização autoral a partir de uma coleção de textos referentes a solicitações feitas por cidadãos ao governo brasileiro por meio do sistema e-SIC. Dentre os diversos modelos de classificação avaliados, a LSTM (*Long Short-Term Memory*) apresentou o melhor desempenho para a tarefa de predição de faixa etária, alcançando uma F1 ponderada igual a 0,67.

3. Metodologia

Estudos apresentados na literatura já demonstraram que o estilo de escrita de um autor pode ser usado na identificação de suas características demográficas, tais como gênero, idade, nível de escolaridade e outras. Desse modo, vários trabalhos na literatura avaliaram o uso de classificadores para predição dessas características. A abordagem tradicionalmente adotada por esses trabalhos envolve a classificação plana monorrótulo, ou seja, um classificador é treinado a partir de um conjunto de textos rotulados para uma determinada característica para ser capaz de predizê-la para novas instâncias (textos).

Partindo da hipótese de que explorar a interdependência entre diferentes características do autor de um texto poderia colaborar para a melhoria do desempenho preditivo da tarefa de classificação de faixa etária, a metodologia proposta neste trabalho envolve a utilização de uma abordagem de classificação hierárquica onde a hierarquia de classes é composta tanto por classes relacionadas com as faixas etárias como com o gênero dos autores dos textos.

A hipótese adotada baseia-se na suposição de que o estilo de escrita de uma pessoa do gênero masculino que pertence a uma determinada faixa etária é diferente do estilo de outra pessoa da mesma faixa etária, mas do gênero feminino. Desse modo, a abordagem hierárquica proposta neste estudo propicia o treinamento de modelos de classificação de faixa etária exclusivos para cada gênero, possibilitando que esses modelos capturem as particularidades do estilo de escrita para cada faixa etária associada a cada gênero.

Dentre as abordagens de classificação hierárquica propostas na literatura [Silla and Freitas 2011], este trabalho fará uso da classificação local por nó pai. Isso significa que a resolução do problema de classificação de faixa etária envolverá o treinamento de três classificadores, sendo um para distinção de gênero e outros dois para distinguir entre as faixas etárias, conforme a estrutura hierárquica de classes apresentada

na Figura 1. Nessa figura, as classes estão representadas pelos círculos, enquanto os classificadores locais estão representados pelos retângulos pontilhados.

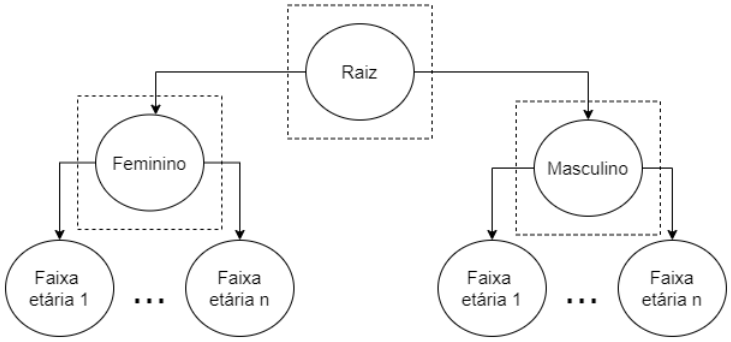

Figura 1. Hierarquia de classes utilizada na abordagem proposta

Desse modo, dado um texto (instância) para ser classificado com relação à faixa etária do seu autor, para decidir qual dos classificadores de faixa etária será utilizado, necessita-se primeiro inferir o gênero do autor do texto. Portanto, utilizando-se o classificador de gênero (nó raiz da hierarquia), se o gênero predito for o masculino, o classificador de faixa etária treinado somente com instâncias do gênero masculino será utilizado na predição da faixa etária. Caso contrário, ou seja, se o gênero predito for o feminino, então o classificador de faixa etária treinado somente com instâncias do gênero feminino será utilizado na predição da faixa etária. Essa abordagem hierárquica também é conhecida na literatura como abordagem *top-down*, uma vez que o resultado da classificação no nó superior da hierarquia (no caso, o classificador de gênero) é utilizado para definir qual classificador do nível inferior (neste caso, os classificadores de faixa etária) será utilizado. Essa estratégia evita o problema de predições inconsistentes, que ocorrem quando as predições realizadas em diferentes níveis hierárquicos não respeitam as restrições da hierarquia de classes. Esse fato justifica a escolha dessa abordagem para avaliação da classificação hierárquica neste trabalho.

