DOUGLAS HENRIQUE SILVA

CLASSIFICAÇÃO DE GÊNEROS E FAIXAS ETÁRIAS EM REDES SOCIAIS ONLINE POR MEIO DE TÉCNICAS DE APRENDIZAGEM MULTIDIMENSIONAL

LAVRAS – MG

2020

DOUGLAS HENRIQUE SILVA

CLASSIFICAÇÃO DE GÊNEROS E FAIXAS ETÁRIAS EM REDES SOCIAIS ONLINE POR MEIO DE TÉCNICAS DE APRENDIZAGEM MULTIDIMENSIONAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, área de concentração em Sistemas Inteligentes, para a obtenção do título de Mestre.

Prof. DSc. Demóstenes Zegarra Rodríguez
Orientador
Prof. DSc. Erick Galani Maziero
Coorientador
Profa. DSc. Renata Lopes Rosa
Coorientadora

LAVRAS – MG
2020

Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos da Biblioteca Universitária da UFLA

Silva, Douglas Henrique.

Classificação de gêneros e faixas etárias em redes sociais online por meio de técnicas de aprendizagem multidimensional / Douglas Henrique Silva. – Lavras : UFLA, 2020.

70 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2020.

Orientador: Prof. DSc. Demóstenes Zegarra Rodríguez.

Bibliografia.

1. Classificação de Gêneros e Faixas Etárias. 2. Métodos de Transformação. 3. Aprendizagem Multidimensional. I. Rodríguez, Demóstenes Zegarra. II. Maziero, Erick Galani. III. Rosa, Renata Lopes.

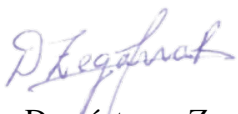
DOUGLAS HENRIQUE SILVA

CLASSIFICAÇÃO DE GÊNEROS E FAIXAS ETÁRIAS EM REDES SOCIAIS ONLINE POR MEIO DE TÉCNICAS DE APRENDIZAGEM MULTIDIMENSIONAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, área de concentração em Sistemas Inteligentes, para a obtenção do título de Mestre.

APROVADO em 5 de Outubro de 2020.

Prof. DSc. Demóstenes Zegarra Rodríguez	UFLA
Prof. DSc. Wilian Soares Lacerda	UFLA
Prof. DSc. Jadson Castro Gertrudes	UFOP



Prof. DSc. Demóstenes Zegarra Rodríguez
Orientador

LAVRAS – MG
2020

Dedico este trabalho aos meus pais Luiz Henrique da Silva e Roseli Rêgo da Silva, à minha irmã Ana Luiza Silva e à minha noiva Talita Rodrigues de Souza.

AGRADECIMENTOS

Agradeço a Deus, que permitiu e concedeu-me a realização deste trabalho.

Agradeço aos meus pais e a minha irmã, pelo amor, apoio, carinho, compreensão, incentivo e paciência nos momentos difíceis.

Agradeço a minha noiva, pela amizade, companheirismo e amor incondicionais.

Agradeço aos professores Demóstenes Zegarra Rodríguez e Renata Lopes Rosa, pela dedicação, paciência, ensinamentos, orientações, pesquisas e revisões, durante a minha trajetória no mestrado.

Agradeço ao professor Erick Galani Maziero, pela atenção, apoio técnico, ensinamentos e revisões, no desenvolvimento deste trabalho.

Agradeço aos meus amigos íntimos, pelo apoio e incentivo nos momentos difíceis.

Agradeço aos amigos, funcionários e professores do Programa de Pós-graduação em Engenharia de Sistemas e Automação (PPGESISA), e Departamento de Ciência da Computação (DCC), pelo aprendizado e convívio. Em especial a Vânia Batista dos Santos, por dividir o conhecimento comigo.

Agradeço à Universidade Federal de Lavras (UFLA), pela infraestrutura oferecida e por estimular a interação e participação nas atividades acadêmicas.

Agradeço à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo financiamento da bolsa do Programa de Apoio à Pós-Graduação – Engenharia de Sistemas e Automação.

Muito obrigado!

RESUMO

Devido ao grande volume de conteúdo gerado por usuários nas Redes Sociais Online (RSO), as organizações têm aplicado técnicas de análise de sentimento ou de mineração de opinião para obter informações sobre pessoas ou entidades de interesse. Uma entidade pode ser produtos, serviços, pessoas, instituições governamentais e não-governamentais, políticas públicas, entre outros tipos. A classificação de gêneros e faixas etárias dá suporte a análise de sentimento e de opinião, pois auxiliam na obtenção de um sentimento ou uma opinião mais precisa. Entretanto, informações sobre perfis de usuários podem estar ocultas ou preenchidas erroneamente nas RSO. Na literatura, várias abordagens são utilizadas com o intuito de efetuar a classificação de gêneros e faixas etárias. Porém, neste trabalho um novo conjunto de características é utilizado, por meio de uma aprendizagem multidimensional. Assim, o objetivo principal deste trabalho é desenvolver um novo modelo de classificação de gêneros e faixas etárias com dados extraídos da RSO *Twitter*, usando os métodos de transformação *Classifier Chains (CC)* e *Label Powerset (LP)*, e através de técnicas de aprendizagem de máquina baseadas em regras, álgebra linear e probabilidade. Este estudo trabalha com uma nova base de dados contendo 8000 instâncias extraídas do *Twitter*. Os melhores subconjuntos de características de perfis de usuários são avaliados, assim como os modelos de aprendizagem multidimensional utilizando diferentes métricas de desempenho. Por meio dos experimentos, obteve-se um modelo de classificação multidimensional na fase de teste, com 0,999 de micro-média F1 para gêneros e 0,923 para faixas etárias. Os resultados da classificação de gêneros superou a maioria dos trabalhos relacionados, e o desempenho da classificação de faixas etárias é bastante competitivo.

Palavras-chave: Classificação de Gêneros. Classificação de Faixas etárias. Métodos de Transformação. Aprendizagem Multidimensional. Classificação Multidimensional.

ABSTRACT

Due to the large volume of content generated by users on Online Social Networks (OSN), organizations have applied sentiment analysis or opinion mining techniques to obtain information about people or entities of interest. An entity can be products, services, people, governmental and non-governmental institutions, public policies, among other types. The classification of genders and age groups supports the analysis of sentiment and opinion, as they help to obtain a more precise feeling or opinion. However, information about gender and age-group may be hidden or incorrectly filled out in the OSN. In the literature, several approaches are used in order to classify genders and age groups. However, in this work, a new set of features is used to classify genders, and age groups, through multidimensional learning. Thus, the main objective of this work is to develop a new model of classification of genders and age groups with data extracted from OSN Twitter, using the transformation methods Classifier Chains (CC) and Label Powerset (LP), and through machine learning techniques based on rules, linear algebra, and probability. This study works with a new database containing 8000 instances extracted from Twitter. The best subsets of user profile features are evaluated, as well as multidimensional learning models using different performance metrics. Through the experiments, a multidimensional classification model was obtained in the test phase, with 0.999 of F1 micro-average for genders and 0.923 for age groups. The results of the classification of genders surpassed most of the related works, and the performance of the classification of age groups is quite competitive.

Keywords: Gender Classification. Age-group Classification. Transformation Methods. Multidimensional Learning. Multidimensional Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e motivação

As organizações têm utilizado o conteúdo das mídias sociais para obter informações sobre pessoas ou demais entidades de interesse. Uma entidade pode ser produtos, serviços, pessoas, instituições governamentais e não-governamentais, políticas públicas, ou de um outro tipo. Atualmente, o grande volume de conteúdo gerado pelas pessoas tem sido utilizado por técnicas automáticas de análise de sentimentos e mineração de opinião, para analisar e avaliar tais informações. No entanto, a presença de opiniões em frases curtas com vários idiomas, gírias, ironia, sarcasmo, emojis, *emoticons* e ausência da forma culta do idioma, tem dificultado a identificação de opiniões corretas dos usuários, afetando diretamente os resultados de análise. (LIU, 2012; GUIMARÃES et al., 2017; ASGHAR et al., 2018).

Uma possível solução para tratar a dificuldade da identificação de opiniões é integrar as informações do perfil do usuário na análise de sentimentos. De acordo com Park et al. (2016), as mulheres são mais calorosas, amigáveis e focadas em pessoas nas mídias sociais. Em Schwartz et al. (2013), foi constatado que adolescentes usam mais gírias e *emoticons*. Segundo Nguyen et al. (2013), verificou-se que a idade pode influenciar diretamente o sentimento final de uma frase. Logo, as características de escrita e o modo de expor os sentimentos são distintos entre gêneros e idades. Assim, tais estudos concluem que as informações de perfis de usuário como gênero e idade podem reduzir a dificuldade de identificação de opiniões (GUIMARÃES et al., 2017).

No entanto, nem sempre é possível obter dados de perfis de usuário, pois eles podem estar ocultos (LI; LI; JI, 2018) ou preenchidos erroneamente. Nesse cenário, a análise de autoria, que é o estudo das características linguísticas e computacionais de documentos escritos por indivíduos, é aplicada para classificar automaticamente faixas etárias, gêneros, etnias e níveis de escolaridade. Na literatura, em função do conteúdo analisado, três abordagens trabalham a caracterização de gêneros e faixas etárias: classificação baseada em conteúdo textual; classificação baseada em características de páginas dos perfis de usuários; e híbrida. A maioria dos trabalhos encontrados na literatura, aplica a classificação baseada em conteúdo textual, mas a classificação baseada em características de páginas dos perfis de usuários tem alcançado excelentes resultados na classificação de gêneros e faixas etárias.

Nesse contexto, em Guimarães et al. (2017) foram propostos modelos baseados em dados de páginas dos perfis de usuários do *Twitter* para classificar as faixas etárias: adolescente e adulto. Os modelos de classificação das técnicas *Multilayer Perceptron*, Árvore da Decisão, *Random Forest*, *Support Vector Machine (SVM)* e *Convolutional Neural Network (CNN)* são aplicados nos experimentos. O melhor resultado de medida F foi obtido com o CNN, após eliminação de características irrelevantes, alcançando valores de 0,940 na fase de validação.

Em Kiratsa et al. (2018), os perfis de usuários do *Facebook* foram analisados, com o objetivo de classificar o gênero do proprietário de uma frase. O objetivo foi atingido por meio da construção de modelos de classificação, baseados em características das seguintes interações dos usuários: a quantidade de curtidas de cada assunto relacionado e os subassuntos curtidos de cada assunto. O melhor desempenho alcançado foi de 0,973 de acurácia e 0,974 de medida F. Tais valores foram obtidos pelo classificador Árvore de Decisão *Adaboost*, após a seleção de quatro e sete características da primeira e segunda interações, respetivamente.

Diferentemente dos trabalhos acima, este trabalho propõe o aprendizado multidimensional, aplicando as técnicas *Classifier Chains* (CC) e *Label Powerset* (LP), para classificar os gêneros e as faixas etárias de usuários da Web. A classificação multidimensional é o aprendizado supervisionado em que cada instância é associada a um subconjunto de classes (READ; MARTINO; LUENGO, 2014).

Em Marquardt et al. (2014), foram utilizadas as classificações de gêneros e faixas etárias utilizando as técnicas CC e LP. Entretanto, os modelos de classificação não alcançaram uma alta acurácia para a Rede Social Online (RSO) *Twitter*. Os resultados de CC foram 0,334 de acurácia em inglês e 0,316 de acurácia em espanhol. Os resultados de LP foram 0,327 de acurácia em inglês e 0,337 de acurácia em espanhol.

Portanto, o aprimoramento da classificação simultânea de gêneros e faixas etárias usando características de perfis de usuários, pode auxiliar na obtenção de sentimentos mais precisos nas análises de sentimento e afetividade. Além disso, soluciona o problema de elaboração de perfis de autoria, que busca identificar gêneros e faixas etárias dos autores de RSO.

1.2 Objetivo

O objetivo desta dissertação é aprimorar o desempenho da classificação multidimensional de gêneros e faixas etárias de usuários do *Twitter*, utilizando os métodos de transformação

Classifier Chains (CC) e *Label Powerset (LP)*. Para isso, as características de perfis de usuários extraídas da RSO *Twitter* são utilizadas na formação dos modelos de classificação de gêneros e faixas etárias. Tais características de perfis de usuários incluem o uso de responder *tweets* (*retweets*), ícones de emoção, gírias, *Uniform Resource Locator (URL)* para compartilhar informações da mídia, o número de caracteres de cada mensagem, o número de pessoas que o usuário segue, o número de seguidores que o usuário possui, o número total de *tweets* postados na RSO e o assunto abordado em cada mensagem.

1.2.1 Objetivos específicos

Os objetivos específicos são:

- Construir um conjunto de dados para classificar gêneros e faixas etárias, contendo 8000 instâncias de determinados assuntos, com dados extraídos da RSO *Twitter*. Tais assuntos abordam: convênio médico, alimentação saudável, carboidrato, novena religiosa, trabalho profissional e mãe;

- Definir os melhores subconjuntos de características de perfis de usuários da RSO *Twitter*, através de métricas de avaliação;

- Avaliar os modelos de aprendizagem multidimensional através de diferentes métricas de desempenho: Precisão, Revocação, medida F1, macro-média F1, micro-média F1 e *Hamming Loss*, para obter uma análise imparcial dos resultados. Além disso, os melhores modelos de aprendizagem multidimensional serão analisados por meio de métricas de avaliação e de testes estatísticos não paramétricos: *Friedman Aligned Ranks* e *Finner*.

1.3 Organização do texto

Este trabalho tem a seguinte estrutura. Capítulo 2 contém a descrição dos trabalhos e tópicos relacionados ao assunto tema deste trabalho. Capítulo 3 contém a descrição de métodos e técnicas da metodologia proposta. O capítulo 4 apresenta a descrição da avaliação experimental, dos resultados e das análises dos modelos de classificação. Por fim, o capítulo 5 apresenta a conclusão com as considerações finais e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo abordará os tópicos de estudo que compõem este trabalho e apresentará os trabalhos relacionados. O primeiro tópico (2.1), aborda a caracterização de gêneros e faixas etárias. O segundo (2.2), o contexto para selecionar as características mais relevantes do conjunto de dados. O terceiro (2.3), a classificação de gêneros e faixas etárias através do aprendizado de máquina. O quarto (2.4), as métricas de classificação e o método de avaliação dos modelos. O último, os trabalhos relacionados com este trabalho.

2.1 Caracterização de gêneros e faixas etárias

Neste trabalho, a classificação de gêneros e faixas etárias está relacionada à análise da autoria sobre a perspectiva de caracterização. Conforme Alhijawi, Hriez e Awajan (2018), a análise de autoria é o estudo das características linguísticas e computacionais dos documentos escritos por indivíduos. A caracterização da autoria busca coletar características demográficas, como sexo, idade e nível educacional de um autor. Na literatura, dependendo do conteúdo analisado, três abordagens resolvem a caracterização de gêneros e faixas etárias: classificação baseada em conteúdo textual, classificação baseada em características de páginas dos perfis de usuários, e híbrida, que utiliza ambas abordagens como em (ALSUKHNI; ALEQUR, 2016).

A classificação baseado no conteúdo textual utiliza técnicas de análise de texto para encontrar características e padrões de escrita existentes, que ajudem a classificar gêneros ou faixas etárias dos usuários. Em outras palavras, o objetivo é extrair características e encontrar padrões de escrita existentes usando dados não estruturados para classificar gêneros ou faixas etárias dos usuários. Por exemplo, Park et al. (2016) cita um padrão encontrado no *Facebook*, no qual as mulheres usam mais pronomes pessoais, advérbios de intensidade, palavras de emoção e têm maior probabilidade de discutir a vida social e familiar. Nesse contexto, abordagens baseadas em conteúdo textual aplicam recursos sintáticos, linguísticos e lexicais de palavras no aprendizado de máquina (evidenciado na seção 2.6.1).

A classificação baseada em características de páginas dos perfis de usuários utiliza um conjunto de dados estruturado, que consiste em instâncias e atributos com seus respectivos valores e tipos. De acordo com o tipo de dados, as instâncias são pré-processadas, com técnicas de análise de texto, técnicas estatísticas e conversores de dados, antes das técnicas de classificação. Como exemplo, em Guimarães et al. (2017) foram extraídas treze características para distinguir as faixas etárias, adolescente e adulto. As características mais relevantes foram: gê-

nero; assunto; usos de gíria e link direcionado; e a quantidade de caracteres em cada mensagem. Segundo os autores, o uso de gírias é mais comum na faixa etária adolescente (SCHWARTZ et al., 2013), enquanto o uso de URL é mais comum na faixa etária adulto. O assunto e o número de caracteres são específicos para cada faixa etária. Além disso, conforme Kiratsa et al. (2018), características baseadas em assuntos foram usadas na classificação de gêneros.

Na literatura, são encontradas outras características de páginas dos perfis de usuários. De acordo com Alowibdi, Buy e Yu (2013b), foram extraídos do *Twitter* e aplicados para classificar os gêneros: nome do perfil, nome do usuário, cor do plano de fundo do perfil, cor do texto, cor do link, cor do preenchimento da barra lateral e cor da borda da barra lateral. Também em (ALOWIBDI; BUY; YU, 2013a), os cinco recursos baseados em cores extraídos do *Twitter* foram aplicados para classificar os gêneros.

Nesse contexto, este trabalho baseia-se em características de páginas dos perfis de usuário, pois visa classificar os gêneros e faixas etárias usando características extraídas das páginas de perfis dos usuários do *Twitter*.

