Uma Abordagem Híbrida para Predição de Gênero a partir de Textos em Português

João Pedro M. de Morais[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

joao.morais@estudante.ufla.br,luiz.hcm@ufla.br

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 hybrid approach, which combines a proposed heuristic and a classifier, for the gender prediction problem using only textual content written in the Portuguese language.*

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 híbrida, que combina uma heurística com um classificador, para a predição do gênero do autor de um texto escrito em português utilizando somente o conteúdo textual.*

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. Como a maioria dos trabalhos da literatura se concentra em alguns poucos idiomas amplamente utilizados, como por exemplo o inglês [Hsieh et al., 2018], ainda há um deficit de pesquisas e, consequentemente, recursos e ferramentas computacionais para idiomas como o português.

Desse modo, visando contribuir nessa área de estudo, o presente trabalho apresenta uma abordagem híbrida, que combina uma heurística com 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 heurística proposta neste trabalho realiza a análise morfossintática de um texto a fim de encontrar indícios que permitam inferir o gênero do seu autor. Quando a heurística não consegue inferir o gênero com um certo grau de confiança, então um classificador treinado a partir de um corpus é utilizado para realizar essa tarefa.

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 heurística e a abordagem aqui propostas poderiam contribuir com a melhoria do desempenho preditivo alcançado pelos trabalhos da literatura que foram utilizados como referência. Os resultados experimentais mostraram que a heurística e a abordagem propostas neste trabalho contribuíram para o aumento do desempenho preditivo dos classificadores avaliados, comprovando a eficácia e eficiência das mesmas para o problema em questão.

O restante deste trabalho está organizado da forma descrita a seguir. A Seção 2 apresenta uma breve revisão dos conceitos relacionados com este trabalho. Em seguida, a Seção 3 descreve os trabalhos relacionados da literatura. A Seção 4 detalha a abordagem proposta. Os experimentos computacionais e a análise dos resultados obtidos são apresentados na Seção 5. Por fim, a Seção 6 apresenta as conclusões e direcionamentos para trabalhos futuros.

2. Referencial Teórico

A tarefa de predição de gênero a partir de textos geralmente envolve a utilização de técnicas de Processamento de Linguagem Natural (PLN) para a realização do pré-processamento e da transformação dos textos (representação por meio de vetores numéricos) de modo que eles possam ser processados pelos classificadores. Portanto, a seguir são apresentados tanto as tarefas de PLN adotadas para realizar o pré-processamento e a transformação dos textos quanto os classificadores utilizados neste trabalho.

Tokenização: visa a segmentação de um texto, ou seja, separa a sequência de caracteres de um texto utilizando os espaços, as 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. Enquanto na tokenização

lexical cada palavra corresponde a um *token* do texto, na tokenização sentencial cada *token* é representado por uma sentença. Por exemplo, a tokenização lexical do texto "Olá, que tal tomarmos um café?" produzirá o seguinte conjunto de *tokens*: {[Olá] [,] [que] [tal] [tomarmos] [um] [café] [?]}.

Remoção de *stopwords*: consiste na remoção de palavras irrelevantes para o entendimento do sentido de um texto e que, portanto, não trazem informações relevantes para construção dos modelos preditivos. Pronomes, preposições e artigos são alguns exemplos de *stopwords*.

N-Gramas: um texto pode ser analisado a partir dos seus n-gramas. Um n-grama é uma sequência contígua de n itens obtidos em um texto. Esses itens podem ser, por exemplo, palavras ou caracteres. O tamanho dessa sequência (n) é definido pelo usuário, sendo comum a utilização de $n = 1$ (unigramas), $n = 2$ (bigramas) e $n = 3$ (trigramas). Por exemplo, considerando o texto "Olá, que tal tomarmos um café?" nós podemos extrair o seguinte conjunto de bigramas de caracteres: {[Ol] [lá] [á,] [,q] [qu] [ue] [et] [ta] [al] [lt] [to] [om] [ma] [ar] [rm] [mo] [os] [su] [um] [mc] [ca] [af] [fé] [é?]}.

Representação dos Textos: para que os algoritmos de mineração de dados possam processar textos, faz-se necessária uma representação matemática adequada dos mesmos. Uma forma de representação é por meio de um modelo vetorial, onde um vetor com valores numéricos é utilizado para representar um texto. Neste trabalho, duas formas de vetorização de texto foram utilizadas, a saber: 1) Cada valor numérico do vetor representa a importância de um termo para um texto de uma coleção, sendo a medida *tf-idf* utilizada para quantificar essa importância, 2) Utilização da técnica *Word2Vec* para obtenção de vetores contínuos (*word embeddings*) com dimensões predefinidas para a representação das palavras de um texto e posterior combinação desses vetores de palavras para obtenção do vetor de representação do texto.