A abordagem hierárquica proposta neste trabalho para a predição de faixa etária foi comparada com a abordagem plana monorrótulo utilizada em trabalhos da literatura. Para isso, bases de dados textuais rotuladas com o gênero e a faixa etária dos autores dos textos foram utilizadas.

A Figura 2 ilustra o funcionamento da abordagem por classificação plana, ou seja, dado um texto para ser classificado, ele é pré-processado e vetorizado para ser submetido a um modelo de classificação para a realização da predição.

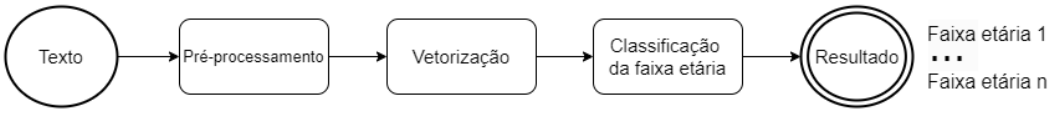

Figura 2. Abordagem de classificação tradicional (plana)

No caso da abordagem por classificação hierárquica, a classificação de um texto segue as etapas ilustradas na Figura 3. Nesse caso, após o pré-processamento do texto,

ele é vetorizado considerando-se o conjunto de textos utilizado no treinamento do classificador de gênero. Vale ressaltar que, para o treinamento do classificador de gênero, são utilizados os textos associados a todas as faixas etárias. Após a vetorização, o texto é submetido ao classificador de gênero e, dependendo da predição desse classificador, o fluxo de processamento segue para o classificador de faixa etária treinado somente com textos associados ao gênero masculino ou para o classificador de faixa etária treinado somente com textos associados ao gênero feminino. Qualquer que seja o fluxo seguido pelo texto, antes de ser submetido ao classificador de faixa etária, a sua versão pré-processada passa novamente pelo processo de vetorização, mas agora considerando somente o conjunto de textos utilizado no treinamento do classificador de faixa etária em questão.

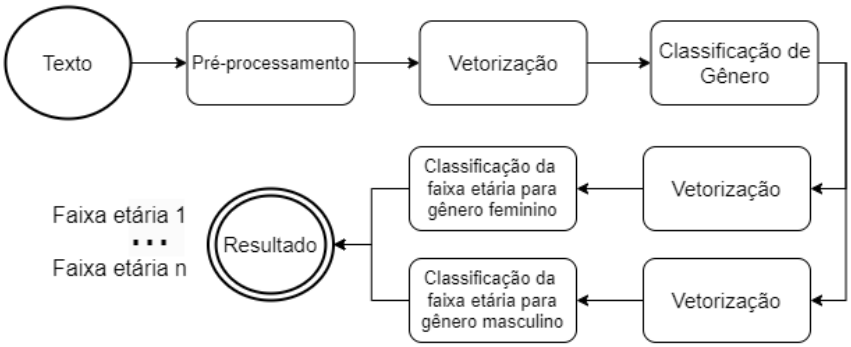

Figura 3. Abordagem de classificação hierárquica proposta

A avaliação das abordagens plana e hierárquica foi realizada utilizando as técnicas *holdout* e validação cruzada *k-fold*. Nos dois casos, o particionamento das bases de dados foi realizado de modo estratificado considerando a combinação dos valores dos atributos gênero e faixa etária, de modo que cada partição de dados ficasse com a mesma proporção de instâncias da base original para cada combinação (gênero, faixa etária). Vale ressaltar que os mesmos pares de partições de treino e teste foram utilizados pelas abordagens plana e hierárquica. Por fim, a métrica de avaliação adotada neste estudo comparativo foi a macro-F1, a mesma utilizada nos trabalhos relacionados apresentados na literatura.

4. Experimentos Computacionais

A avaliação da abordagem hierárquica proposta neste trabalho foi realizada a partir de experimentos comparativos com abordagens planas propostas na literatura em [Hsieh et al. 2018] e [Dias 2019]. Os detalhes dos experimentos realizados serão apresentados da seguinte maneira: a Seção 4.1 descreve as bases de dados utilizadas na avaliação comparativa, a configuração experimental é apresentada na Seção 4.2 e, por fim, na Seção 4.3, são discutidos os resultados obtidos.