2.2 Seleção de características

Como em (GUIMARÃES et al., 2017), este trabalho propõe a identificação de um subconjunto de características relevantes para as previsões de gêneros e faixas etárias de usuários. Essa tarefa é conhecida como seleção de características e procura avaliar o desempenho de um subconjunto de atributos selecionados de acordo com uma função de avaliação. O objetivo é reduzir a dimensionalidade do conjunto de dados, removendo atributos redundantes e irrelevantes. A seleção de características abrange três abordagens: filtro, *wrapper* e embutida.

As abordagens embutidas (internas) pertencem ao processamento de alguns algoritmos de classificação; isto é, elas são uma etapa do processamento desses algoritmos. Assim, o algoritmo de classificação define qual subconjunto de atributos usar e qual subconjunto ignorar. Essas abordagens são geralmente aplicadas para construir árvores de decisão (TAN; STEINBACH; KUMAR, 2005).

As abordagens de *wrapper* usam um algoritmo de classificação de dados de interesse para selecionar o melhor subconjunto de atributos do conjunto de dados. Geralmente, um algoritmo heurístico seleciona um subconjunto de atributos e o algoritmo de classificação cria um modelo. Em seguida, uma função de avaliação de subconjuntos avalia o modelo. Esse pro-

cesso se repete até atender a um critério de parada. O modelo que obtém o melhor desempenho tem seu conjunto de atributos retornado (TAN; STEINBACH; KUMAR, 2005). Uma desvantagem da abordagem wrapper é o tempo computacional. O tempo de execução dessa abordagem, depende muito do algoritmo de classificação de interesse. Uma desvantagem da abordagem *wrapper* é o tempo computacional. O tempo de execução dessa abordagem é dependente do tempo de treinamento algoritmo de classificação de interesse.

As abordagens de filtro aplicam uma abordagem independente do algoritmo de classificação, para selecionar um subconjunto de atributos do conjunto de dados. Essa abordagem filtra um subconjunto de atributos, de acordo com um critério de avaliação (TAN; STEINBACH; KUMAR, 2005).

Nesse contexto, a técnica aplicada para fazer a seleção de características pertence à abordagem filtro, pois não depende de algoritmo de classificação para definir os melhores atributos.

2.2.1 Critério de Avaliação

A medida estatística qui-quadrado (PEARSON, 1900) foi escolhida como critério de avaliação da técnica baseada em filtro, para medir a dependência de cada característica em relação às classes. Essa medida compara as frequências observadas e esperadas das amostras de uma classe. As frequências obtidas a partir dos dados da amostra são as observadas e as frequências calculadas a partir delas são as esperadas (ADENIRAN; JADAH; MOHAMMED, 2020).

Os valores das frequências da equação 2.1 estão contidas em uma tabela de contingência com n linhas e m colunas. Para cada célula da tabela de contingência, o quadrado da diferença entre as frequências, obtidas (o_i) e esperadas (e_i), é calculado e dividido pela frequência esperada (e_i). A soma dos valores de cada divisão é o valor do qui-quadrado (X^2).

$$X^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{2.1}$$

Como exemplo, uma pequena amostra contendo informações de gêneros e relacionadas ao uso de *hashtag* foi extraída do conjunto de dados para calcular o qui-quadrado. As frequências observadas e esperadas da Tabela 2.1 foram aplicadas na equação 2.1. O valor estimado de X^2 é aproximadamente $3,77$.

Tabela 2.1 – Tabela de contigência

Gêneros	**Uso de *hashtag***		**Total**
	Sim	**Não**	
Femenino	25 (19,45)	324 (329,55)	349
Masculino	10 (15,55)	269 (263,45)	279
Total	35	593	628

As frequências esperadas estão entre parenteses.

Fonte: Autor (2020)

2.3 Algoritmos e técnicas de aprendizado de máquina

O aprendizado de máquina é uma área da inteligência artificial, cujo objetivo é desenvolver algoritmos que aprendem padrões presentes em dados de entrada. Esses algoritmos têm uma fase de aprendizado, na qual é gerado um modelo ou função que define padrões em dados.Três abordagens de aprendizado são encontrados na literatura: supervisionado, não supervisionado e semissupervisionado (BAEZA-YATES; RIBEIRO-NETO, 2013).

Neste trabalho, a caracterização da autoria é um problema de classificação e uma tarefa supervisionada de aprendizado de máquina. O aprendizado supervisionado usa informações anteriores de exemplos de treinamento, fornecidos por seres humanos ou adquiridos através de assistência humana, como dados de entrada. Assim, na tarefa de classificação, o aprendizado ocorre através de um conjunto de classes e exemplos de instâncias para cada classe. As classes com suas instâncias são determinadas em um processo de rotulagem por especialistas humanos e constituem um conjunto de treinamento. Por fim, esse conjunto é usado por uma função, para aprender a classificar novas instâncias (BAEZA-YATES; RIBEIRO-NETO, 2013).

Dois paradigmas de classificação foram encontrados na literatura para caracterizar gêneros e faixas etárias: classificação de rótulo único e classificação multirrótulo. A classificação de rótulo único representa a abordagem supervisionada tradicional, na qual cada instância pertence a uma classe. Por outro lado, a classificação multirrótulo associa cada instância a um conjunto de rótulos (ZHANG; ZHOU, 2014). De acordo com Read, Martino e Luengo (2014), a classificação multirrótulo pode ser considerada um problema multidimensional específico para classes binárias.

A classificação de rótulo único é amplamente aplicada na classificação de gêneros e faixas etárias (evidenciado na seção 2.6). A classificação multirrótulo foi encontrada em Marquardt et al. (2014), cujo objetivo era classificar simultaneamente gênero e faixa etária. Os

autores aplicaram o algoritmo SVM com dois métodos de transformação de problemas de rótulos múltiplos: LP e CC.

Os dois métodos transformam problemas multirrótulo ou multidimensional em cenários de aprendizagem tradicionais, e consideram uma correlação entre rótulos múltiplos. O LP transforma um problema de classificação multirrótulo ou multidimensional em um único problema de classificação multi-classe (ZHANG; ZHOU, 2014). O CC usa classificadores de rótulo único em cadeia, cuja saída de um classificador aplicado a um rótulo é usada como entrada para o classificador de outro rótulo (ZHANG; ZHOU, 2014). Esses métodos foram aplicados com as seguintes técnicas de classificação:

- Árvore de Decisão: é uma técnica de classificação, em que o conjunto de dados de treinamento é usado para criar regras de classificação. As regras são organizadas em ramificações de uma árvore. Cada ramo da raiz à folha é uma regra e elas classificam as novas instâncias (BAEZA-YATES; RIBEIRO-NETO, 2013).

- *Random Forest*: é um comitê de classificadores de árvores de decisão. Assim, têm como base, regras de classificação. Essa técnica combina previsões de várias árvores de decisão. A construção de cada árvore ocorre a partir das características presentes no conjunto de dados, distribuídas em vetores por uma probabilidade fixa (TAN; STEINBACH; KUMAR, 2005).

- *Gradient Tree Boosting*: é um comitê de classificadores de árvores de decisão (CHEN; GUESTRIN, 2016). Também, têm como base, regras de classificação. Uma técnica baseada em *Boosting* combina n modelos com base na ideia de que eles se complementam e valoriza a contribuição de cada modelo de acordo com sua confiabilidade. A implementação prática dessa técnica ocorre através de um processo interativo, onde um novo modelo é criado com base no desempenho do modelo anterior. Assim, cada novo modelo é um especialista, atribuindo pesos mais altos aos casos analisados incorretamente nos modelos anteriores (WITTEN; FRANK; HALL, 2011).

- *Support Vector Machine (SVM)*: é uma técnica de classificação de vetores binários em um espaço n-dimensional. O objetivo é encontrar o melhor hiperplano de decisão (H_x), que separa as instâncias em duas classes distintas. Ao executar a classificação de dados,

cada instância X é um vetor no espaço, onde $X = \{x_1, x_2, \cdots, x_n\}$, x_n é o valor do enésimo dado na instância, e n é o número de atributos no conjunto de dados (BAEZA-YATES; RIBEIRO-NETO, 2013).

- *Complement Naïve Bayes*: é uma técnica de classificação probabilística. Essa técnica é um aprimoramento da técnica *Multinomial Naïve Bayes (MNB)*. Diferentemente do MNB, um cálculo estatístico do complemento da classe define os pesos dos recursos do modelo (RENNIE et al., 2003). Os classificadores probabilísticos atribuem a cada par de classe e instância $[X_j, c_p]$, uma probabilidade $P(c_p|X_j)$ de que X_j pertença à classe c_p (BAEZA-YATES; RIBEIRO-NETO, 2013). O classificador atribui X_j a classe c_p mais provável.

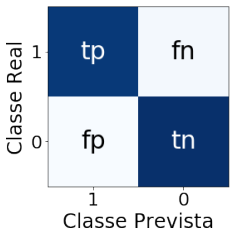
2.4 Medidas de classificação

Os desempenhos dos modelos de classificação por gênero e faixa etária foram computados por matrizes de confusão, medidas tradicionais, medidas baseadas em rótulos e pela *Hamming Loss* baseada em exemplo. As medidas baseadas em rótulos e tradicionais são as médias micro ou macro das respectivas métricas: precisão, revocação e medida F1. Essas medidas têm uma baixa correlação de Pearson com a *Hamming Loss* (PEREIRA et al., 2018).

2.4.1 Matriz de confusão

A Figura 2.1 exemplifica uma matriz de confusão. Os valores dos verdadeiros positivos (*true positives - tp*) e verdadeiros negativos (*true negatives - tn*), quantificam o número total de respostas corretas do classificador. Os valores falsos positivos (*false positives - fp*) e falsos negativos (*false negatives - fn*), quantificam os erros totais do classificador.

Figura 2.1 – Matriz de confusão



Fonte: Adaptado de Hunter (2007)

2.4.2 Medida baseada em exemplo

A *Hamming Loss* é uma média da diferença entre os valores reais e esperados dos rótulos e está definida na equação 2.2 (ASIM; REHMAN; SHOAIB, 2017). Portanto, q_i é um conjunto de rótulos fornecido por um classificador, y_i é o conjunto de rótulos verdadeiros, N é o número de instâncias e L é o tamanho do conjunto de rótulos.

$$Hamming\ Loss_{(q_i,\ y_i)} = \frac{1}{N}\sum_{i=1}^{N}\frac{xor(q_i,\ y_i)}{L} \tag{2.2}$$

2.4.3 Medidas baseadas em rótulos

As medidas baseadas em rótulos são obtidas a partir da adaptação das medidas tradicionais de aprendizado de máquina e seguem a notação proposta em (ASIM; REHMAN; SHOAIB, 2017). Assim, os dados da matriz de confusão para cada rótulo e o número total de possíveis conjuntos de rótulos (Q) foram utilizados neste trabalho para calcular as médias micro e macro de F1, definidas em (2.3) e (2.4)

A macro-média F1 (macro-μ F1) calcula a média de F1 em todos os rótulos em (2.3).

$$F1_{\ macro-\mu} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{2\,tp}{2\,tp+fp+fn} \tag{2.3}$$

A micro-média F1 (micro-μ F1) usa os valores da revocação e precisão das micro-médias para calcular seu valor em (2.4) (BAEZA-YATES; RIBEIRO-NETO, 2013).

$$F1_{\ micro-\mu} = \frac{\sum_{i=1}^{Q}2\,tp}{\sum_{i=1}^{Q}2\,tp+\sum_{i=1}^{Q}fp+\sum_{i=1}^{Q}fn} \tag{2.4}$$

2.4.4 Medidas tradicionais

Na avaliação individual de cada rótulo foram usadas as métricas tradicionais precisão (P), revocação (R) e F1 apresentadas nas equações 2.5, 2.6, e 2.7. As micro-médias de P, R e F1 estão representadas nas equações 2.8, 2.9, e 2.10.

A precisão é a fração das instâncias classificadas corretamente em uma classe (c), dividida por todas as instâncias classificadas em c, de acordo com o conjunto de testes (BAEZA-YATES; RIBEIRO-NETO, 2013).

$$P = \frac{tp}{tp+fp} \tag{2.5}$$

A revocação é a fração das instâncias classificadas corretamente em uma classe (c), dividida por todas as instâncias reais de c, de acordo com o conjunto de testes (BAEZA-YATES; RIBEIRO-NETO, 2013).

$$R = \frac{tp}{tp+fn} \tag{2.6}$$

A métrica F1 combina as equações de precisão e revocação, equilibrando a importância relativa de cada métrica (BAEZA-YATES; RIBEIRO-NETO, 2013).

$$F1 = \frac{2tp}{2tp+fn+fp} \tag{2.7}$$

A micro-média da precisão (micro-μ P) em (2.8) é a fração da soma de todas as instâncias classificadas corretamente em todas as classes de C, cujo denominador é a soma de todas as instâncias atribuídas às classes de C pelo classificador (BAEZA-YATES; RIBEIRO-NETO, 2013).

$$P_{\,micro-\mu} = \frac{\sum_{i=1}^{C} tp}{\sum_{i=1}^{C} (tp+fp)} \tag{2.8}$$

A micro-média da revocação (micro-μ R) em (2.9) é a fração de todas as instâncias classificadas corretamente em todas as classes de C, cujo denominador é a soma de todas as instâncias pertencentes às classes de C (BAEZA-YATES; RIBEIRO-NETO, 2013).

$$R_{\,micro-\mu} = \frac{\sum_{i=1}^{C} tp}{\sum_{i=1}^{C} (tp+fn)} \tag{2.9}$$

A micro-média F1 (micro-μ F1) definida em (2.4) é aplicada como tradicional, substituindo o número total de rótulos Q pelo número total de classes C em 2.10.

$$F1_{\,micro-\mu} = \frac{\sum_{i=1}^{C} 2\,tp}{\sum_{i=1}^{C} 2\,tp + \sum_{i=1}^{C} fp + \sum_{i=1}^{C} fn} \tag{2.10}$$

2.5 Métodos de avaliação de modelos

A validação cruzada avalia a generalização dos modelos de classificação. Neste trabalho, foi utilizada a validação cruzada estratificada, que garante a mesma distribuição de classes nos conjuntos avaliados. Essa validação divide o conjunto de dados em subconjuntos f. Cada

subconjunto é aplicado como conjunto de testes uma vez e $f-1$ vezes como conjuntos de treinamentos. A avaliação dos f classificadores ocorre de forma independente, usando medidas de avaliação baseadas em rótulos e exemplos, que são estimadas através de medidas de avaliação (BAEZA-YATES; RIBEIRO-NETO, 2013).

2.6 Trabalhos relacionados

Esta seção apresenta os trabalhos relacionados a classificação de gêneros e faixas etárias, encontrados na literatura, por meio da aprendizagem de máquina. A seção está dividida em função do conteúdo: textual; características de páginas dos perfis de usuários; e híbrida, com os dois primeiros conteúdos.

2.6.1 Classificação de gêneros e faixas etárias através de conteúdo textual

Tam e Martell (2009) apresentaram os resultados da aplicação de uma análise estatística e categorização de textos para classificar as faixas etárias de um autor, através de suas publicações em um *chat*. Foi implementado um método que gera *stopwords* e que escolhe *n-gram* com base na sua distribuição relativa entre classes. Os modelos dos classificadores *Naïve Bayes* e SVM foram aplicados na predição das faixas etárias. Para gerar os resultados, os dados de um conjunto de registros de *chats*, descrito em Lin (2007), foi aplicado nos experimentos. O modelo do SVM obteve o maior desempenho e alcançou 0,996 de medida F em dados de teste.

Cheng et al. (2009) investigaram a classificação de gênero de autores de email. O conjunto de dados de email da Eron[1], versão 2005, foi aplicado nos experimentos. Também, uma técnica psico-linguística, pistas ligadas ao gênero e características estilométricas tradicionais foram aplicados para apoiar a solução. Os gêneros dos autores foram preditos por classificadores baseados em Árvore de Decisão e SVM. Os maiores valores alcançados nos experimentos foram uma acurácia média de 0,822 e uma macro-média F1 de 0,8145.

Argamon et al. (2009) aplicaram aprendizado de máquina à categorização de texto para classificar o perfil de autoria. Um conjunto de dados contendo postagens de autores de *blogs* em inglês foi usado para predizer gêneros e faixas etárias. As idades estão representadas em três faixas etárias: 13 a 17, 23 a 27 e 33 a 47 anos. Recursos baseados em conteúdo e em estilo foram aplicados nos experimentos. Os gêneros e as faixas etárias foram classificados por

[1] https://www.cs.cmu.edu/∼./enron/

um modelo de Regressão Multinomial Bayesiana. A maior acurácia alcançada para gênero foi 0,7610 e a maior acurácia alcançada para faixa etária foi 0,7770.

Zhang, Dang e Chen (2011) propuseram uma estrutura de classificação de gêneros de participantes de fóruns Web, através de características extraídas de estilos de escrita, tópicos de interesse e conteúdo de publicações feminino e masculino. Os experimentos foram executados em um conjunto de dados extraídos de um fórum político de mulheres islâmicas, onde foram comparados os desempenhos de diferentes conjuntos de características. De acordo com os autores, os melhores resultados foram obtidos com a aplicação de quatro tipos de características: características lexicais; características sintáticas; características estruturais; e características de conteúdo específico. Além disso, foi aplicado seleção de características. A média da acurácia foi 0,8600 e a micro-média F1 0,8645.