Classificadores: Máquina de Vetores de Suporte (*Support Vector Machine* – SVM), Regressão Logística e *Multilayer Perceptron* são três métodos [Witten et al., 2011] comumente utilizados para a classificação de textos. Enquanto o SVM é um classificador linear binário não probabilístico, a Regressão Logística é um método estatístico que permite a construção de modelos que calculam a probabilidade de cada classe do problema a partir do conjunto de valores de atributos que caracterizam uma instância. Já o *Multilayer Perceptron* é uma rede neural composta por três ou mais camadas de neurônios que geralmente é treinada utilizando-se o algoritmo de retropropagação de erro.

3. Trabalhos Relacionados

A seguir são descritos trabalhos recentes da literatura que focaram na predição de gênero para textos escritos na língua portuguesa.

O PAN[1] é uma tradicional competição com foco na estilometria textual que, ao longo das suas edições, tem agregado muitas contribuições a essa área de estudo. A partir da base de dados de *tweets* em português disponibilizada na edição de 2017 do PAN, vários pesquisadores apresentaram propostas para a tarefa de predição de gênero. Basile et al. [2017], vencedores da competição naquele ano, apresentaram uma abordagem que utilizou n-gramas de caracteres (n=3 a n=5) e n-gramas de palavras (n=1 e n=2)

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

como atributos e o esquema de pesos *tf-idf* para produzir a representação dos textos. Para classificação dos textos, o classificador SVM com kernel linear (com parâmetro $C = 0.5$) foi escolhido. A técnica k-validação cruzada (k=*5*) foi utilizada na avaliação do classificador, que alcançou 84,5% de acurácia na predição de gênero. Markov et al. [2017a] utilizaram abordagem similar, trocando os n-gramas convencionais pelos n-gramas tipados propostos por Sapkota et al. [2015], e alcançaram 84% de acurácia. Miura et al. [2017] focaram no uso de Redes Neurais Recorrentes e Redes Neurais Convolucionais e obtiveram 81,2% de acurácia.

Com foco em textos mais longos, o trabalho de Hsieh et al. [2018] utilizou uma parte do *b5-corpus* [Ramos et al., 2018], o qual é composto por postagens em português na rede social Facebook. Como pré-processamento dos textos os autores realizaram a remoção de *stopwords* e a tokenização. Após experimentarem diferentes alternativas, a representação vetorial dos textos que alcançou melhores desempenhos para a predição de gênero foi obtida utilizando-se o *tf-idf* dos 3 mil termos mais frequentes nos textos. Utilizando Regressão Logística (com parâmetro $C = 25.0$) e k-validação cruzada (k=*10*), os autores alcançaram F1 = 88% na predição de gênero.

No trabalho [Dias and Paraboni, 2020] os autores avaliam os resultados da tarefa de predição de gênero a partir de modelos de classificação treinados com textos de um único domínio e com textos de domínios distintos. Utilizando bases de dados textuais de diferentes domínios no idioma português brasileiro (BlogSetBR, BRMoral e E-SIC) os autores observaram 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. Como pré-processamento dos textos os autores realizaram a tokenização. Nesse trabalho, 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. Os *word embeddings* foram obtidos tanto a partir de modelos pré-treinados com 100 e 600 dimensões, como de modelos construídos com os textos utilizados nos experimentos[2]. Com a técnica de avaliação *holdout* (80% para treinamento e 20% para teste), 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.

Por fim, Krüger and Hermann [2019] realizaram uma revisão da literatura sobre a tarefa de predição de gênero considerando os trabalhos publicados entre janeiro de 2017 e janeiro de 2019. Os autores constataram que os melhores resultados dessa tarefa para a língua portuguesa, entre 80% e 85% de acurácia, estão muito aquém daqueles obtidos para a língua inglesa, que alcança uma acurácia de 93,4% em [Markov et al., 2017b], evidenciando a necessidade de melhoria dos resultados para a língua portuguesa.

4. Abordagem Proposta

Assim como pode ser observado nos trabalhos relacionados descritos na Seção 3, a resolução do problema de predição de gênero a partir de texto envolve tipicamente 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. Geralmente as diferenças

[2]Modelos pré-treinados obtidos em [Hartmann et al., 2017]

entre os trabalhos está na escolha das técnicas utilizadas em cada uma dessas três etapas, principalmente nas etapas de vetorização do texto e de treinamento do modelo de classificação.