4.1. Bases de Dados

Três bases de dados textuais, rotuladas com o gênero e a faixa etária dos autores dos textos, previamente utilizadas em outros trabalhos da literatura, foram selecionadas para a realização dos experimentos. Essas bases de dados, que contêm textos em português do Brasil, possuem características diversas com relação ao seu conteúdo e à origem dos textos (redes sociais, *blogs* etc.). Segue uma descrição de cada uma das bases de dados:

b5-post: base que reúne 516 textos extraídos de postagens de usuários distintos no Facebook, com cada texto reunindo até 1000 postagens de um usuário. Neste trabalho, utilizamos somente o subconjunto dos textos da base original que continha as informações de gênero e faixa etária de seu autor. Criada por [Ramos et al. 2018], essa base faz parte de um corpus com textos informais sobre diversos temas e foi utilizada em [Hsieh et al. 2018]. A Tabela 1 apresenta a distribuição das instâncias de cada faixa etária para cada gênero.

Tabela 1. Distribuição das instâncias na base b5-post

Faixa Etária	# Masculino	# Feminino	# Total
18 - 20	66	116	182
23 - 25	95	94	189
28 - 61	67	78	145
# Total	228	288	516

BRMoral: base que reúne textos opinativos gerados por 510 autores distintos. Os textos abordam temas como religião, política, legalização de drogas, pena de morte e outros. Essa base de dados, disponibilizada por [Santos and Paraboni 2019], foi utilizada em [Dias 2019] com a distribuição das instâncias nas faixas etárias e gêneros apresentada na Tabela 2.

Tabela 2. Distribuição das instâncias na base BRMoral

Faixa Etária	# Masculino	# Feminino	# Total
0 - 23	122	65	187
24 - 30	131	51	182
31 - 99	83	58	141
# Total	336	174	510

BlogSetBR: base formada por 2604 textos de *blogs* publicados por diferentes usuários. Cada texto reúne uma ou mais publicações de um autor. Criada por [dos 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, também utilizada em [Dias 2019], possui a distribuição das instâncias nas faixas etárias e gêneros mostrada na Tabela 3.

Tabela 3. Distribuição das instâncias na base BlogSetBR

Faixa Etária	# Masculino	# Feminino	# Total
10 - 25	394	275	669
26 - 40	596	425	1021
> 40	574	340	914
# Total	1564	1040	2604

4.2. Configuração Experimental

Para se ter uma comparação justa entre a abordagem de classificação hierárquica proposta neste trabalho e a abordagem plana adotada em trabalhos da literatura, ao invés

de simplesmente serem utilizados os resultados reportados nos trabalhos da literatura, as abordagens propostas nos mesmos foram reimplementadas para serem utilizadas como base de comparação (baseline). Desse modo, garantiu-se que as mesmas partições de dados (treino e teste) e o mesmo processo de calibração de hiperparâmetros dos métodos de classificação fossem utilizados pelas duas abordagens.

No processo de reimplementação da abordagem plana, a mesma configuração experimental (pré-processamento, vetorização do texto, classificador e método de avaliação) utilizada nos trabalhos de referência foi adotada nos experimentos conduzidos neste estudo. A Tabela 4 apresenta um resumo da configuração adotada por cada trabalho de referência.

Tabela 4. Configuração experimental dos trabalhos de referência

Trabalho de Referência	Base de Dados	Pré-processamento	Vetorização do texto	Classificador	Método de Avaliação
[Dias 2019]	BlogsetBR	Remoção de stopwords Remoção de tags HTML Transf. em caixa baixa	TF-IDF	CNN	Houldout 80/20 (treino/teste)
[Dias 2019]	BRMoral	Remoção de stopwords Remoção de tags HTML Transf. em caixa baixa	TF-IDF	Regressão Logística	Validação Cruzada com 10 partições
[Hsieh et al. 2018]	b5-post	Remoção de stopwords	TF-IDF	Regressão Logística	Validação Cruzada com 10 partições

Além disso, a mesma configuração experimental utilizada pelos trabalhos de referência foi utilizada na avaliação da abordagem hierárquica. Ainda que a abordagem hierárquica faça uso de três classificadores, um para predição de gênero e dois para predição de faixa etária, o mesmo algoritmo de classificação adotado em cada trabalho de referência é utilizado na criação dos três modelos. Por fim, vale ressaltar que a calibração de hiperparâmetros dos classificadores foi realizada para as duas abordagens envolvidas na comparação, sempre utilizando somente os dados das partições de treino. No caso da avaliação feita a partir da técnica de validação cruzada, o ajuste de hiperparâmetros foi realizado para cada uma das dez partições de treino.