Peersman, Daelemans e Vaerenbergh (2011) investigaram a possibilidade de classificar as faixas etárias e os gêneros em mensagens curtas de bate-papo na RSO belga *Netlog*. A abordagem se apoia na categorização de texto, seleção de características e no método de classificação SVM. Experimentos avaliando a classificação das faixas etárias, incluindo os gêneros, foram executados e a abordagem alcançou 0,888 de acurácia. Em contraste com outros trabalhos (MARQUARDT et al., 2014; GUIMARÃES et al., 2017), a inclusão dos gêneros na classificação das faixas etárias apresentou baixa relevância de acordo com os autores.

Cheng, Chandramouli e Subbalakshmi (2011) investigaram a classificação de gêneros de autores em textos curtos. O conjunto de dados de email da Eron foi aplicado para avaliar classificação. Os autores avaliaram cinco conjuntos de características relacionados aos gêneros: baseadas em caracteres, baseadas em métricas estatísticas, baseadas em recursos sintáticos, baseadas na estrutura das frases e em palavras gramaticais. Apoiados nos conjuntos de características, os modelos de classificação SVM, Regressão Logística Bayesiana e Árvore de Decisão *Adaboost* foram projetados para predição dos gêneros. O maior valor alcançado nos experimentos foi 0,851 de acurácia.

Marquardt et al. (2014) apresentaram os resultados de duas abordagens de classificação multirrótulo, cujo objetivo foi predizer simultaneamente o gênero e a faixa etária de um autor. Na primeira abordagem o problema de multirrótulo é reduzido a um problema de rótulo único aplicando a transformação *Label Powerset*. A outra abordagem é uma transformação *Classifier Chains*: a saída de um classificador aplicado para gêneros é usado como entrada para um

classificador de faixas etárias. Os modelos de treinamentos foram construídos a partir de quatro mídias sociais diferentes: *blogs*, *tweets*, resenha de hotéis e postagens de uma mídia social não especificada. O dados de cada mídia social são em inglês e espanhol, salvo as resenhas de hotéis, que são em inglês. Nos experimentos, foram aplicados recursos de sentimentos e *emoticons*. Também, foram aplicados *Linguistic Enquiry and Word Count (LIWC)*, uma ferramenta de análise de texto. A classificação de faixas etárias obteve resultados abaixo de 0,5 acurácia, independente da abordagem. Já na classificação de gêneros dos autores, a abordagem apoiada em *Label Powerset* obteve os melhores resultados, com exceção no domínio da mídia social desconhecida em inglês. A acurácia alcançada para cada mídia social e idioma em relação aos gêneros, foram as seguintes: 0,6871 para *blogs* em inglês; 0,8068 para *blogs* em espanhol; 0,7115 para *tweets* em inglês; 0,7472 para *tweets* em espanhol, 0,5739 para mídia social desconhecida em inglês; 0,6462 para mídia social desconhecida em espanhol; e 0,6546 para revisões.

Company e Wanner (2014) apresentaram a classificação de gêneros e faixas etárias de autores, através de uma coleção de recursos baseadas em caracteres, palavras, sentenças, dicionário e recursos sintáticos. Além disso, recursos baseados em número de erros ortográficos por palavra, porcentagem de marcadores discursivos, freqüência de palavrões e abreviaturas, uso de voz passiva e recursos baseados em dicionário, ajudaram a melhorar a predição das faixas etárias. Três conjuntos de dados foram usados: o primeiro conjunto de dados foi construído para o trabalho com dados extraídos do *blog* de opinião do *New York Times*; o segundo conjunto de dados é composto de publicações informais de *blogs* e foi utilizado e descrito em (MUKHERJEE; LIU, 2010); e o terceiro conjunto de dados é composto de publicações informais extraídas do *blogger.com*, foi compilado e descrito em (SCHLER et al., 2006) e utilizado em (ARGAMON et al., 2009). Um método baseado na abordagem *Bagging* foi utilizado para classificar. O primeiro conjunto de dados alcançou 0,8283 de acurácia para gêneros, enquanto no segundo, foi obtido 0,9708 de acurácia. Em relação ao terceiro, foram alcançados 0,6809 de acurácia para gêneros e 0,6292 de acurácia para faixas etárias.

Sinha e Sinha (2015) utilizaram um sistema de inferência *fuzzy* de rede adaptável para classificar os gêneros em um conjunto de dados de *blogs*. Como entrada do modelo de classificação, foi submetido um vetor de características composto por três medidas: formal; preferencial de gênero; e estilística. Essas medidas são calculadas a partir de recursos semânticos identificados pelo *Part-Of-Speech Tagger* (*POS Tagger*), uma técnica de marcação de recursos

semânticos em dados textuais. Um experimento foi realizado pelos autores e o resultado obtido foi 0,8656 de acurácia para o teste de classificação de gêneros.

Pentel (2015) investigou o efeito de características diferentes na categorização de textos curtos para classificar as faixas etárias de um autor. Os dados textuais foram extraídos de mídias sociais, como *Facebook*, comentários de *blogs* e fóruns da Web. Contudo, 14 características diferentes foram extraídas e apenas valores numéricos e normalizados foram aplicados. Os modelos foram construídos a partir dos classificadores SVM, Regressão Logística e algoritmos Bayesianos. Nos experimentos, o modelo do classificador SVM obteve uma média de 0,968 de medida F e uma média de 0,965 de acurácia na classificação das faixas etárias de 7 à 15 anos e de 20 à 48 anos.

Aravantinou et al. (2015) abordaram o problema da classificação de gêneros de autores em *blogs*, com aplicação de um conjunto de identificação de idioma e de marcação de recursos semânticos *POS Tagger*. O conjunto de dados aplicado foi introduzido em (MUKHERJEE; LIU, 2010). Nos experimentos, oito modelos de classificação baseados em Árvore de Decisão, SVM e prendizagem sob demanda, foram aplicados. O melhor resultado alcançado foi 0,7050 de acurácia, com seleção de características e o modelo baseado em *Random Forest*.

Simaki et al. (2015) apresentaram uma metodologia para classificar gêneros de autores da Web, aplicando características de texto baseadas em sociolinguística. O conjunto de dados aplicado é composto por comentários de usuários de diferentes fontes da Web que incluem diversas áreas temáticas. Nos experimentos, oito modelos de classificação baseados em Árvore de Decisão, SVM, Rede Neural e Comitê de Classificadores foram aplicados. O melhor resultado alcançado foi 0,8436 de acurácia, com diferentes configurações de características e o classificador *Multilayer Perceptron*.

Silessi, Varol e Karabatak (2016) propuseram uma abordagem que combina algoritmos de aprendizado de máquina com processamento de textos para classificar os gêneros de mensagens de texto *Short Message Service (SMS)*. O conjunto de dados de SMS foi baixado do site da *National University of Singapore* [2]. Foram aplicados os algoritmos do ambiente *Weka*: *Naïve Bayes*, J48 e *Multilayer Perceptron*. As tarefas de processamento de textos aplicadas, foram: *lower case*, *tokenizer*, *stopwords*, *n-gram* e *stemming*. Nos experimentos executados, o algoritmo J48 obteve melhor performance e alcançou a média de 0,7239 de acurácia.

[2] https://www.comp.nus.edu.sg/entrepreneurship/innovation/osr/corpus/

Bayot e Gonçalves (2016) investigaram perfis de autores de *tweets* em inglês e espanhol, com o intuito de promover uma avaliação cruzada com outras mídias sociais. Segundo os autores a definição do perfil consiste na classificação por gêneros e faixas etárias. A idade é representada nas seguintes faixas etárias: 18 a 24, 25 a 34, 35 a 49, 50 a 64 e 65 a mais anos. O conjunto de treinamento usados pelos autores foi extraído de *tweets* e aplicado em (PARDO et al., 2016). Já o conjunto de dados para avaliação foram extraídos de *blogs*, resenhas de hotéis e outras mídias sociais. Esses dados foram usados em (PARDO et al., 2014). Nos experimentos, foram comparados os resultados após aplicação do *Term Frequency - Inverse Document Frequency (TF-IDF)* e uma rede neural de duas camadas que processa textos, denominada *Word2Vec*. O modelo do classificador SVM atingiu 0,265 de acurácia para mídias sociais e 0,477 de acurácia para *blogs*, na avaliação da classificação das faixas etárias em espanhol. Em relação à língua inglesa, alcançou 0,313 de acurácia para mídias sociais, 0,449 de acurácia para *blogs* e 0,240 de acurácia para revisões. Usando o mesmo classificador para a predição de gêneros na avaliação, alcançou-se 0,572 de acurácia para mídias sociais e 0,670 de acurácia para *blogs*, ambos em espanhol. Já para língua inglesa, alcançou 0,511 de acurácia para mídias sociais, 0,653 de acurácia para *blogs* e 0,510 de acurácia para revisões.

Modaresi, Liebeck e Conrad (2016) apresentaram uma abordagem para a tarefa de identificação do perfil de um autor, em função de gêneros e faixas etárias. As identificações dos gêneros e faixas etárias foram abordados como problemas de classificação. Foram aplicados recursos estilísticos e lexicais antes de treinar um modelo de Regressão Logística. No treinamento dos classificadores foram usados dados textuais extraídos do *Twitter* em inglês e espanhol. Os modelos foram avaliados aplicado um conjunto contendo dados de *tweets*, *blogs*, revisões e outras mídias sociais. Os melhores resultados para as tarefas de classificação de gêneros e faixa etárias em função da acurácia, foram 0,7564 para o inglês e 0,5179 para o espanhol, respectivamente.

AlSukhni e Alequr (2016) investigaram a classificação de gêneros de autores no *Twitter* em língua árabe. Foram avaliados a remoção de *stopwords* e a aplicação de *stemming*. Além disso, foram avaliados o número de palavras, comprimento médio das palavras e nomes de autores dos *tweets*, como característica adicionais. Os modelos dos classificadores *Naïve Bayes*, SVM, MNB, J48 e o *k-Nearest Neighbor* foram usados para classificar. O melhor resultado alcançado foi 0,9993 de acurácia, com as características adicionais e o classificador MNB.

Martinc et al. (2017) propuseram uma abordagem para classificar gêneros de usuários do *Twitter*. O conjunto de treinamento aplicado foi proposto no *Profiling Analysis 2017 (PAN 2017)* com quatro idiomas diferentes, agrupado por autores e rotulado por gêneros e idiomas. Antes de treinar o modelo foram aplicados conjuntos de recursos com *n-gram*, *POS Tagger*, lista de *emoticons* e listas de palavras dos idiomas. Foram testados diferentes classificadores, contudo, o modelo do classificador Regressão Logística alcançou o melhor resultado com uma acurácia de 0,8600 para o idioma em português na classificação dos gêneros.

Dwivedi et al. (2017) propuseram dois métodos para classificar os gêneros de autores de *blogs* baseados em dados textuais: o primeiro é um sistema manual de extração de características, o outro é um método de *deep learning*. Os experimentos foram executados em dois conjunto de dados. O primeiro é um conjunto de dados proposto em Mukherjee e Liu (2010), enquanto o segundo é um conjunto de dados composto por posts filtrados do *blog Authorship Corpus*[3]. No primeiro conjunto de dados, o sistema manual de extração de características obteve 0,7460 de acurácia e o método de *deep learning* obteve 0,7191. Já no segundo conjunto de dados, o sistema manual de extração de características obteve 0,7214 de acurácia, enquanto o método de *deep learning* obteve 0,8030 de acurácia.

Markov et al. (2017) sugeriram um método de classificação de gêneros e faixas etárias de autores, criando um modelo de Regressão Logística, a partir de recursos extraídos com o *Doc2Vec*. Foram realizados experimentos de classificação de gêneros e faixas etárias em uma mídia social e com o cruzamento de mídias sociais distintas. As coleções de dados usadas na classificação foram aplicadas em (PARDO et al., 2015) (PAN 2015) e (PARDO et al., 2016) (PAN 2016) com mensagens do *Twitter* em inglês, espanhol, holandês e italiano. Já na classificação de faixas etárias e gêneros com dados de mídias sociais cruzadas, foram aplicados os conjuntos do PAN 2016 e PAN 2015, para treino e teste, respectivamente. O conjunto do PAN 2016 é composto por *tweets* em inglês e espanhol. O conjunto do PAN 2015 contém revisões em inglês, publicações de mídias sociais e *blogs*, em inglês e espanhol. Os melhores resultados em uma mídia social foram alcançados em espanhol, 0,5044 de acurácia para faixas etárias e 0,7720 de acurácia para gêneros. Nas mídias sociais cruzadas, a classificação das faixas etárias obteve resultados abaixo dos 0,5 de acurácia; e a classificação dos gêneros alcançou 0,5175 de acurácia para revisões em inglês, 0,6477 de acurácia para *blogs* e 0,5590 de acurácia para

[3] http://u.cs.biu.ac.il/ koppel/BlogCorpus.htm

mídias sociais em espanhol.

Surendran et al. (2017) apresentaram um conjunto de recursos para predizerem as faixas etárias e os gêneros. O conjunto é composto de recursos baseados em *n-gram*, palavras e caracteres amplamente utilizados, sintaxe, semântica, métricas de legibilidade e humor. Os dados para avaliar foram extraídos da RSO *Twitter*. Os seguintes modelos foram aplicados para classificar as faixas etárias e o gêneros: CNN, SVM, *Random Forest*, Árvore de Decisão e *Naïve Bayes*. A CNN alcançou uma acurácia de 0,977 para gêneros e 0,901 para faixas etárias.

Isbister, Kaati e Cohen (2017) investigaram a possibilidade de classificar os gêneros de autores de *blogs* em cinco idiomas diferentes: inglês, sueco, francês, espanhol e russo. Um conjunto de *blogs* extraídos do serviço da *Google Blogger*[4], contendo dados sobre os blogueiros, foi classificada por código de países e aplicado nos experimentos. Os autores aplicaram recursos do LIWC, uma ferramenta de análise de texto, no conjunto de dados. O objetivo do uso dessa ferramenta foi solucionar a desvantagem de dependência da linguagem nas abordagens com conteúdo textual. Além disso, uma seleção de características e o classificador SVM foram usados para classificar os gêneros. A média da acurácia alcançada para cada idioma foram as seguintes: 0,7961 para o inglês, 0,7706 para o sueco, 0,7380 para o francês, 0,7424 para o espanhol, 0,7661 para o russo. Apesar dos resultados promissores, a abordagem fica limitada aos idiomas contidos na ferramenta.

Bsir e Zrigui (2018a) propuseram uma abordagem para o problema de classificação de gêneros por meio de uma rede neural recorrente bidirecional do tipo *Long Short-Term Memory (LSTM)*, combinada com a técnica de incorporação de palavras *Word2Vec*. O conjunto de dados aplicado foi coletado do *Twitter* e contém *tweets* escritos em árabe de 2400 autores. Os autores aplicaram uma validação cruzada para avaliar a classificação de gêneros, e o melhor resultado alcançado para os dados de teste foi 0,7923 de acurácia.

Liu e Cocea (2018) propuseram o uso de uma abordagem *fuzzy* para identificar gêneros. A identificação de gêneros nesse trabalho é tratada como uma tarefa de classificação generativo, segundo os autores. O conjunto de dados de *blogs* proposto em (MUKHERJEE; LIU, 2010) foi aplicado. Nos experimentos, a técnica de extração de característica *n-gram* foi aplicada, com dimensões um, dois e três. Em função da acuŕacia, a abordagem *fuzzy* foi comparada com as três abordagens de classificação discriminante: SVM, *Naïve Bayes* e C4.5. Os resultados demons-

[4] https://www.blogger.com/

tram que a abordagem *fuzzy* superou as abordagens discriminantes e alcançou o valor máximo de 0,892 de acurácia.

Soler-Company e Wanner (2018) propuseram um modelo baseado em características lingüísticas profundas para classificar os gêneros. As características são baseadas em seis grupos de recursos: caracteres, palavras, frases, dicionários, recursos sintáticos, e recursos discursivos (e.g., organização textual, tópicos e outros). Um conjunto de dados, denominado "*Blog-Dataset*", com publicações jornalísticas de vinte três autores de *blogs* britânicos, foi aplicado nos experimentos. Também, diferentes combinações de recursos e o classificador *LibSVM* do *Weka*, foram usados nos experimentos. O melhor resultado alcançado foi 0,8997 de acurácia, com todos os recursos aplicados.

Briedienė e Kapočiutė-Dzikienė (2018) avaliaram métodos de classificação para predizer automaticamente dados de perfis de usuários lituanos: sexo, idade, escolaridade, estado civil e personalidade; através de textos curtos e não normativos. O conjunto de dados usado, foi extraído da RSO *Facebook* entre 2016 e 2017. Foram aplicados *n-gram* de caracteres e seqüências de palavras, antes de aplicar cinco classificadores baseados em aprendizagem supervisionada e similaridade. Os melhores desempenhos alcançados, 0,843 de acurácia para gêneros e 0,527 de acurácia para faixas etárias, foi obtido pelo classificador MNB.