Ao invés de usar apenas um classificador para predizer o gênero do autor de um texto, neste trabalho propõe-se a utilização de uma abordagem híbrida na qual, além de um classificador, uma heurística é usada para auxiliar na predição do gênero.

As Seções 4.1 e 4.2 apresentam com mais detalhes a abordagem híbrida proposta e o funcionamento da heurística de gênero que a compõe.

4.1. Abordagem Híbrida

A Figura 1 ilustra o funcionamento da abordagem híbrida proposta. Dado um texto para ser classificado, primeiramente esse texto é processado pela heurística, que realiza a análise morfossintática do mesmo a fim de encontrar trechos do texto que indiquem o gênero do seu autor. Se a heurística conseguir definir o gênero com um certo grau de confiança, ele é retornado como o resultado da predição. Caso contrário, ou seja, se a heurística não encontrar qualquer trecho que forneça a indicação do gênero ou se o grau de confiança no gênero definido a partir do processamento do texto estiver abaixo de um limite pré-definido, então o texto original é pré-processado e vetorizado para ser submetido a um modelo de classificação para a realização da predição.

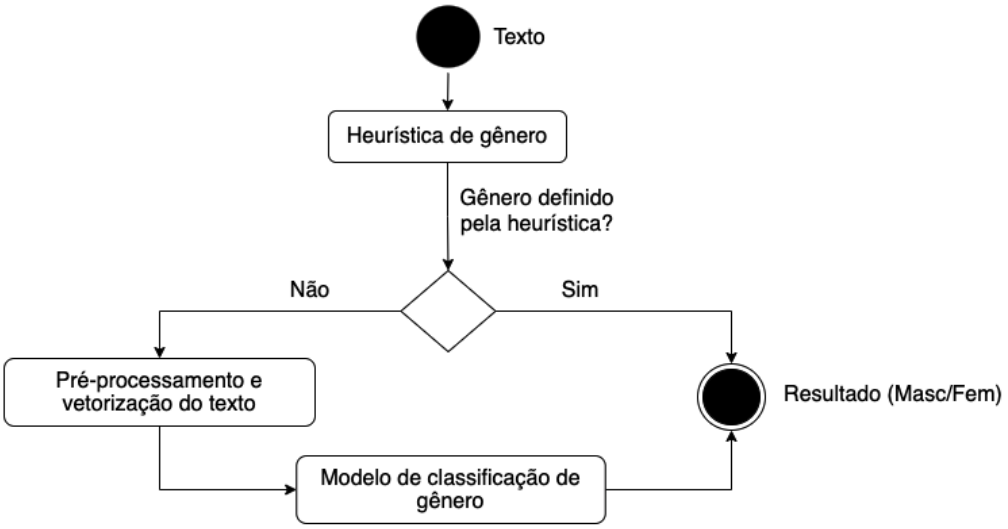

Figura 1. Abordagem híbrida proposta

4.2. Heurística de Gênero

A língua portuguesa é considerada morfologicamente rica, sendo a flexão de gênero muito presente em adjetivos, substantivos, pronomes etc. A heurística proposta neste trabalho explora essa riqueza morfológica da língua portuguesa com a finalidade de encontrar expressões em um texto que indiquem o gênero do seu autor.

O primeiro processamento realizado pela heurística visa a criação de uma estrutura de árvore para cada frase do texto em análise. Essa etapa tem como objetivo capturar as relações sintáticas entre as palavras de cada frase. Além disso, uma análise morfológica

de cada palavra da frase é realizada com o intuito de recuperar informações como a classe gramatical e o gênero (quando couber) da mesma. A Figura 2 ilustra esse processo.