A implementação foi realizada em *Python* e fez uso das bibliotecas *NLTK*, *Scikit-learn* e *TensorFlow*. As bases de dados utilizadas e mais detalhes sobre a configuração experimental encontram-se disponíveis em `https://github.com/ciceribeiroo/hier_class_age_range`.

4.3. Resultados

A Tabela 5 mostra os resultados (macro-F1) alcançados pelas abordagens de classificação plana e hierárquica segundo a configuração experimental descrita na Seção 4.2. Nessa tabela, os resultados reportados pelos artigos de referência são apresentados na coluna 'Reportado'. Em seguida, a coluna 'Baseline' mostra os resultados obtidos a partir da reimplementação da abordagem de classificação plana de cada trabalho de referência. Por fim, os resultados obtidos a partir da abordagem hierárquica proposta neste trabalho são apresentados na coluna 'Abordagem Hierárquica'. Como os resultados das duas últimas colunas são valores médios de dez execuções (validação cruzada com 10 partições ou 10 execuções do *holdout*), é apresentado também o desvio padrão.

Como pode ser observado na Tabela 5, para todas as bases de dados avaliadas, o resultado médio de macro-F1 alcançado a partir da abordagem hierárquica foi superior

Tabela 5. Resultados (macro-F1 médio) das abordagens plana e hierárquica

Base de Dados	Reportado	*Baseline*	Abordagem Hierárquica
BlogsetBR	0,45	0,43±0,01	**0,45±0,02**
BRMoral	0,41	0,45±0,10	**0,47±0,06**
b5-post	0,59	0,56±0,06	**0,60±0,06**

àquele obtido pela abordagem de classificação plana proposta nos trabalhos de referência, sugerindo, portanto, que a hipótese levantada neste trabalho é verdadeira.

Dada a natureza da abordagem hierárquica *top-down* utilizada neste trabalho, a classificação em uma determinada faixa etária é realizada pelo classificador escolhido a partir do resultado da classificação do gênero do autor do texto. Desse modo, o desempenho do primeiro classificador da estrutura hierárquica (classificador de gênero) obviamente influencia no desempenho da classificação de faixa etária. Nesse contexto, é natural esperar que o desempenho preditivo da classificação de faixa etária aumente quanto melhor for o desempenho do classificador de gênero.

Os resultados apresentados na Tabela 6 nos permitem observar como o desempenho do classificador de gênero afeta o desempenho da classificação de faixa etária. Nessa tabela, na segunda e terceira colunas tem-se o desempenho preditivo do classificador de gênero (macro-F1 médio) e o ganho de desempenho (macro-F1 médio) da abordagem hierárquica em relação à abordagem plana, respectivamente. Os resultados mostram que, para a base de dados em que o desempenho da predição de gênero foi maior (b5-post), o ganho apresentado pela abordagem hierárquica também foi superior.

Tabela 6. Resultado da predição de gênero *versus* ganho de desempenho

Base de Dados	Classificador de Gênero	Ganho de Desempenho
BlogsetBR	0,71	0,02
BRMoral	0,62	0,02
b5-post	0,87	0,04

5. Conclusão

No campo da mineração de dados textuais, a caracterização autoral, que tem como foco a inferência de características de autores a partir de seus textos, cumpre um papel relevante para as áreas de computação forense, marketing digital, comércio eletrônico e outras.

Mesmo com os avanços alcançados até o momento na área de caracterização autoral, os resultados reportados na literatura para a predição de faixa etária ainda são modestos e, portanto, mais investimento em pesquisa se faz necessário para melhorar o desempenho preditivo dessa tarefa, principalmente para idiomas como o português, onde a quantidade de trabalhos e de recursos computacionais disponíveis ainda é relativamente pequena quando comparada àquela disponível para outros idiomas.

Portanto, este trabalho contribui nessa área de estudo propondo e avaliando uma abordagem de classificação hierárquica para a tarefa de predição de faixa etária. O objetivo foi avaliar se a utilização da classificação hierárquica envolvendo o gênero dos autores

dos textos poderia contribuir para melhorar o desempenho preditivo da classificação de faixa etária apresentado na literatura para textos escritos na língua portuguesa.

Os experimentos computacionais mostraram que a abordagem hierárquica proposta obteve um ganho, em termos de macro-F1 médio, para as três bases de dados utilizadas na comparação com a abordagem por classificação plana proposta na literatura para predição da faixa etária. Portanto, conclui-se que a utilização dessa abordagem hierárquica tem potencial para obter resultados melhores do que a abordagem de classificação plana.