Bsir e Zrigui (2018b) abordaram o problema de detecção de gêneros de usuários árabes em mídias sociais, com recursos estilísticos de texto e uma rede neural recursiva. Os recursos aplicados foram: recursos lexicais, recursos de sintaxe, *n-gram*, frequência de caracteres, lista de emoticons, *stopwords*. A rede neural que foi aplicada, é uma variante da arquitetura *Gated Recurrent Units (GRU)*. Dois conjuntos de dados foram usados nos experimentos. O primeiro, extraído do *Facebook*, foi aplicado em Bsir e Zrigui (2017). O segundo, extraído do *Twitter* e proposto no PAN 2017, foi aplicado em Pardo et al. (2017). Os resultados alcançados foram 0,790 de acurácia no *Twitter* e 0,621 de acurácia no *Facebook*.

Al-Ghadir e Azmi (2019) propuseram a classificação de gêneros para textos em árabes com o apoio de duas listas de recursos. A primeira lista representa a pontuação Tf-Idf de cada palavra em ordem decrescente. A segunda lista contém o radical das palavras mais usadas por ambos os gêneros. Um conjunto de publicações de um fórum Web[5] foram usadas nos experimentos. Os classificadores SVM e 1-NN (vizinho mais próximo) predisseram os gêneros, com

[5] www.eqla3.com

variações de tamanho nas listas de recursos. O melhor resultado alcançado, uma acurácia de 0,9316, foi conseguido através da recomendação dos autores com 100 recursos por listas.

Bacciu et al. (2019) apresentaram uma abordagem para a tarefa de criação de perfil do PAN 2019. O autores classificaram os gêneros de usuários de *tweets*. Conjuntos de dados em inglês e espanhol foram usados na avaliação. Os autores aplicaram técnicas de tokenização e *stemming* nas instâncias do conjunto de dados em inglês. Também, aplicaram técnicas de tokenização e lematização nas instâncias do conjunto de dados em espanhol. Além disso, as técnicas de extração de características n-gram, Tf-Idf e *Latent Semantic Analysis (LSA)* foram aplicadas. Três modelos construídos por classificadores diferentes foram comparados. O modelo baseado em SVM alcançou os maiores resultados, 0,8548 de acurácia em inglês e 0,7130 de acurácia em espanhol.

Cimino e Dell'Orletta (2019) propuseram três abordagens para a tarefa de criação de perfil do PAN 2019. A primeira abordagem é baseada em um classificador SVM. A segunda é baseada em um classificador *Bidirectional Encoder Representations from Transformers (BERT)* e a terceira abordagem é baseada em uma rede neural hierárquica *Gated Recurrent Unit - Long Short Term Memory (GRU-LSTM)*. Um conjunto de dados com *tweets* em inglês foi usado para treinar os modelos. As instâncias do conjunto de dados foram rotuladas com *bot*, *male* ou *female*. Antes de usar o conjunto de dados para treinar os modelos, foram usados um marcador morfo-sintático, um léxico de polaridade de sentimentos e um léxico de incorporação de palavras para *tweets* em inglês. O marcador morfo-sintático foi aplicado somente para abordagem baseada em SVM. O maior resultado para criação de perfis de gêneros foi alcançado pelo Hierárquico GRU-LSTM, com uma acurácia de 0,7898.

Wu et al. (2019) propuseram abordagens neurais para classificar os gêneros nas mídias sociais, baseado no conteúdo e emoções das mensagens de *microblog*. Dois conjuntos de dados construídos com conteúdos das mídias sociais chinesa foram aplicados: *Natural Language Processing and Chinese Computing 2018 (NLPCC2018)* e *Sina Weibo*. Os autores aplicaram os modelos denominados *Neural Gender Prediction (NGP)*, *Neural Gender Prediction Emotion-aware Message Representations (NGP-EaRM)* e *Neural Gender Prediction Emotion-aware Message Encoder (NGP-EaME)*. Os desempenhos foram medidos em acurácia e medida F. Os três modelos alcançaram desempenhos superiores a outros sete usados como linha de base. Os maiores desempenhos no conjunto de dados NLPCC2018 foram alcançados pelo

NGP-EaME, 0,9351 de acurácia e 0,9350 de medida F. Os maiores desempenhos no conjunto de dados *Sina Weibo* foram alcançados pelo NGP-EaRM, 0,6903 de acurácia e 0,6796 de medida F.

Li et al. (2019) propuseram um novo método de classificação de gêneros e faixas etárias de usuários, usando informações sociais e semânticas. O conjunto de dados aplicados nos experimentos foi extraído da RSO *Sina Weibo* e contém três tipos dados de relações sociais, *microblogs* e atributos do usuário. Um modelo baseado em *Text Attention Neural Network (TA-NN)* e dois modelos baseados em *Social network Attention Neural Network (SA-NN)* foram desenvolvidos pelos autores. Na classificação de gêneros, ocorreu um empate de 0,8742 de acurácia, entre as abordagens TA-NN e SA-NN. Além disso, a abordagem SA-NN alcançou 0,6556 de acurácia; maior desempenho na classificação das faixas etárias com três classes.

Maslennikova et al. (2019) apresentaram um tarefa de classificação de faixas etárias para textos em italiano. O conjunto de dados aplicado, foi construído com postagens de fóruns públicos. Um classificador baseado na abordagem SVM linear construiu os modelos de classificação. Os autores avaliaram os pré-processamentos lexicais, morfossintáticos e sintáticos. Além disso, os experimentos avaliaram a classificação com duas classes e cinco classes, em diferentes domínios. Os experimentos com pré-processamento lexicais alcançaram ótimos desempenhos em todos os domínios, sendo que o domínio carro, proporcionou os maiores desempenhos com 0,54 de medida F em cinco classes e 0,87 de medida F em duas classes.

Pascucci, Masucci e Monti (2019) apresentaram a importância do suporte à estilometria computacional e ao aprendizado de máquina, para classificar gêneros e faixas etárias de autores, em textos de *cyberbullying*. A classificação de faixas etárias nesse trabalho, é um problema com 10 classes. O classificador aplicado na classificação de gêneros foi o *Random Forest* e o classificador aplicado na detecção de faixas etárias foi o Mínimo Sequencial Suporte. Os resultados dos desempenhos foram medidos em precisão, revocação e medida F. Os maiores desempenhos para classificação de gêneros foram 0,686, 0,688 e 0,679, respectivamente. E os maiores desempenhos para classificação de faixas etárias foram 0,500, 0,549 e 0,459, respectivamente.

Zheng et al. (2019) propuseram um modelo de *Sentiment Representation Learning* baseado em *Multilayer Perceptron* (SRL-MLP) para classificar gêneros. Um conjunto de dados de *microblogs* Chineses foi usado no trabalho. Os autores aplicaram um modelo baseado em LSTM, treinado com uma coleção de revisão de mercadorias, para classificar a polaridade de

sentimentos. O resultado da classificação de sentimentos foi concatenado com vetores de documentos extraídos do conjunto de *microblog*. A concatenação foi usada para treinar a rede MLP. Nos experimentos, o modelo alcançou uma acurácia máxima de 0,8973.

Bassem e Zrigui (2020) avaliaram modelos de *deep learning* com incorporação de palavras para prever os gêneros de autores árabes de *tweets*. O conjunto de dados aplicado nos experimentos foi usado no PAN 2018. Os modelos criados foram construídos pelos algoritmos CNN, GRU e LSTM. Os desempenhos dos modelos foram medidos e comparados em acurácia, revocação e medida F. O modelo GRU alcançou os maiores desempenhos em duas medidas, 0,796 de acurácia e 0,786 de revocação. O modelo LSTM obteve o maior desempenho na medida F, com o valor de 0,795.

Dong, Mihalcea e Radev (2020) classificaram gêneros de usuários usando dados textuais de perfis de *blogs*. O trabalho avaliou o desempenho do classificador Regressão Logística, após aplicação de técnicas de expansão de características. O maior desempenho alcançado pelo classificador foi 0,813 de acurácia.

Gupta et al. (2020) propuseram um método de incorporação de palavras para melhorar o desempenho de tarefas de classificação de dados textuais. Os autores aplicaram um conjunto de dados de *blogs* rotulados com o gêneros dos autores, para avaliar o metodologia proposta neste domínio. Sete algoritmos de classificação foram aplicados. Os desempenhos dos modelos foram comparados pela métrica *Area Under the ROC Curve (AUC)*. O maior desempenho foi do modelo *Support Vector Machine with Optimized Word Embeddings (SVM-OptEm)*, que alcançou 0,7992 de AUC.

Zhong et al. (2020) realizaram experimentos sobre a compressão do modelo e análise de atributos dos usuários na estrutura *Capsule Network*. Os autores analisaram os dados de um conjunto de *microblogs* e os atributos do usuário, para classificar uma diferença emocional entre os gêneros. Em seguida, um método de aprendizado de migração foi aplicado para rotular *tags* emocionais nos exemplos do conjunto de *microblogs*. Por fim, os autores propuseram o algoritmo de classificação de gêneros *Sentiment Polarity with Capsule Networks (SPT-CapsNet)*, uma adaptação do *Capsule Networks* baseado em análise de sentimentos. Nos experimentos com o conjunto de dados *microblog*, o modelo do SPT-CapsNet alcançou uma acurácia de 0,8589 e superou o *Capsule Network*.

2.6.2 Classificação de gêneros e faixas etárias através de características de páginas dos perfis de usuários

Em (ALOWIBDI; BUY; YU, 2013b) a classificação de gêneros foi baseada no nome do perfil, nome de usuário, cor de fundo do perfil, cor do texto, cor do link, cor de preenchimento da barra lateral e cor da borda da barra lateral. Essas características foram extraídas do perfil de usuários do *Twitter* e comparadas empiricamente para encontrar os seus pontos fortes e fracos. Foram realizados diferentes experimentos com aprendizagem de máquina e com conjuntos de características distintas. O melhor resultado foi uma acurácia média de 0,8250, cujo experimento foi realizado com a característica nome do perfil, uso do agrupamento de palavras *trigrams* e um modelo de Árvore de Decisão. Com as caracterśiticas baseadas em cores, o melhor resultado foi uma acurácia média de 0,7400, cujos experimentos usaram uma quantização de todas as cores e o modelo *Naïve Bayes Decision Tree*. Também em (ALOWIBDI; BUY; YU, 2013a), as cinco características baseadas em cores extraídas da RSO *Twitter* foram aplicadas para classificar os gêneros sem depender do idioma. O melhor resultado foi uma acurácia média de 0,7740, cujos experimentos usou uma quantização de todas as cores e o modelo *Naïve Bayes Decision Tree*. De acordo com os autores, uma das vantagens dessa abordagem é a ampla aplicabilidade para diferentes idiomas.

Outra abordagem baseada em características de perfis de usuário foi aplicada em (LI; LI; JI, 2018), para classificar os gêneros através de dados extraídos da RSO *Sina Weibo*. Os autores usaram dados estatísticos e exploraram o comportamento de republicação homofílico. Os dados são baseados em uma proporção feminina de 20 comunidades, no número de perfis que o usuário segue, no número de seguidores que o usuário possui, no número de *blogs* publicados pelo usuário, se é um usuário verificado ou não, se é um usuário avançado ou não e o nível da conta do usuário. Os autores propuseram quatro métodos para classificar os gêneros, entre esses, o método Média Aritmética Comunitária que alcançou os melhores resultados, com uma acurácia média de 0,877 e uma medida F média de 0,899. Contudo, uma rede neural artificial que foi aplicada, alcançou uma acurácia média de 0,884 e uma medida F média de 0,882. Os autores citam como boas características específicas do trabalho: a captura da homofilia contida nas rotas de difusão de conteúdo online; a classificação de gêneros independente da linguagem; a robustez para pequenos conjuntos de dados; e o bom desempenho competitivo em dados limitados. Contudo, os autores consideram as hipóteses de não ser fácil obter as informações

exigidas e de não ser possível obter as relações de repostagem.

Kiratsa et al. (2018) analisaram os perfis de usuários do *Facebook* com o objetivo de classificação de gêneros desses. O objetivo foi atingido por meio da construção de modelos de classificação, baseados em características das seguintes interações dos usuários: a quantidade de curtidas de cada assunto das páginas de usuários e os subassuntos curtidos de cada assunto. Os assuntos abordam: filmes, programas de TV, música, livros, esportes, jogos, eventos e atletas de equipes. Nos experimentos, dez classificadores foram aplicados. Os melhores desempenhos, 0,9730 de acurácia e 0,9741 de medida F, foram do classificador Árvore de Decisão *Adaboost*.

Em (GUIMARÃES et al., 2017), os autores usaram dados extraídos de páginas de perfis de usuários do *Twitter*, para classificar as faixas etárias, adolescente e adulto. Os dados são baseados nas informações dos perfis de usuários e em suas características de escrita. Os modelos de classificação dos algoritmos *Multilayer Perceptron*, Árvore da Decisão, *Random Forest*, SVM e CNN foram aplicadas nos experimentos. O melhor resultado foi com a CNN, após eliminação de características irrelevantes. Desse modo, o valor máximo alcançado na medida F foi 0,940 na fase de validação.

2.6.3 Classificação de gêneros e faixas etárias usando conteúdo textual e características de páginas dos perfis de usuários

Em (ALSUKHNI; ALEQUR, 2016) investigaram a identificação de gêneros de autores de *tweets* em árabe, usando diferentes técnicas de classificação. O conjunto de dados usado pelos autores, contém *tweets*, nomes dos autores dos *tweets*, tamanho médio das palavras de cada *tweet* e número de palavras de cada *tweet*. Os autores avaliaram os algoritmos de classificação com pré-processamento e sem pré-processamento nos dados textuais. As seguintes técnicas de pré-processamento foram aplicadas: remoção de *stopwords*, tokenização, *stemming* e normalização. Cinco algoritmos foram usados no trabalho: *Naïve Bayes (NB)*, SVM, MNB, J48 e KNN. Os melhores resultados alcançados foram sem pré-processamento e com a adição dos nomes dos autores dos *tweets*. Os classificadores MBN, SVM e J48 atingiram uma acurácia acima de 0,980.

Pandya et al. (2018) abordaram o problema de identificação de faixas etárias de usuários do *Twitter*, como uma tarefa de classificação. Os autores usaram dois conjuntos de dados, um em holandês e outro em inglês, aplicados em (NGUYEN et al., 2013) e (MORGAN-LOPEZ et

al., 2017), respectivamente. O modelo de classificação foi criado com recursos: lingüísticos, estilométricos e léxicos. Também, foram usados metadados de perfis de usuários, de páginas do *Twitter* e de *tweets*, tais como URL e *hashtags*. Um modelo baseado na rede CNN foi construído e aplicado nos experimentos. Os melhores resultados alcançados foram 0,810 de micro-média F1 para inglês e 0,820 de micro-média F1 para holandês.

Pandya et al. (2020) abordaram o problema de previsão de faixas etárias em conjuntos de dados com conteúdos do *Twitter*. Três conjuntos de dados: holandês, inglês 1 e inglês 2 foram aplicados. Os conjuntos de dados contém URL, *hashtags*, palavras e frases de *tweets*. Os autores aplicaram um classificador baseado em CNN e compararam os desempenhos obtidos com os de outros classificadores baseados em SVM, *Random Forest* e Regressão Logística. Os desempenhos foram avaliados em micro-média F1. O CNN alcançou os maiores desempenhos, com 0,82 em holandês, 0,86 em inglês 1 e 0,82 em inglês 2.

2.7 Considerações finais

Neste capítulo, foram apresentados a definição das tarefas que compõem este trabalho e o estado da arte que ele está inserido. Além disso, a descrição textual buscou justificar as técnicas escolhidas para solucionar o problema proposto e o tema do trabalho: caracterização de gêneros e faixas etárias através de aprendizagem supervisionada multidimensional. Na metodologia, será apresentado as ferramentas e procedimentos seguidos para criar e validar os modelos propostos.