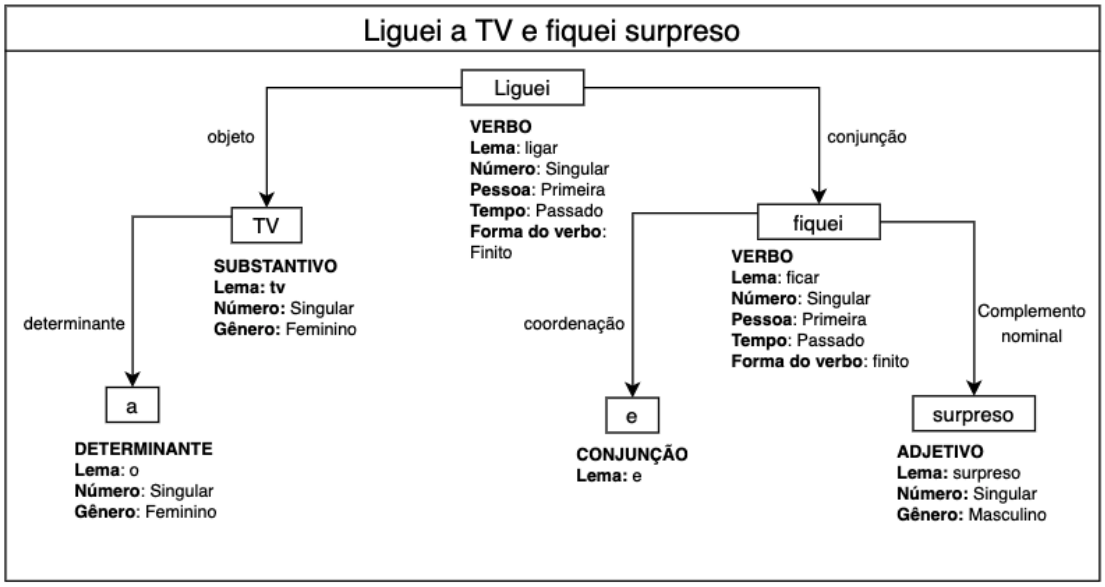

Figura 2. Exemplo de análise morfossintática

Segundo Faraco and Moura [2010], verbos de ligação (por exemplo, ser, estar, permanecer etc.) têm como função estabelecer uma relação entre o sujeito da frase e um termo que expressa características do mesmo. Dado esse fato, numa segunda etapa, a heurística busca por verbos de ligação nas frases em análise com o objetivo de definir o gênero do autor do texto. Desse modo, esse verbo deve estar na primeira pessoa do singular. É importante mencionar que, a fim de desconsiderar trechos do texto que realizam citações a terceiros, frases iniciadas com aspas são ignoradas. 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 é analisar as palavras associadas tanto aos nós filhos como ao nó pai daquele que representa esse verbo de ligação na árvore de dependência sintática. No entanto, dentre as palavras analisadas, somente serão consideradas na etapa seguinte aquelas que atenderem as seguintes condições:

1. A palavra encontra-se após o verbo de ligação na frase.
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 é realizada caso alguma palavra seja identificada na etapa anterior, ou seja, a heurística utiliza o(s) gênero(s) da(s) palavra(s) encontrada(s) para definir o gênero do autor do texto. Quando apenas uma palavra é identificada na terceira etapa, o gênero da mesma é atribuído ao autor do texto. Por exemplo, na frase utilizada na Figura 2, 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.

No caso de termos múltiplas palavras para a definição do gênero do autor do texto, 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 1.

$$\text{Gênero} = \begin{cases} Feminino & \text{se } (f > m) \text{ e } (\frac{f}{f+m} > c) \\ Masculino & \text{se } (m > f) \text{ e } (\frac{m}{m+f} > c) \end{cases} \quad (1)$$

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.

5. Experimentos Computacionais

Um conjunto de experimentos computacionais foi realizado com o objetivo de avaliar a abordagem proposta neste trabalho. 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 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 dos Santos et al. [2018] a partir de mais de 7 milhões de textos coletados da plataforma Blogspot. 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, 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 and 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 et al., 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 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.

5.2. Configuração Experimental

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 (técnicas e algoritmos com seus respectivos parâmetros) foi exatamente a mesma adotada no trabalho utilizado como referência, cujos detalhes foram apresentados na Seção 3.

No caso da abordagem proposta, para a execução da heurística, o único

Tabela 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

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 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[3].

A Tabela 2 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 2. Resumo da configuração experimental

Base de Dados	Referência	Vetorização do Texto	Classificador	Técnica de Avaliação
b5-corpus	[Hsieh et al., 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 and Paraboni, 2020]	Word2Vec	Multilayer Perceptron	Holdout 80%/20%
BRmoral	[Dias and Paraboni, 2020]	Word2Vec	Multilayer Perceptron	Holdout 80%/20%
E-SIC	[Dias and Paraboni, 2020]	Word2Vec	Multilayer Perceptron	Holdout 80%/20%

5.3. Resultados e Discussões

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 execução dos experimentos se devem à aleatoriedade existente

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

no processo de subdivisão das bases para geração dos conjuntos de treinamento e teste dos classificadores.