Algumas possibilidades de trabalhos futuros são apresentadas a seguir. Como existem diferentes abordagens de classificação hierárquica propostas na literatura, uma extensão natural deste trabalho seria avaliar o desempenho da classificação hierárquica com outras abordagens não utilizadas neste estudo, como, por exemplo, com a classificação local por nó ou com a classificação local por nível. Além disso, como essas abordagens locais envolvem o treinamento de múltiplos classificadores, poderiam ser avaliados diferentes algoritmos de classificação em diferentes nós (ou níveis) da estrutura hierárquica. Por fim, uma vez que este trabalho fez uso de um modelo simples de vetorização de textos, o *Bag of Words* com TF-IDF, modelos mais elaborados que capturam contexto, semântica e relações entre palavras poderiam ser avaliados em conjunto com a abordagem de classificação hierárquica.

Agradecimentos

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

Referências

Álvarez-Carmona, M. Á., López-Monroy, A. P., Montes-y-Gómez, M., Pineda, L. V., and Escalante, H. J. (2015). Inaoe's participation at pan'15: Author profiling task. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*. CEUR-WS.org.

Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., and Nissim, M. (2017). N-gram: New groningen author-profiling model. *Conference and Labs of the Evaluation Forum*, abs/1707.03764.

Delmondes Neto, J. P. (2021). Caracterização autoral interdomínio a partir de textos. Master's thesis, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo.

Dias, R. F. S. (2019). Caracterização autoral a partir de textos utilizando redes neurais artificiais. Master's thesis, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, SP, Brasil.

DOMO (2023). Data never sleeps 11.0. `https://www.domo.com/learn/infographic/data-never-sleeps-11`. Acessado em: 11 março de 2025.

dos Santos, H., Woloszyn, V., and Vieira, R. (2018). Blogset-br: A brazilian portuguese blog corpus. In *LREC*, pages 661–664, Miyazaki, Japan. European Language Resources Association (ELRA).

Flores, A. M., Pavan, M. C., and Paraboni, I. (2022). User profiling and satisfaction inference in public information access services. *Journal of Intelligent Information Systems*, 58:67–89.

Goldenberg, D., Kofman, K., Albert, J., Mizrachi, S., Horowitz, A., and Teinemaa, I. (2021). Personalization in practice: Methods and applications. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, page 1123–1126, New York, NY, USA. Association for Computing Machinery.

Guimarães, R. G., Rosa, R., Gaetano, D. D., Rodríguez, D., and Bressan, G. (2017). Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816.

Hsieh, F., Dias, R., and Paraboni, I. (2018). Author profiling from facebook corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2566–2570, Miyazaki, Japan. European Language Resources Association (ELRA).

López-Monroy, A. (2014). Using intra-profile information for author profiling. In *Notebook for PAN at CLEF 2014*. CEUR-WS.org.

López-Monroy, A., Montes, M., Escalante, H. J., Villaseñor-Pineda, L., and Villatoro-Tello, E. (2013). Inaoe's participation at pan'13: Author profiling task: Notebook for pan at clef 2013. *CEUR Workshop Proceedings*, 1179.

Pizarro, J. (2019). Using n-grams to detect bots on twitter. In *Conference and Labs of the Evaluation Forum*, Lugano, Switzerland. CEUR-WS.org.

Ramos, R., Neto, G., Silva, B. B. C., Monteiro, D. S., Paraboni, I., and Dias, R. (2018). Building a corpus for personality-dependent natural language understanding and generation. *LREC 2018*, pages 1138–1145.

Santos, W. and Paraboni, I. (2019). Moral stance recognition and polarity classification from twitter and elicited text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1069–1075, Varna, Bulgaria. INCOMA Ltd.

Silla, C. and Freitas, A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72.

Silva, D. H. (2020). Classificação de gêneros e faixas etárias em redes sociais online por meio de técnicas de aprendizagem multidimensional. Master's thesis, Universidade Federal de Lavras, Lavras, MG, Brasil.

Statista (2025). Internet usage worldwide - statistics facts. `https://www.statista.com/topics/1145/internet-usage-worldwide/#topicOverview`. Acessado em: 11 março de 2025.

Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., and Ohkuma, T. (2018). Text and image synergy with feature cross technique for gender identification: Notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*. CEUR-WS.org.

Vollenbroek, M. B. O., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., and Nissim, M. (2016). Gronup: Groningen user profiling. In *Conference and Labs of the Evaluation Forum*. CEUR-WS.org.