3 METODOLOGIA

A Figura 3.1 contém uma representação de alto nível da metodologia proposta neste trabalho para classificação de gêneros e faixas etárias. A metodologia foi implementada na linguagem *Python*, com a aplicação de algoritmos das bibliotecas de aprendizado de máquina *Scikit-Learn*, *Scikit-Multilearn* e *XGBoost*. As etapas para construir os modelos de classificação foram as seguintes:

1. Conjunto de dados: o conjunto de dados aplicado foi dividido em conjunto de treino e conjunto de teste;

2. Transformação de características: normaliza e converte os dados para as entradas de algoritmos de aprendizado de máquina;

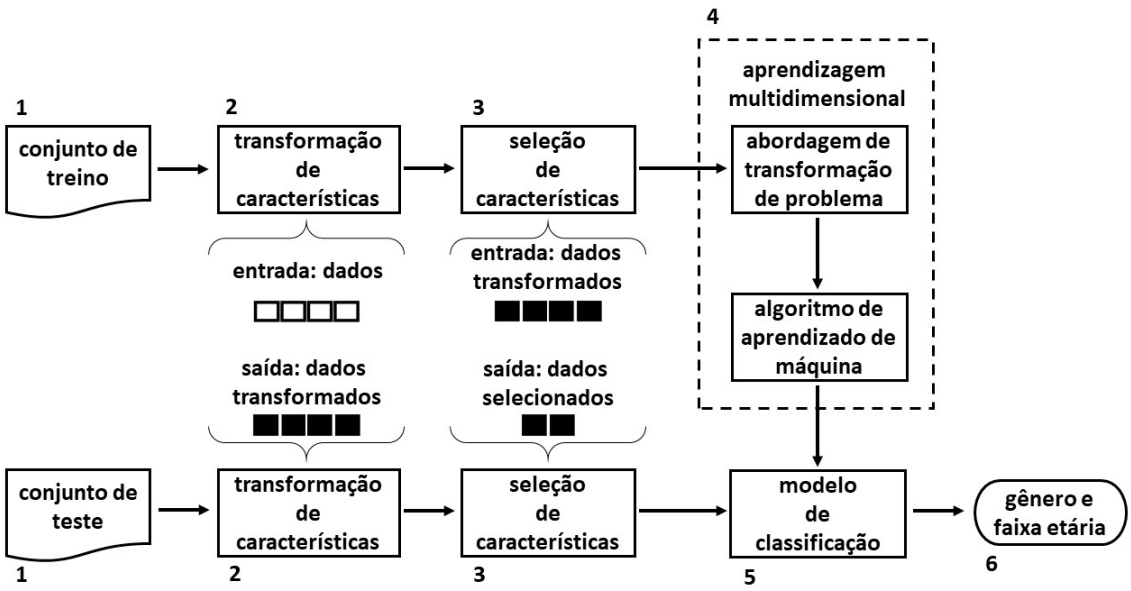
3. Seleção de características: seleciona um subconjunto de atributos que permitem uma melhor distinção entre instâncias de cada classe de rótulos;

4. Aprendizagem multidimensional: Treina os modelos para classificar gêneros e faixas etárias. Assim, os algoritmos de aprendizado de máquina foram aplicados com as transformações CC e LP, para cada subconjunto de atributos que foi selecionado. Nesta etapa, usou-se um conjunto de treinamento para encontrar os melhores parâmetros de cada modelo, através da validação cruzada com dez pastas;

5. Modelo de classificação: é o resultado final obtido da aprendizagem multidimensional após a aplicação do conjunto de treinamento. Os exemplos do conjunto de testes, sem os subconjuntos de classes alvos, foram enviados ao modelo para serem classificados;

6. Gênero e faixa etária: cada subconjunto de classes, composto por gênero e faixa etária, é o resultado da aplicação de um modelo de classificação, conforme cada exemplo do conjunto de teste. Esses subconjuntos são comparados com os subconjuntos de classes alvos para medir os resultados dos modelos.

Figura 3.1 – Metodologia para criar o modelo de classificação.



Fonte: Autor (2020)

3.1 Conjunto de dados

Este trabalho utilizou dados extraídos das páginas de perfis dos usuários do *Twitter* para classificar as faixas etárias adolescente e adulto, e os gêneros masculino e feminino dos usuários. Oito mil mensagens foram extraídas de determinados assuntos via *Application Programming Interface (API)* do *Twitter*. As palavras-chave utilizadas para coleta das mensagens, tem como propósito abranger um grande número de usuários com características de perfis diferentes, que remetem a temas como responsabilidade, esporte, saúde, religião, trabalho e família. A Tabela 3.1 mostra a quantidade de mensagens coletadas por palavras-chave.

Tabela 3.1 – Palavras-chave usadas para coleta das mensagens

Palavras-chave	**Quantidade de mensagens**
“convênio médico”	1333
“alimentação saudável”	1322
“carboidrato”	1334
“novena”	1335
“trabalho profissional”	1321
“mãe”	1355

Fonte: Autor (2020)

Um conjunto de dados de assuntos similares, com 6280 instâncias foi utilizada em (GUIMARÃES et al., 2017) para classificar as classes do rótulo faixa etária: adolescente e adulto.

Além da faixa etária, neste trabalho, o atributo gênero foi utilizado como rótulo. Além disso, as 8000 instâncias foram utilizadas, e elas têm a seguinte distribuição de classes: 50,41% pertencem à classe adolescente (*teenager* - valor 1 da faixa etária do rótulo); 49,59% pertencem à classe adulto (*adult* - valor 0 do rótulo da faixa etária); 60,60% das instâncias pertencem à classe feminina (female - valor 0 do rótulo de gênero); e 39,40% pertencem à classe masculina (male - valor 1 do rótulo de gênero). Além disso, o conjunto está estruturado em *Attribute-Relation File Format (ARFF)* e possui a seguinte estrutura:

```
@relation 'twitter: -C 2 -R'
@attribute gender {0,1}
@attribute age-group {0,1}
@attribute rt {YES,NO}
@attribute at {NO,YES}
@attribute hashtag {NO,YES}
@attribute slang {NO,YES}
@attribute punctuation {NO,YES}
@attribute url {NO,YES}
@attribute characters numeric
@attribute follow numeric
@attribute followers numeric
@attribute tweets numeric
@attribute topic numeric

@data
0,1,NO,NO,NO,NO,NO,YES,68,0,6,5984,3
```

Os dados apresentados na estrutura acima se referem ao uso (*YES*) ou não uso (*NO*), classes (0 ou 1) ou quantidade dos seguintes atributos:

- *gender* (gênero): atributo rótulo do conjunto multidimensional, ele assume duas classes referentes ao sexo: 0 ou 1, feminino ou masculino, respectivamente;
- *age-group* (faixa etária): atributo rótulo do conjunto multidimensional, assume duas classes: 0 ou 1, adulto ou adolescente, respectivamente;

- *rt*: considera se a sentença é um *retweet*;
- *at* (arroba): é um marcador, considera o uso ou não uso do símbolo @;
- *hashtag*: é um marcador, considera o uso ou não do símbolo #;
- *slang* (gíria): considera o uso ou não uso de gírias, que pertencem a um dicionário;
- *punctuation* (pontuação): considera o uso ou não de pontuação, caracteres (“ ” ? ! ... : () /) e símbolos que expressam emoções;
- url: considera que a mensagem tem ou não um link (URL) apontando para outra página ou algum anexo;
- *characters* (caracteres): é um parâmetro numérico que representa o número de caracteres na frase;
- *follow* (seguir): é um parâmetro numérico que representa o número de usuários que um usuário específico segue;
- *followers* (seguidores): é um parâmetro numérico que representa o número de seguidores que cada usuário possui;
- *tweets*: é um parâmetro numérico que representa o número total de *tweets* que o usuário escreveu em seu perfil;
- *topic* (assunto): é um identificador numérico do assunto principal da frase.

O conjunto de dados foi dividido em 80% para treino e 20% para teste. Essa divisão foi estratificada e manteve a proporção de instâncias para cada subconjunto de classes. Assim, a estratificação foi baseada em combinações das classes dos rótulos. Como nas abordagens de transformação, se considerou uma correlação entre os rótulos nessa divisão.

3.2 Transformação de características

A transformação de recursos consiste em converter e modificar todos os valores de um atributo. O objetivo é preparar os dados para os algoritmos de aprendizado de máquina. Neste trabalho, a função *normalize* do pacote *sklearn.preprocessing* realizou a normalização de dados numéricos, e a classe *LabelEncoder* do pacote *sklearn.preprocessing* realizou a conversão de

dados nominais para numéricos. A normalização dos dados consiste em colocar os valores no intervalo de $[0,1]$. Por exemplo, o atributo *characters* tem um valor médio de 87 caracteres e um intervalo de 13 a 140 caracteres por mensagem. Aplicando a normalização para o intervalo de 0 a 1, o valor médio torna-se aproximadamente 0,62.

3.3 Seleção de características

A seleção de características ocorre após a transformação de características. Neste trabalho, a abordagem filtro foi escolhida empiricamente. Essa abordagem selecionou os subconjuntos de características com melhor desempenho na classificação do que a abordagem *wrapper*, em testes experimentais. Além disso, não é inerente a um algoritmo de classificação como a abordagem embutida (TAN; STEINBACH; KUMAR, 2005).

A classe *SelectKBest* do pacote *sklearn.feature_selection* executa esta tarefa. A classe recebe como parâmetros uma instância da função *chi2* do pacote *sklearn.feature_selection* e um valor natural de K referente ao número de características a serem selecionadas. Conforme descrito na seção 2.2, o qui-quadrado mede a falta de independência entre características e classe. Dessa forma, a função *chi2* calcula a pontuação do qui-quadrado e as K características com as pontuações mais altas são selecionadas, ou seja, as características mais dependentes são selecionadas. A Tabela 3.2 mostra as classificações dos atributos mais relevantes por rótulos.

Tabela 3.2 – Rank de atributos por rótulo

(a) Rank do Gênero

Rank	**Atributo**	X^2
1	*slang*	191,225
2	url	108,077
3	*hashtag*	88,364
4	rt	25,127
5	*characters*	13,975
6	*followers*	8,356
7	*punctuation*	8,010
8	*follow*	7,170
9	*tweets*	2,935
10	*at*	0,499
11	*topic*	0,483

(b) Rank da Faixa Etária

Rank	**Atributo**	X^2
1	*slang*	750,200
2	url	527,100
3	*at*	21,990
4	*punctuation*	21,100
5	rt	19,670
6	*followers*	2,834
7	*follow*	1,458
8	*topic*	0,737
9	*tweets*	0,241
10	*hashtag*	0,193
11	*characters*	0,166

Fonte: Autor (2020)

A seleção é feita por rótulo, de acordo com os K melhores atributos, tal que $\{k \in \mathbb{N} : 1 \leq k \leq 9\}$. Consequentemente, dois subconjuntos com K características são selecionadas e unidos. O limite superior $K = 9$ é o controle, porque a união dos subconjuntos selecionados de cada rótulo mantém todos as características do conjunto de dados. Portanto, a redução das características ocorre somente, a partir de $K = 8$. A Tabela 3.3 mostra os subconjuntos selecionados como uma função de K.

Tabela 3.3 – Seleção de características

k-value	**Subconjunto de Gênero**	**Subconjunto de faixa etária**	**União de Subconjuntos**
9	{rt, *hashtag*, *slang*, *punctuation*, url, *characters*, *follow*, *followers*, *tweets*}	{rt, *at*, *slang*, *punctuation*, url, *follow*, *followers*, *tweets*, *topic*}	{rt, *at*, *hashtag*, *slang*, *punctuation*, url, *characters*, *follow*, *followers*, *tweets*, *topic*}
8	{rt, *hashtag*, *slang*, *punctuation*, url, *characters*, *follow*, *followers*}	{rt, *at*, *slang*, *punctuation*, url, *follow*, *followers*, *topic*}	{rt, *at*, *hashtag*, *slang*, *punctuation*, url, *characters*, *follow*, *followers*, *topic*}
7	{rt, *hashtag*, *slang*, *punctuation*, url, *characters*, *followers*}	{rt, *at*, *slang*, *punctuation*, url, *follow*, *followers*}	{rt, *at*, *hashtag*, *slang*, *punctuation*, url, *characters*, *follow*, *followers*}
6	{rt, *hashtag*, *slang*, url, *characters*, *followers*}	{rt, *at*, *slang*, *punctuation*, url, *followers*}	{rt, *at*, *hashtag*, *slang*, *punctuation*, url, *characters*, *followers*}
5	{rt, *hashtag*, *slang*, url, *characters*}	{rt, *at*, *slang*, *punctuation*, url}	{rt, *at*, *hashtag*, *slang*, *punctuation*, url, *characters*}
4	{rt, *hashtag*, *slang*, url}	{*at*, *slang*, *punctuation*, url}	{rt, *at*, *hashtag*, *slang*, *punctuation*, url}
3	{*hashtag*, *slang*, url}	{*at*, *slang*, url}	{*at*, *hashtag*, *slang*, url}
2	{*slang*, url}	{*slang*, url}	{*slang*, url}
1	{*slang*}	{*slang*}	{*slang*}

Fonte: Autor (2020)

3.4 Modelos de aprendizagem de máquina

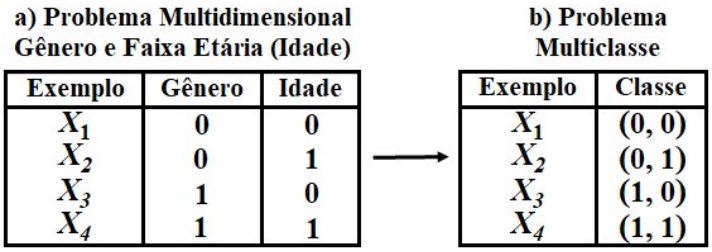
Os recursos obtidos com a seleção das características serviram de base para a construção dos modelos de classificação. Os modelos de classificação bidimensional são produtos de algoritmos de classificação e das classes *LabelPowerset* e *ClassifierChains* do pacote *sk-multilearn.problem_transform*. A classe *LabelPowerset* transforma um problema multirrótulo em multi-classe. Assim, com duas classes em cada atributo rótulo do conjunto de dados e quatro combinações possíveis, foi gerado quatro classes para cada modelo de classificação: *feminino-adolescente*; *masculino-adolescente*; *feminino-adulto*; e *masculino-adulto*. A Tabela 3.4 apresenta a distribuição das instâncias em função dessas classes e a Figura 3.2 exemplifica a transformação LP feita pela classe *LabelPowerset*.

Tabela 3.4 – Distribuição das instâncias em função das classes na transformação LP

Classe	Porcentagem
feminino-adolescente	38,20%
masculino-adolescente	12,21%
feminino-adulto	22,40%
masculino-adulto	27,19%

Fonte: Autor (2020)

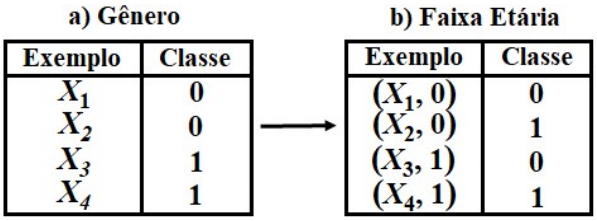
Figura 3.2 – Transformação LP



Fonte: Autor (2020)

A classe *ClassifierChains* cria um classificador para cada atributo rótulo, e cada previsão de um classificador predecessor é usada como entrada para os classificadores sucessores. A Figura 3.3 exemplifica a transformação CC feita pela classe *ClassifierChains*. Assim, dois classificadores são treinados nessa abordagem - o primeiro para os gêneros e o segundo para a faixas etárias. Em Guimarães et al. (2017) e Marquardt et al. (2014), o gênero é um atributo relevante na classificação da faixa etária; portanto, o primeiro classificador treinado é o que representa os gêneros.

Figura 3.3 – Transformação CC



Fonte: Autor (2020)

Os cinco algoritmos aplicados para construir os modelos foram os seguintes: *Decision Tree Classifier (DTC)* do pacote *sklearn.tree*; *Random Forest Classifier (RFC)* do pacote *sklearn.ensemble*; *Extreme Gradient Boosting (XGB)* do pacote *xgboost*; *Support Vector Classification (SVC)* do pacote *sklearn.svm*; e o *Complement NB (CNB)* do pacote *sklearn.naive_bayes*. Alguns parâmetros dos algoritmos de aprendizado tiveram seus valores alterados. Os melhores valores para os parâmetros de cada algoritmo foram definidos a partir de um *grid-search*, implementada pela classe *Grid-SearchCV* do pacote *sklearn.model_selection*.

Os modelos avaliados dos classificadores CC-DTC e LP-DTC são resultados de 104.976 configurações de valores de parâmetros em cada treinamento. Os parâmetros com seus respectivos valores ou faixas de valores foram:

- *criterion*: [‘gini’, ‘entropy’]
- *splitter*: [‘best’, ‘random’]
- *min_samples_split*: [2, $\cdots$, 8]
- *min_samples_leaf*: [1, $\cdots$, 4]
- *min_weight_fraction_leaf*: [0, $\cdots$, 0.5]
- *max_features*: [None, $\cdots$, ’log2’]
- *max_leaf_nodes*: [None, $\cdots$, 4]
- *min_impurity_decrease*: [0, $\cdots$, 0.5]
- *min_impurity_split*: [1e-8, $\cdots$, 1e-6]

Em relação aos modelos dos classificadores CC-RFC e LP-RFC, são resultados de 157.464 configurações de valores de parâmetros em cada treinamento. Os parâmetros com seus respectivos valores ou faixas de valores foram:

- *n_estimators*: [14, $\cdots$, 100]
- *criterion*: [‘gini’, ‘entropy’]
- *max_depth*: [None, $\cdots$, 8]
- *min_samples_split*: [2, $\cdots$, 8]
- *min_samples_leaf*: [1, $\cdots$, 4]
- *max_features*: [‘auto’, ‘sqrt’, ‘log2’]
- *max_leaf_nodes*: [None, $\cdots$, 4]
- *min_impurity_split*: [1e-8, $\cdots$, 0]
- *class_weight*: [None, ‘balanced’, ‘balanced_subsample’]

Os modelos de classificação CC-XGB e LP-XGB são resultados de 92.160 configurações de valores de parâmetro em cada treinamento. Os parâmetros com seus respectivos valores ou faixas de valores foram:

- *learning_rate*: [0.1, 0.5, 1, 10]
- *subsample*: [0.5, 1]
- *scale_pos_weight*: [0.5, 1]
- *reg_lambda*: [0.5, 1]
- *reg_alpha*: [0.5, 1],
- *colsample_bytree*: [0.5, 1]
- *colsample_bynode*: [0.5, 1]
- *colsample_bylevel*: [0.5, 1]

- *booster*: [gbtree', 'dart']
- *max_depth*: [10, 20, 30]
- *n_estimators*: [10, 15, 30]

Em relação aos modelos dos classificadores CC-SVC e LP-SVC, são resultados de 648 configurações de valores de parâmetros em cada treinamento. Nos modelos baseados em SVC, os parâmetros foram combinados de acordo com o kernel, gerando quatro subespaços de pesquisa. Os quatro subespaços de parâmetros com seus respectivos valores ou intervalos de valores foram:

- *kernel linear*:
  - *C*: [10, $\cdots$, 1000]
  - *tol*: [0.001, $\cdots$, 0.1]
- *kernel rbf*:
  - *C*: [10, $\cdots$, 1000]
  - *gamma*: scale
  - *tol*: [0.001, $\cdots$, 0.1]
- *kernel sigmoid*:
  - *C*: [10, $\cdots$, 1000]
  - *gamma*: scale
  - *coef0*: [0.001, $\cdots$, 0.1]
  - *tol*: [0.001, $\cdots$, 0.1]
- *kernel poly*:
  - *C*: [10, $\cdots$, 1000]
  - *gamma*: scale
  - *coef0*: [0.001, $\cdots$, 0.1]

  - *tol*: [0.001, $\cdots$, 0.1]

Os modelos dos classificadores CC-CNB e LP-CNB são resultados de 396 configurações de valores de parâmetros em cada treinamento. Os parâmetros com seus respectivos valores ou faixas de valores foram:

- *alpha*: [0.0, $\cdots$, 1.0]
- *fit_prior*: [False, True]
- *class_prior*: None
- *norm*: [False, True]

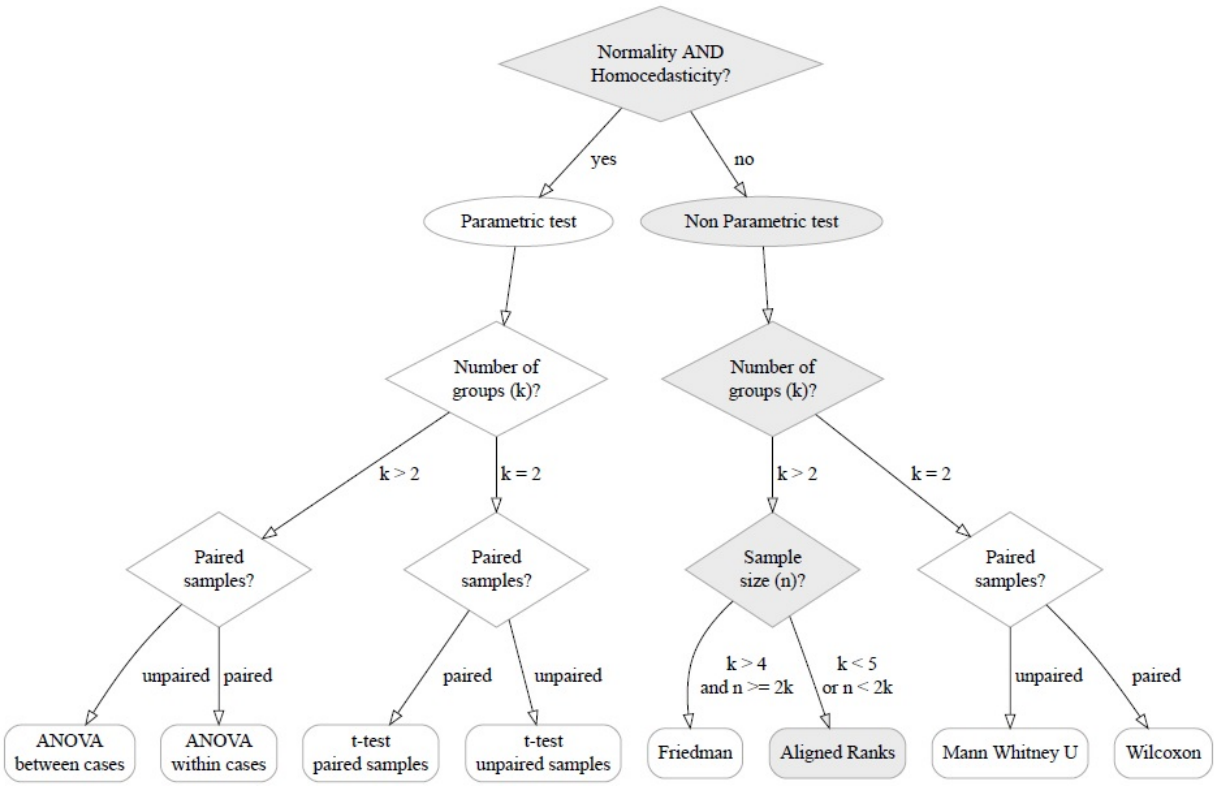
3.5 Avaliação de modelos e análise de resultados

As métricas mencionadas na seção 2.4 avaliaram os modelos de classificação nas fases de treinamento e teste. Além disso, o método de validação cruzada descrito na seção 2.5 avaliou o desempenho dos modelos na fase de treinamento. A classe *IterativeStratification* do pacote *skmulti-learn.model_selection*, com $f = 10$ pastas, produziu a validação. O desempenho médio dos modelos é a média da soma dos desempenhos das pastas foram medidos em macro-média F1, micro-média F1 e *Hamming Loss*.

Os desempenhos dos modelos, para cada métrica, foram comparados na plataforma Web *Statistical Tests for Algorithms Comparison (STAC)* (RODRÍGUEZ-FDEZ et al., 2015). Os valores das métricas macro-média F1 e micro-média F1 tiveram os sinais invertidos, pois a plataforma STAC assume que quanto menor o resultado de um algoritmo num problema, melhor o desempenho desse algoritmo (RODRÍGUEZ-FDEZ et al., 2015). Os desempenhos dos modelos de treinamento por seleção de características, medido em cada pasta na validação cruzada, foram importados para plataforma através da janela *Import data*. Em seguida, a aplicação *Assistant* da plataforma STAC foi executada e gerou a Árvore de Decisão, com a indicação do teste mais adequado, como exposto na Figura 3.4. Assim, de acordo com os valores dos dados inseridos, a quantidade de modelos analisados ($k = 10$) e o tamanho da amostra ($n = f = 10$), foi estimado o teste estatístico de *Friedman Aligned Ranks* (HODGES J. L.AND LEHMANN, 2012) para cada métrica de avalição. O teste estimado foi realizado com nível de significância $\alpha = 0,05$, verificando a hipótese nula de que as médias dos resultados dos dez modelos são

as mesmas. Com a rejeição da hipótese nula, o teste *post hoc* de *Finner* (FINNER, 1993) foi realizado para comparações múltiplas com nível de significância $\alpha = 0,05$, para a hipótese nula de que a média dos resultados de cada par de algoritmos são iguais.

Figura 3.4 – Processo *Assistant* da plataforma STAC.



Fonte: Rodríguez-Fdez et al. (2015)

Em relação à avaliação da classificação do conjunto de teste, foram medidos na etapa 6 da Figura 3.1. O desempenho absoluto de cada modelo foi calculado utilizando a macro-média F1, micro-média F1 e *Hamming Loss*. O desempenho absoluto das classes feminino e masculino do rótulo gênero e das classes adulto e adolescente do rótulo faixa etária, também foram computadas. As matrizes de confusão extraídas pela função *multilabel_confusion_matrix* do pacote *sklearn.metrics* ajudam a calcular as métricas na fase de teste, em cada rótulo. Os valores das variáveis tp, fp e fn das matrizes, substituem as variáveis das métricas Precisão, Revocação e medida F1 das classes dos rótulos. Além disso, eles substituem as variáveis das métricas gerais de avaliação, micro-média P, micro-média R e micro-média F1.

4 AVALIAÇÃO EXPERIMENTAL E RESULTADOS

Neste capítulo serão apresentados as descrições da avaliação experimental e resultados dos modelos de classificação de gêneros e faixas etárias. A execução dos experimentos foram realizados em uma máquina com 6GB de memória RAM e um processador Intel Core i5-4200U 2.6GHz de quarta geração. Além disso, o sistema operacional instalado na máquina era o Ubuntu 18.04. Os tempos de execução dos experimentos de cada modelo na fase de treinamento estão no Apêndice A.

4.1 Avaliação dos modelos de classificação nos treinamentos

Os objetivos da avaliação do treinamento são validar a capacidade de generalização dos modelos de classificação e identificação do melhor modelo, usando o conjunto de treinamento. Nos experimentos, os subconjuntos de dados com os *k* melhores atributos tiveram seus desempenhos avaliados. A Tabela 4.1 mostra o desempenho médio em macro-média F1, micro-média F1 e *Hamming Loss*.

Tabela 4.1 – Avaliação da seleção de características baseada no valor k - Fase de treinamento

(a) Macro-média F1

k-value	CC-DTC	LP-DTC	CC-RFC	LP-RFC	CC-XGB	LP-XGB	CC-SVC	LP-SVC	CC-CNB	LP-CNB
9	**0,964**	**0,965**	**0,965**	0,964	**0,965**	0,964	0,820	**0,822**	**0,632**	0,629
8	**0,964**	**0,965**	**0,965**	**0,965**	0,964	0,964	**0,824**	0,821	0,598	0,620
7	**0,964**	**0,965**	0,964	0,964	**0,965**	**0,965**	0,818	**0,822**	0,597	**0,630**
6	**0,964**	**0,965**	**0,965**	0,964	**0,965**	**0,965**	0,775	0,782	0,598	0,600
5	**0,964**	**0,965**	**0,965**	0,964	**0,965**	**0,965**	0,734	0,740	0,598	0,602
4	0,703	0,702	0,709	0,694	0,701	0,699	0,703	0,702	0,605	0,604
3	0,550	0,597	0,598	0,597	0,611	0,598	0,546	0,597	0,519	0,589
2	0,550	0,586	0,586	0,586	0,611	0,553	0,550	0,550	0,502	0,571
1	0,337	0,586	0,586	0,586	0,611	0,586	0,319	0,586	0,303	0,274

(b) Micro-média F1

k-value	CC-DTC	LP-DTC	CC-RFC	LP-RFC	CC-XGB	LP-XGB	CC-SVC	LP-SVC	CC-CNB	LP-CNB
9	**0,959**	**0,959**	**0,959**	**0,959**	0,959	0,958	0,822	**0,823**	**0,633**	0,630
8	**0,959**	**0,959**	**0,959**	**0,959**	0,959	0,958	**0,825**	**0,823**	0,596	0,622
7	**0,959**	**0,959**	**0,959**	**0,959**	0,959	**0,960**	0,820	**0,823**	0,596	**0,633**
6	**0,959**	**0,959**	**0,959**	**0,959**	**0,960**	0,959	0,779	0,786	0,597	0,600
5	**0,959**	**0,959**	**0,959**	**0,959**	**0,960**	0,959	0,737	0,743	0,597	0,603
4	0,709	0,708	0,711	0,698	0,709	0,705	0,709	0,708	0,605	0,604
3	0,603	0,596	0,597	0,596	0,613	0,597	0,602	0,596	0,526	0,587
2	0,603	0,585	0,585	0,585	0,613	0,601	0,603	0,603	0,519	0,570
1	0,539	0,585	0,585	0,585	0,613	0,585	0,479	0,585	0,420	0,400

(c) *Hamming loss*

k-value	CC-DTC	LP-DTC	CC-RFC	LP-RFC	CC-XGB	LP-XGB	CC-SVC	LP-SVC	CC-CNB	LP-CNB
9	**0,037**	**0,037**	0,037	**0,037**	**0,036**	0,038	0,167	0,166	**0,344**	**0,357**
8	**0,037**	**0,037**	0,037	**0,037**	0,037	0,038	**0,163**	**0,165**	0,378	0,382
7	**0,037**	**0,037**	0,037	**0,037**	**0,036**	**0,036**	0,169	**0,165**	0,379	0,382
6	**0,037**	**0,037**	**0,036**	**0,037**	**0,036**	0,037	0,208	0,199	0,378	0,404
5	**0,037**	**0,037**	0,037	**0,037**	**0,036**	0,037	0,253	0,248	0,378	0,406
4	0,272	0,274	0,280	0,325	0,276	0,275	0,272	0,273	0,346	0,409
3	0,374	0,379	0,378	0,379	0,558	0,378	0,373	0,379	0,370	0,389
2	0,374	0,391	0,391	0,391	0,558	0,376	0,374	0,374	0,351	0,406
1	0,435	0,391	0,391	0,391	0,558	0,391	0,396	0,391	0,357	0.565

Os melhores resultados de cada modelo, em cada métrica, estão em negrito.

Fonte: Autor (2020)

Os resultados dos experimentos demonstram que os modelos LP-CNB e CC-CNB, baseados no teorema de Bayes, alcançaram as melhores médias em cada métrica, quando $k \in \{7,9\}$. Os modelos CC-SVC e LP-SVC, baseados em SVM, atingiram as melhores médias em cada métrica, quando $k \in \{7,8,9\}$. Em relação aos modelos CC-DTC, LP-DTC, CC-RFC, LP-RFC, CC-XGB e LP-XGB, alcançaram as melhores médias, quando $k \in \{5,6,7,8,9\}$. O fato dos últimos modelos terem uma etapa de seleção de característica interna durante a sua construção, justifica o excelente desempenho em vários subconjuntos. Portanto, infere-se que os sete atributos selecionados nos subconjuntos de $k = 5$ obtêm os melhores padrões de classificação.

Os desempenhos dos modelos com os melhores subconjuntos de características de cada algoritmo foram comparados pelo teste estatístico não paramétrico de *Friedman Aligned Ranks* e pelo teste *post hoc* de *Finner*. As Tabelas 4.2, 4.3, 4.4 e 4.5 mostram a comparação do desempenho médio de cada métrica para os subconjuntos de $k = 9$, $k = 8$, $k = 7$ e $k = 5$, respectivamente.

O teste de *Friedman Aligned Ranks* mostrou que o desempenho médio em cada métrica, difere entre os modelos de $k = 9$ (Tabela 4.2), com $p\text{-}value < 0,001$. O desempenho médio dos modelos CC-DTC, LP-DTC, CC-RFC, LP-RFC, CC-XGB e LP-XGB foram similares em relação ao teste de *Finner* em cada métrica, e superiores ao desempenho de CC-CNB e LP-CNB. O modelo LP-XGB teve um empate estatístico no desempenho médio com os modelos CC-SVC e LP-SVC, nas métricas macro-média F1 e *Hamming Loss*. Também, as médias dos modelos CC-SVC, LP-SVC, CC-CNB e LP-CNB tiveram resultados iguais pelo teste de *Finner* em cada métrica.

Tabela 4.2 – Desempenho médio dos modelos na classificação com $k = 9$ - Fase de treinamento

Modelo	Macro-média F1	Micro-média F1	*Hamming Loss*
LP-DTC	0,965 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
CC-DTC	0,964 ± 0,002 a	0,959 ± 0,002 a	0,037 ± 0,002 a
LP-RFC	0,964 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
CC-RFC	0,965 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
LP-XGB	0,964 ± 0,001 a b	0,958 ± 0,001 a	0,038 ± 0,001 a b
CC-XGB	0,965 ± 0,001 a	0,959 ± 0,001 a	0,036 ± 0,001 a
LP-SVC	0,822 ± 0,003 b c	0,823 ± 0,003 b	0,166 ± 0,004 b c
CC-SVC	0,820 ± 0,002 b c	0,822 ± 0,002 b	0,167 ± 0,003 b c
LP-CNB	0,629 ± 0,004 c	0,630 ± 0,004 b	0,357 ± 0,003 c
CC-CNB	0,632 ± 0,001 c	0,633 ± 0,001 b	0,344 ± 0,001 c

As médias com as mesmas letras em uma coluna não diferem no nível de probabilidade de 5% pelo teste de *Finner*. Os melhores desempenhos em cada métrica estão com a letra (a).

Fonte: Autor (2020)

Nos experimentos para $k = 8$ (Tabela 4.3), o teste de *Friedman Aligned Ranks* mostrou que o desempenho médio em cada métrica difere entre os modelos, com $p\text{-}value < 0,001$. Os modelos CC-DTC, LP-DTC, CC-RFC, LP-RFC, CC-XGB e LP-XGB têm as mesmas médias para cada métrica, com desempenhos superiores aos dos modelos CC-SVC, LP-SVC, CC-CNB e LP-CNB. Além disso, esses últimos modelos empataram pelo teste de *Finner* em cada métrica.

Tabela 4.3 – Desempenho médio dos modelos na classificação com $k = 8$ - Fase de treinamento

Modelo	Macro-média F1	Micro-média F1	*Hamming Loss*
LP-DTC	0,965 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
CC-DTC	0,964 ± 0,002 a	0,959 ± 0,002 a	0,037 ± 0,002 a
LP-RFC	0,965 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
CC-RFC	0,965 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
LP-XGB	0,964 ± 0,002 a	0,958 ± 0,002 a	0,038 ± 0,002 a
CC-XGB	0,964 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
LP-SVC	0,821 ± 0,002 b	0,823 ± 0,002 b	0,165 ± 0,002 b
CC-SVC	0,824 ± 0,003 b	0,825 ± 0,003 b	0,163 ± 0,003 b
LP-CNB	0,620 ± 0,002 b	0,622 ± 0,002 b	0,382 ± 0,002 b
CC-CNB	0,598 ± 0,001 b	0,596 ± 0,001 b	0,378 ± 0,001 b

As médias com as mesmas letras em uma coluna não diferem no nível de probabilidade de 5% pelo teste de *Finner*. Os melhores desempenhos em cada métrica estão com a letra (a).