A Tabela 3 apresenta os resultados obtidos a partir dos experimentos realizados. 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 proposta neste trabalho são mostrados na coluna 'Abordagem Proposta'. 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. Por fim, a coluna 'Cobertura da Heurística' mostra o percentual das instâncias de teste que foram classificadas pela heurística implementada na abordagem proposta.

Tabela 3. Resultados dos experimentos

Base de Dados	Referência	*Baseline*	Abordagem Proposta	Métrica	Cobertura da Heurística
b5-corpus	88,0%	88,5%	**89,6%**	F1	47,64%
PAN-17	84,5%	84,9%	• **88,7%**	Acurácia	55,75%
B2W-Reviews	–	68,3%	• **68,6%**	F1	3,66%
BlogSetBR	78,0%	77,2%	**80,6%**	F1	31,18%
BRmoral	74,0%	**74,5%**	**74,5%**	F1	4,31%
E-SIC	79,0%	78,0%	**78,5%**	F1	5,07%

Antes de analisarmos os resultados obtidos, é importante destacar que a Tabela 3 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 3, 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 ou igual 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*, de modo semelhante, para duas delas a abordagem proposta obteve desempenho superior e para uma o desempenho foi igual 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 da heurística

foi maior, indicando o seu potencial de contribuição na melhoria do desempenho da tarefa de classificação de gênero.

Por fim, outra correlação interessante que pode ser observada a partir desses resultados está relacionada com a quantidade média de palavras por texto e o domínio das bases de dados. As bases de dados com maior quantidade de textos relacionados a questões pessoais do autor (situação tipicamente encontrada em textos de *blogs* e redes sociais) e com maior quantidade média de palavras por texto (BlogSetBR, b5-corpus e PAN-17) foram as que tiveram as maiores coberturas da heurística e os maiores ganhos de desempenho em relação aos baselines.

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 para predição de gênero de autores de textos escritos na língua portuguesa.

Os resultados obtidos a partir da abordagem proposta mostraram que explorar as especificidades da língua portuguesa 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 ou equivalente àquelas apresentadas na literatura.

Como trabalho futuro pretende-se aprimorar a heurística proposta com objetivo de aumentar a cobertura da mesma para bases de dados de diferentes domínios.

AGRADECIMENTOS

Os autores agradecem à Stilingue Inteligência Artificial Ltda pelo apoio financeiro concedido por meio do Acordo de Parceria 006/2020 - UFLA.

Referências

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*, 2017.

Evanildo Bechara. *Moderna Gramática Portuguesa*. Editora Nova Fronteira, 2009.

Rafael Dias and Ivandré Paraboni. Cross-domain author gender classification in brazilian portuguese. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020.

Henrique D. P. dos Santos, Vinicius Woloszyn, and Renata Vieira. BlogSet-BR: A Brazilian Portuguese Blog Corpus. In *11th International Conference on Language Resources and Evaluation*, 2018.

Faraco Carlos Emílio Faraco and Francisco Marto Moura. *Gramática*. 2010.

Yongyan Guo, Jiayong Liu, Wenwu Tang, and Cheng Huang. Exsense: Extract sensitive information from unstructured data. *Computers & Security*, 2021.

Nathan S. Hartmann, Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 2017.

Fernando Hsieh, Rafael Dias, and Ivandré Paraboni. Author profiling from facebook corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

Stefan Krüger and Ben Hermann. Can an online service predict gender? on the state-of-the-art in gender identification from texts. In *IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering*, 2019.

Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. Language-and subtask-dependent feature selection and classifier parameter tuning for author profiling. In *Conference and Labs of the Evaluation Forum*, 2017a.

Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, and Alexander Gelbukh. The winning approach to cross-genre gender identification in russian at rusprofiling. 2017b.

Yasuhide Miura, Tomoki Taniguchi, Motoki Taniguchi, and Tomoko Ohkuma. Author profiling with word+ character neural attention network. In *Conference and Labs of the Evaluation Forum*, 2017.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "how old do you think i am?"a study of language and age in twitter. In *International Conference On Web and Social Media*, 2013.

Ricelli Ramos, Georges Neto, Barbara Silva, Danielle Monteiro, Ivandré Paraboni, and Rafael Dias. Building a corpus for personality-dependent natural language understanding and generation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*, 2017.

Livy Real, Marcio Oshiro, and Alexandre Mafra. B2w-reviews01 an open product reviews corpus. In *XII Symposium in Information and Human Language Technology and Collocates Events*, 2019.

Wesley Santos and Ivandré Paraboni. Moral stance recognition and polarity classification from Twitter and elicited text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2019.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 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*, 2015.

Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Amsterdam, 3 edition, 2011.