Fonte: Autor (2020)

Nos experimentos com $k = 7$ (Tabela 4.4), o teste de *Friedman Aligned Ranks* mostrou que o desempenho médio em cada métrica difere entre os modelos de classificação com $p\text{-}value < 0,001$. As médias de desempenho dos modelos CC-DTC, LP-DTC, CC-RFC, LP-RFC, CC-XGB e LP-XGB empataram em cada métrica pelo teste de *Finner*, mas são superiores aos desempenhos médios dos modelos CC-SVC, LP-SVC, CC-CNB e LP-CNB. Além disso, esses últimos modelos apresentam as mesmas médias, segundo o teste de *Finner*, nas métricas macro-

média F1 e *Hamming Loss*. Na micro-média F1, o desempenho médio do modelo LP-SVC é superior estatisticamente ao desempenho médio do modelo CC-CNB.

Tabela 4.4 – Desempenho médio dos modelos na classificação com $k = 7$ - Fase de treinamento

Modelo	Macro-média F1	Micro-média F1	*Hamming Loss*
LP-DTC	0,965 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
CC-DTC	0,964 ± 0,002 a	0,959 ± 0,002 a	0,037 ± 0,002 a
LP-RFC	0,964 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
CC-RFC	0,964 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
LP-XGB	0,965 ± 0,001 a	0,960 ± 0,002 a	0,036 ± 0,001 a
CC-XGB	0,965 ± 0,002 a	0,959 ± 0,002 a	0,036 ± 0,001 a
LP-SVC	0,822 ± 0,002 b	0,823 ± 0,002 b	0,165 ± 0,002 b
CC-SVC	0,818 ± 0,004 b	0,820 ± 0,003 b c	0,169 ± 0,004 b
LP-CNB	0,630 ± 0,004 b	0,633 ± 0,004 b c	0,382 ± 0,004 b
CC-CNB	0,597 ± 0,001 b	0,596 ± 0,001 c	0,379 ± 0,001 b

As médias com as mesmas letras em uma coluna não diferem no nível de probabilidade de 5% pelo teste de *Finner*. Os melhores desempenhos em cada métrica estão com a letra (a).

Fonte: Autor (2020)

Também nos experimentos com $k = 5$ (Tabela 4.5), o teste de *Friedman Aligned Ranks* mostrou que o desempenho médio em cada métrica difere entre os modelos de classificação com $p\text{-}value < 0,001$. As médias de desempenho dos modelos CC-DTC, LP-DTC, CC-RFC, LP-RFC, CC-XGB e LP-XGB empataram em cada métrica pelo teste de *Finner*, mas são superiores aos desempenhos médios dos modelos CC-SVC, LP-SVC, CC-CNB e LP-CNB. Esses últimos modelos empataram estatisticamente na macro-média F1. Porém, o desempenho médio do modelo LP-SVC é superior ao do modelo CC-CNB na micro-média F1, e superior ao modelo LP-CNB na métrica *Hamming Loss*.

Tabela 4.5 – Desempenho médio dos modelos na classificação com $k = 5$ - Fase de treinamento

Modelo	Macro-média F1	Micro-média F1	*Hamming Loss*
LP-DTC	0,965 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
CC-DTC	0,964 ± 0,002 a	0,959 ± 0,002 a	0,037 ± 0,002 a
LP-RFC	0,964 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
CC-RFC	0,965 ± 0,001 a	0,959 ± 0,001 a	0,037 ± 0,001 a
LP-XGB	0,965 ± 0,001 a	0,959 ± 0,002 a	0,037 ± 0,002 a
CC-XGB	0,965 ± 0,001 a	0,960 ± 0,001 a	0,036 ± 0,001 a
LP-SVC	0,740 ± 0,002 b	0,743 ± 0,002 b	0,248 ± 0,002 b
CC-SVC	0,734 ± 0,003 b	0,737 ± 0,003 b c	0,253 ± 0,003 b c
LP-CNB	0,602 ± 0,011 b	0,603 ± 0,012 b c	0,406 ± 0,005 c
CC-CNB	0,598 ± 0,001 b	0,597 ± 0,001 c	0,378 ± 0,001 b c

As médias com as mesmas letras em uma coluna não diferem no nível de probabilidade de 5% pelo teste de *Finner*. Os melhores desempenhos em cada métrica estão com a letra (a).

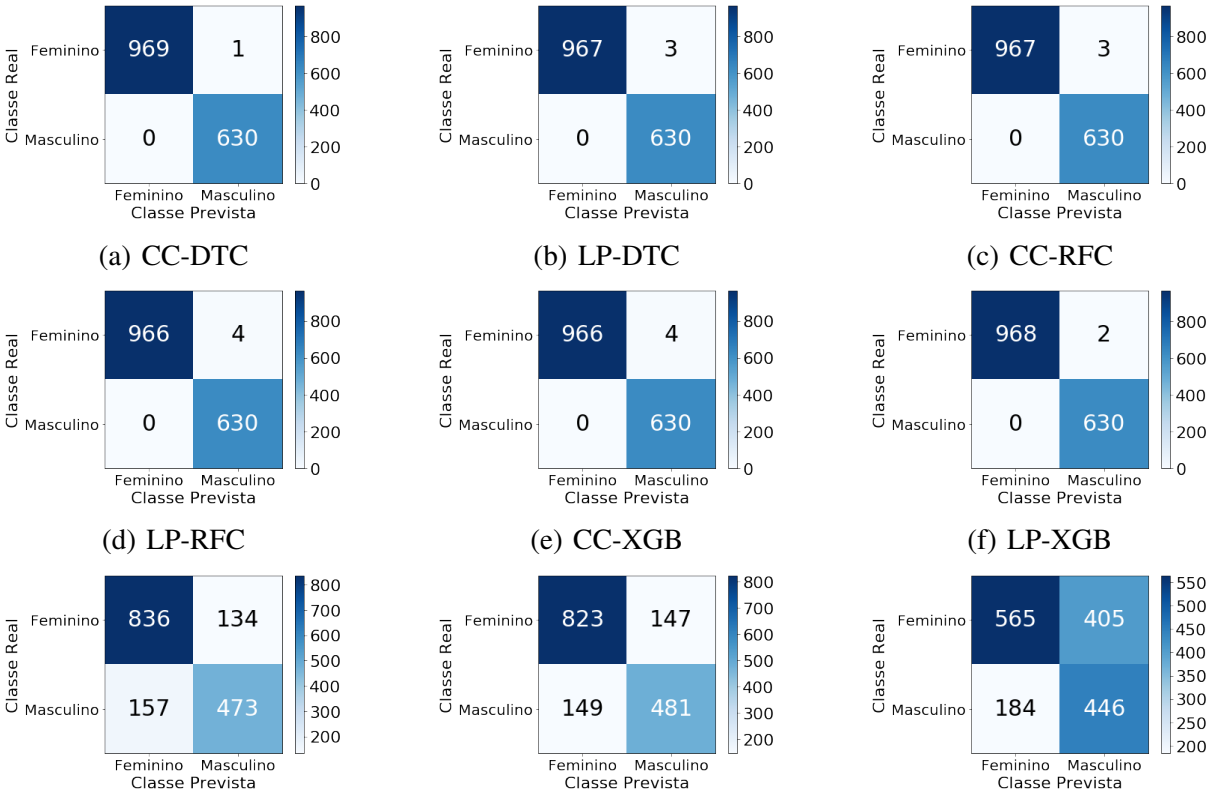
Fonte: Autor (2020)

4.2 Avaliação dos modelos de classificação nos testes

A avaliação de testes dos modelos de classificação mede e compara os desempenhos dos modelos, com dados de teste. Essa avaliação mediu o desempenho dos modelos com os melhores subconjuntos de características de cada algoritmo, definida na avaliação de treinamento, usando as matrizes de confusão e as métricas: Precisão, Revocação, F1, macro-média F1, micro-média F1 e *Hamming Loss*. O modelo LP-CNB com $k = 9$ foi escolhido por ter obtido o melhor resultado na métrica *Hamming Loss* na avaliação do treinamento, e por ter resultados com uma diferença insignificante nas métricas macro-média F1 e micro-média F1, quando comparados com os resultados de $k = 7$. E o desempenho do modelo LP-SVC com $k = 7$ foi escolhido por ser o menor subconjunto com bons resultados na avaliação do treinamento.

A Figura 4.1 apresenta as matrizes de confusão da classificação de gêneros, por modelo. A primeira diagonal contém os exemplos previstos corretamente. Essas matrizes foram geradas por funções do pacote *matplotlib.pyplot* (HUNTER, 2007).

Figura 4.1 – Matrizes de confusão da classificação de gênero - Fase de teste



(a) CC-DTC

(b) LP-DTC

(c) CC-RFC

(d) LP-RFC

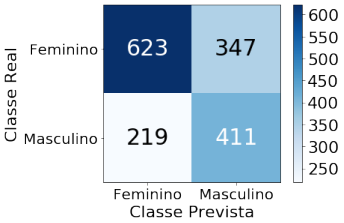
(e) CC-XGB

(f) LP-XGB

(g) CC-SVC

(h) LP-SVC

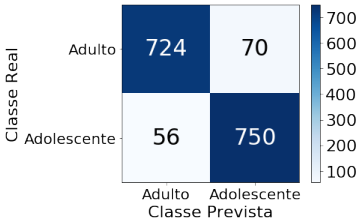
(i) CC-CNB



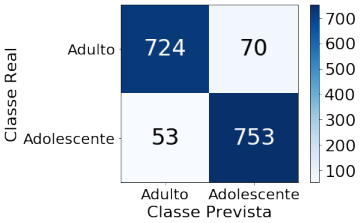
(j) LP-CNB
Fonte: Autor (2020)

A Figura 4.2 apresenta as matrizes de confusão da classificação das faixas etárias, por modelo. A primeira diagonal contém os exemplos previstos corretamente. Essas matrizes foram geradas por funções do pacote *matplotlib.pyplot* (HUNTER, 2007).

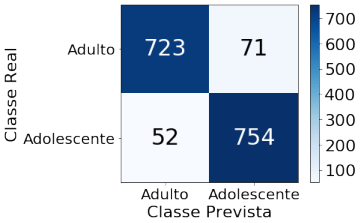
Figura 4.2 – Matrizes de confusão da classificação de faixa etária - Fase de teste



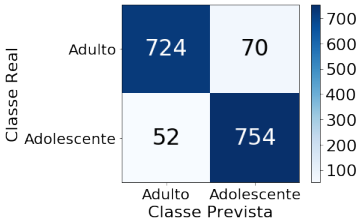
(a) CC-DTC



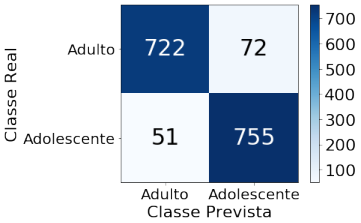
(b) LP-DTC



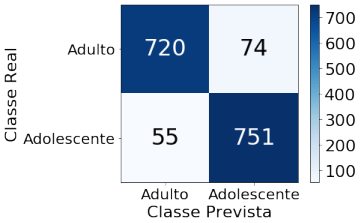
(c) CC-RFC



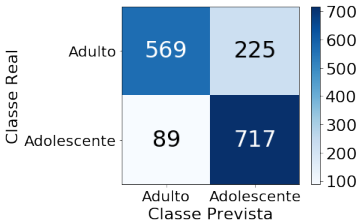
(d) LP-RFC



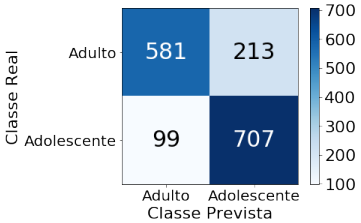
(e) CC-XGB



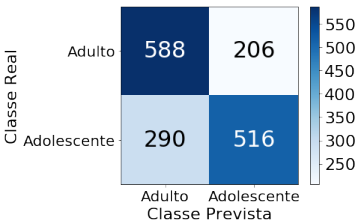
(f) LP-XGB



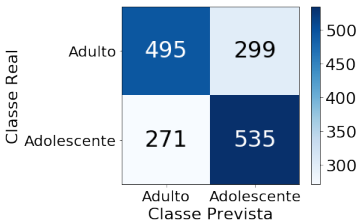
(g) CC-SVC



(h) LP-SVC



(i) CC-CNB



(j) LP-CNB
Fonte: Autor (2020)

As matrizes de confusão demonstram que os modelos de classificação erram mais na predição de faixas etárias, exceto o modelo CC-CNB, que errou mais na predição dos gêneros. Os modelos CC-DTC, LP-DTC, CC-RFC, LP-RFC, CC-XGB e LP-XGB, que usam regras de classificação, são os modelos mais assertivos.

A partir das matrizes de confusão, foi feito uma avaliação individual dos rótulos gênero e faixa etária. Nessa avaliação, desempenhos semelhantes à avaliação multidimensional de treinamento foram alcançados, confirmando o desempenho dos melhores modelos. A medida de comparação entre os modelos foi o valor geral da micro-média F1. As tabelas 4.6 e 4.7 apresentam os resultados.

Tabela 4.6 – Desempenho dos modelos na classificação dos gêneros - Fase de teste

Modelo	Precisão	Revocação	F1	Classe
	1.000	0,998	0,999	Feminino
CC-DTC	0,998	1.000	0,999	Masculino
	0,999	0,999	**0,999**	Geral
	1.000	0,996	0,998	Feminino
LP-DTC	0,995	1.000	0,997	Masculino
	0,998	0,998	0,998	Geral
	1.000	0,996	0,998	Feminino
CC-RFC	0,995	1.000	0,997	Masculino
	0,998	0,998	0,998	Geral
	1.000	0,995	0,997	Feminino
LP-RFC	0,995	1.000	0,997	Masculino
	0,997	0,997	0,997	Geral
	1.000	0,995	0,997	Feminino
CC-XGB	0,993	1.000	0,996	Masculino
	0,997	0,997	0,997	Geral
	1.000	0,997	0,998	Feminino
LP-XGB	0,996	1.000	0,998	Masculino
	0,998	0,998	0,998	Geral
	0,841	0,861	0,851	Feminino
CC-SVC	0,779	0,750	0,764	Masculino
	0,818	0,818	0,818	Geral
	0,846	0,848	0,847	Feminino
LP-SVC	0,765	0,763	0,764	Masculino
	0,815	0,815	0,815	Geral
	0,754	0,582	0,657	Feminino
CC-CNB	0,524	0,707	0,602	Masculino
	0,631	0,631	0,631	Geral
	0,739	0,642	0,687	Feminino
LP-CNB	0,542	0,652	0,592	Masculino
	0,646	0,646	0,646	Geral

O melhor desempenho de acordo com a micro-média F1 está em negrito.
Fonte: Autor (2020)

Os resultados apresentados na Tabela 4.6, referentes à classificação dos gêneros, indicam um empate entre os modelos de classificação CC-DTC, LP-DTC, CC-RDC, LP-RDC, CC-XGB e LP-XGB, mesmo com uma diferença máxima de dois milésimos entre eles. Assim, a superioridade do melhor para o pior modelo, segue a seguinte ordem: modelos que usam regras de classificação; CC-SVC; LP-SVC; LP-CNB; e CC-CNB.

Tabela 4.7 – Desempenho dos modelos na classificação das faixas etárias - Fase de teste

Algoritmo	Precisão	Revocação	F1	Classe
	0,928	0,911	0,919	Adulto
CC-DTC	0,914	0,930	0,922	Adolescente
	0,921	0,921	0,921	Geral
	0,931	0,911	0,921	Adulto
LP-DTC	0,914	0,934	0,924	Adolescente
	0,923	0,923	**0,923**	Geral
	0,932	0,910	0,921	Adulto
CC-RFC	0,913	0,935	0,924	Adolescente
	0,923	0,923	**0,923**	Geral
	0,932	0,911	0,922	Adulto
LP-RFC	0,915	0,935	0,925	Adolescente
	0,923	0,923	**0,923**	Geral
	0,934	0,909	0,921	Adulto
CC-XGB	0,912	0,936	0,924	Adolescente
	0,923	0,923	**0,923**	Geral
	0,929	0,906	0,917	Adulto
LP-XGB	0,910	0,931	0,920	Adolescente
	0,919	0,919	0,919	Geral
	0,864	0,716	0,783	Adulto
CC-SVC	0,761	0,889	0,820	Adolescente
	0,803	0,803	0,803	Geral
	0,854	0,731	0,788	Adulto
LP-SVC	0,768	0,877	0,819	Adolescente
	0,805	0,805	0,805	Geral
	0,669	0,740	0,703	Adulto
CC-CNB	0,714	0,640	0,675	Adolescente
	0,690	0,690	0,690	Geral
	0,646	0,623	0,634	Adulto
LP-CNB	0,641	0,663	0,652	Adolescente
	0,643	0,643	0,643	Geral

O melhor desempenho de acordo com a micro-média F1 está em negrito.

Fonte: Autor (2020)

Os resultados apresentados na Tabela 4.7, referentes à classificação das faixas etárias, também indicam um empate entre os modelos de classificação CC-DTC, LP-DTC, CC-RDC, LP-RDC, CC-XGB e LP-XGB. A diferença máxima entre os desempenhos dos modelos é de

quatro milésimos. Assim, a superioridade do melhor para o pior modelo, segue a seguinte ordem: modelos que usam regras de classificação; LP-SVC; CC-SVC; CC-CNB e LP-CNB.

A comparação do desempenho geral dos modelos em macro-média F1, micro-média F1 e *Hamming Loss*, também confirmou o desempenho dos melhores modelos. A Tabela 4.8 mostra o desempenho geral dos modelos na fase de teste, para cada métrica. Apesar da vantagem insignificante do desempenho do modelo CC-RDC na métrica macro-média F1, os modelos que usam regras de classificação conseguiram os mesmos desempenhos nas outras métricas e foram superiores aos modelos CC-SVC, LP-SVC, CC-CNB e LP-CNB.

Tabela 4.8 – Desempenho dos modelos de classificação - Fase de teste

Modelo	**Macro-média F1**	**Micro-média F1**	***Hamming Loss***
CC-DTC	0,960	**0,956**	**0,039**
LP-DTC	0,960	**0,956**	**0,039**
CC-RFC	**0,961**	**0,956**	**0,039**
LP-RFC	0,960	**0,956**	**0,039**
CC-XGB	0,960	**0,956**	**0,039**
LP-XGB	0,960	**0,956**	**0,039**
CC-SVC	0,793	0,797	0,189
LP-SVC	0,791	0,796	0,190
CC-CNB	0,639	0,639	0,339
LP-CNB	0,622	0,624	0,355

O melhor desempenho para cada métrica está em negrito.

Fonte: Autor (2020)

Por fim, a abordagem proposta foi validada em uma comparação dos resultados da classificação das faixas etárias deste estudo, com os resultados da classificação das faixas etárias obtidas em (GUIMARÃES et al., 2017), na Tabela 4.9.

Tabela 4.9 – Comparação dos resultados com os desempenhos de Guimarães et al. (2017) - Fase de Teste

Modelo	**Micro-média F1**
CC-DTC	0,921
LP-DTC	**0,923**
Decision Tree*	0,840
CC-RFC	**0,923**
LP-RFC	**0,923**
*Random Forest**	0,850
CC-SVC	0,803
LP-SVC	0,805
SVM*	**0,840**

* Nos modelos de Guimarães et al. (2017), o gênero está identificado.

O melhor desempenho de acordo com a micro-média F1 está em negrito.

Fonte: Autor (2020)

Os modelos CC-RFC, LP-RFC, CC-DTC e LP-DTC obtiveram desempenhos superiores aos modelos que usaram as mesmas abordagens de aprendizagem em (GUIMARÃES et al., 2017). O mesmo não aconteceu com os modelos CC-SVC e LP-SVC, pois alcançaram desempenhos inferiores. Porém, os modelos com os melhores desempenhos deste estudo não ultrapassaram o modelo baseado no *deep learning* de Guimarães et al. (2017), que atingiu 0,940 na micro-média F1.

4.3 Análise dos resultados

O aumento ou manutenção dos desempenhos dos modelos CC-DTC, LP-DTC, CC-RFC, LP-RFC, CC-XGB, LP-XGB, CC-SVC e LP-SVC, com a redução dos dados produzidos pela seleção das características, demonstra a importância desta tarefa. Os modelos que usam regras de classificação: CC-DTC, LP-DTC, CC-RFC, LP-RFC, CC-XGB e LP-XGB; obtiveram os melhores resultados com as características: rt, *at*, *hashtag*, *slang*, *punctuation*, *url*, e *characters*. O modelo CC-SVC obteve os melhores resultados com as características: rt, *at*, *hashtag*, *slang*, *punctuation*, url, *characters*, *follow*, *followers* e *topic*. E o modelo LP-SVC, obteve os melhores resultados com as características: rt, *at*, *hashtag*, *slang*, *punctuation*, url, *characters*, *follow*, *followers*. Na abordagem CC, o gênero compõe os subconjuntos selecionados na classificação de faixas etárias. Porém, neste trabalho, não foi avaliado o seu impacto na classificação, uma vez que as avaliações experimentais da inclusão do gênero na classificação de faixas etárias em Guimarães et al. (2017) foram positivas.

Os resultados dos experimentos demonstraram que os modelos CC-SVC, LP-SVC, CC-CNB e LP-CNB são inferiores aos modelos que usam regras de classificação. A única exceção foi o empate estatístico entre os modelos LP-XGB, CC-SVC e LP-SVC, no subconjunto com todos atributos. Uma possível justificativa da inferioridade dos modelos baseados no teorema de Bayes e no SVM, é que são mais eficazes em espaços de alta dimensão e classificação de texto, de acordo com a documentação do *Scikit-learn*. Outra possível justificativa é o maior número de configurações de parâmetros avaliados nos melhores modelos. No entanto, a definição do menor número de parâmetros nos modelos CC-SVC e LP-SVC, otimizou o tempo de busca pelos melhores parâmetros desses modelos. Além disso, os modelos CC-CNB e LP-CNB possuem poucos parâmetros, então o número de configurações é muito menor.

Por último, através da análise comparativa dos resultados da classificação de faixas etá-

rias, foi identificado a superioridade dos modelos que usam regras de classificação deste trabalho, em comparação com os modelos sobre as mesmas abordagens de classificação de Guimarães et al. (2017). A justificativa hipotética para este resultado é a busca pelos melhores parâmetros neste trabalho. Os modelos de Guimarães et al. (2017) mantiveram os valores padrões dos parâmetros. Por outro lado, a superioridade do modelo baseado em SVM de Guimarães et al. (2017), expôs a importância da classificação correta dos gêneros, antes de classificar as faixas etárias.

5 CONCLUSÃO

Este trabalho apresentou as etapas de construção e avaliação dos modelos de classificação de gêneros e faixa etárias. Os modelos classificaram exemplos de treinamento e teste, através de características de páginas dos perfis de usuários do *Twitter*. Os métodos de transformação *Classifier Chains (CC)* e *Label Powerset (LP)* foram usados em conjunto com algoritmos de classificação, para construir os modelos bidimensionais.

Os melhores modelos dos algoritmos *Decision Tree Classifier (DTC)*, *Random Forest Classifier (RFC)* e *Extreme Gradient Boosting (XGB)*, que utilizam regras de classificação, foram superiores aos dos algoritmos *Support Vector Classification (SVC)* e *Complement NB (CNB)*. Além disso, os melhores modelos baseados em regras foram construídos com os valores das seguintes características: *retweet (rt)*; arroba (at); *hashtag*; *slang*; *punctuation*; *url*; e *characters*. Essa redução de características otimiza a extração de dados da RSO em aplicações futuras.

Os bons desempenhos dos modelos em abordagens que usam regras de classificação, indica a eficácia da metodologia proposta. Na fase de teste, os modelos com DTC, RFC e XGB atingiram 0,956 de micro-média F1 e 0,039 de *Hamming Loss*. A maioria dos modelos que usam regras alcançou 0,960 de macro-média F1, exceto o modelo CC-RFC que alcançou 0,961. Na avaliação por rótulos, a superioridade dos modelos que usaram regras, foi mantido. Os desempenhos em micro-média F1 da classificação dos gêneros, foram: 0,999 no modelo CC-DTC; 0,998 nos modelos LP-DTC, CC-RFC e LP-XGB; e 0,997 nos modelos LP-RFC e CC-XGB. Em relação aos desempenhos na classificação das faixas etárias, foram: 0,923 nos modelos LP-DTC, CC-RFC, LP-RFC e CC-XGB; 0,921 no modelo CC-DTC; e 0,919 no modelo LP-XGB. Apesar dos bons desempenhos alcançados, nenhum modelo superou o *deep learning* na classificação de faixas etárias em (GUIMARÃES et al., 2017).

Desses resultados pode-se constatar três implicações: (1) A capacidade de generalização dos modelos; (2) A existência de padrões no conjunto de dados, capazes de distinguir gêneros e faixas etárias; e (3) A contribuição do excelente desempenho da classificação de gêneros na avaliação geral, pois, evitou a propagação de erros para classificação de faixas etárias.

Em trabalhos futuros, os experimentos de Guimarães et al. (2017) serão replicados com o conjunto de dados deste trabalho, para fazer uma nova comparação. Além disso, os modelos criados serão integrados com métricas de análise de sentimento, que fazem uso das caracterís-

ticas de gênero e faixa etária. Também, será aplicado e avaliado a normalização *min-max* na etapa de transformação da característica e avaliá-la. O objetivo será investigar o impacto dessas técnicas e melhorar o desempenho da classificação de gêneros e faixas etárias.

REFERÊNCIAS

ADENIRAN, A. O.; JADAH, H. M.; MOHAMMED, N. H. Impact of information technology on strategic management in the banking sector of Iraq. **Insights into Regional Development**, v. 2, n. 2, p. 592–601, June 2020. Disponível em: <https://ideas.repec.org/a/ssi/jouird/v2y2020i2p592-601.html>.

AL-GHADIR, A. I.; AZMI, A. M. A study of arabic social media users—posting behavior and author's gender prediction. **Cognitive Computation**, v. 11, n. 1, p. 71–86, Feb 2019. ISSN 1866-9964. Disponível em: <https://doi.org/10.1007/s12559-018-9592-7>.

ALHIJAWI, B.; HRIEZ, S.; AWAJAN, A. Text-based authorship identification - a survey. In: **2018 Fifth International Symposium on Innovation in Information and Communication Technology (ISIICT), October 31, 2018 - November 01, 2018**. Amman, Jordan: IEEE, 2018. v. 1, p. 1–7.

ALOWIBDI, J. S.; BUY, U. A.; YU, P. S. Language independent gender classification on twitter. In: ROKNE, J. G.; FALOUTSOS, C. (Ed.). **Advances in Social Networks Analysis and Mining 2013, ASONAM '13, August 25-29, 2013**. Niagara, ON, Canada: ACM, 2013a. p. 739–743. Disponível em: <https://doi.org/10.1145/2492517.2492632>.

ALOWIBDI, J. S.; BUY, U. A.; YU, P. S. Empirical evaluation of profile characteristics for gender classification on twitter. In: **12th International Conference on Machine Learning and Applications, ICMLA 2013, December 4-7, 2013**. Miami, FL, USA: IEEE, 2013b. v. 1, p. 365–369. Disponível em: <https://doi.org/10.1109/ICMLA.2013.74>.

ALSUKHNI, E.; ALEQUR, Q. Investigating the use of machine learning algorithms in detecting gender of the arabic tweet author. **International Journal of Advanced Computer Science and Applications**, The Science and Information Organization, v. 7, n. 7, 2016. Disponível em: <http://dx.doi.org/10.14569/IJACSA.2016.070746>.

ARAVANTINOU, C.; SIMAKI, V.; MPORAS, I.; MEGALOOIKONOMOU, V. Gender classification of web authors using feature selection and language models. In: RONZHIN, A.; POTAPOVA, R.; FAKOTAKIS, N. (Ed.). **Speech and Computer**. Cham: Springer International Publishing, 2015. p. 226–233. ISBN 978-3-319-23132-7.

ARGAMON, S.; KOPPEL, M.; PENNEBAKER, J. W.; SCHLER, J. Automatically profiling the author of an anonymous text. **Commun. ACM**, ACM, New York, NY, USA, v. 52, n. 2, p. 119–123, fev. 2009. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/1461928.1461959>.

ASGHAR, M. Z.; KUNDI, F. M.; AHMAD, S.; KHAN, A.; KHAN, F. K. T-SAF: twitter sentiment analysis framework using a hybrid classification scheme. **Expert Systems**, v. 35, n. 1, 2018. Disponível em: <https://doi.org/10.1111/exsy.12233>.

ASIM, M. N.; REHMAN, A.; SHOAIB, U. Accuracy based feature ranking metric for multi-label text classification. **International Journal of Advanced Computer Science and Applications**, v. 8, n. 10, p. 369–378, 2017.

BACCIU, A. et al. Bot and gender detection of twitter accounts using distortion and LSA. In: CAPPELLATO, L.; FERRO, N.; LOSADA, D. E.; MÜLLER, H. (Ed.). **Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019**. CEUR-WS.org, 2019. (CEUR Workshop Proceedings, v. 2380). Disponível em: <http://ceur-ws.org/Vol-2380/paper_210.pdf>.

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

BASSEM, B.; ZRIGUI, M. Gender identification: A comparative study of deep learning architectures. In: ABRAHAM, A.; CHERUKURI, A. K.; MELIN, P.; GANDHI, N. (Ed.). **Intelligent Systems Design and Applications**. Cham: Springer International Publishing, 2020. p. 792–800. ISBN 978-3-030-16660-1.

BAYOT, R.; GONÇALVES, T. Multilingual author profiling using word embedding averages and svms. In: **10th International Conference on Software, Knowledge, Information Management & Applications, SKIMA 2016, December 15-17, 2016**. Chengdu, China: IEEE, 2016. p. 382–386. Disponível em: <https://doi.org/10.1109/SKIMA.2016.7916251>.

BRIEDIENĖ, M.; KAPOČIUTĖ-DZIKIENĖ, J. An automatic author profiling from non-normative lithuanian texts. In: **Proc. 23rd Conference "Information Society and University Studies" (IVUS 2018)**. CEUR Workshop Proceedings, 2018. v. 2145, p. 99–105. Disponível em: <http://ceur-ws.org/Vol-2145/>.

BSIR, B.; ZRIGUI, M. An empirical method for evaluation of author profiling framework. In: **Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation, PACLIC 2017, November 16-18, 2017**. Cebu City, Philippines: The National University (Phillippines), 2017.

BSIR, B.; ZRIGUI, M. Bidirectional lstm for author gender identification. In: NGUYEN, N. T.; PIMENIDIS, E.; KHAN, Z.; TRAWIŃSKI, B. (Ed.). **Computational Collective Intelligence**. Cham: Springer International Publishing, 2018. p. 393–402. ISBN 978-3-319-98443-8.

BSIR, B.; ZRIGUI, M. Enhancing deep learning gender identification with gated recurrent units architecture in social text. **Computación y Sistemas**, v. 22, n. 3, 2018. Disponível em: <http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3036>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: KRISHNAPURAM, B. et al. (Ed.). **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016**. ACM, 2016. p. 785–794. Disponível em: <https://doi.org/10.1145/2939672.2939785>.

CHENG, N.; CHANDRAMOULI, R.; SUBBALAKSHMI, K. Author gender identification from text. **Digital Investigation**, v. 8, n. 1, p. 78 – 88, 2011. ISSN 1742-2876. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1742287611000247>.

CHENG, N.; CHEN, X.; CHANDRAMOULI, R.; SUBBALAKSHMI, K. P. Gender identification from e-mails. In: **Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009, part of the IEEE Symposium Series on**

Computational Intelligence 2009, March 30, 2009 - April 2, 2009. Nashville, TN, USA: IEEE, 2009. p. 154–158. Disponível em: <https://doi.org/10.1109/CIDM.2009.4938643>.

CIMINO, A.; DELL'ORLETTA, F. A hierarchical neural network approach for bots and gender profiling. In: CAPPELLATO, L.; FERRO, N.; LOSADA, D. E.; MÜLLER, H. (Ed.). **Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019**. CEUR-WS.org, 2019. (CEUR Workshop Proceedings, v. 2380). Disponível em: <http://ceur-ws.org/Vol-2380/paper_177.pdf>.

COMPANY, J. S.; WANNER, L. How to use less features and reach better performance in author gender identification. In: . Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 1315–1319. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/104_Paper.pdf>.

DONG, M.; MIHALCEA, R.; RADEV, D. Extending sparse text with induced domain-specific lexicons and embeddings: A case study on predicting donations. **Computer Speech & Language**, v. 59, p. 157 – 168, 2020. ISSN 0885-2308. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0885230818300123>.

DWIVEDI, V. P.; SINGH, D. K.; JHA, S.; RANVIJAY. Gender classification of blog authors: With feature engineering and deep learning using lstm networks. In: **2017 Ninth International Conference on Advanced Computing (ICoAC), December 14-16, 2017**. Chennai, India: IEEE, 2017. p. 142–148.

FINNER, H. On a monotonicity problem in step-down multiple test procedures. **Journal of the American statistical association**, Taylor & Francis Group, v. 88, n. 423, p. 920–923, 1993.

GUIMARÃES, R. G.; ROSA, R. L.; GAETANO, D. D.; RODRíGUEZ, D. Z.; BRESSAN, G. Age groups classification in social network using deep learning. **IEEE Access**, v. 5, p. 10805–10816, 2017. ISSN 2169-3536.

GUPTA, S.; KANCHINADAM, T.; CONATHAN, D.; FUNG, G. Task-optimized word embeddings for text classification representations. **Frontiers in Applied Mathematics and Statistics**, v. 5, p. 67, 2020. ISSN 2297-4687. Disponível em: <https://www.frontiersin.org/articles/10.3389/fams.2019.00067>.

HODGES J. L.AND LEHMANN, E. L. Rank methods for combination of independent experiments in analysis of variance. In: ____. **Selected Works of E. L. Lehmann**. Boston, MA: Springer US, 2012. p. 403–418. ISBN 978-1-4614-1412-4. Disponível em: <https://doi.org/10.1007/978-1-4614-1412-4_35>.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

ISBISTER, T.; KAATI, L.; COHEN, K. Gender classification with data independent features in multiple languages. In: BRYNIELSSON, J. (Ed.). **European Intelligence and Security Informatics Conference, EISIC 2017, September 11-13, 2017**. Athens, Greece: IEEE Computer Society, 2017. p. 54–60. Disponível em: <https://doi.org/10.1109/EISIC.2017.16>.

KIRATSA, P. I. et al. Gender identification through facebook data analysis using machine learning techniques. In: **Proceedings of the 22Nd Pan-Hellenic Conference on Informatics**. New York, NY, USA: ACM, 2018. (PCI '18), p. 117–120. ISBN 978-1-4503-6610-6. Disponível em: <http://doi.acm.org/10.1145/3291533.3291591>.

LI, D.; LI, Y.; JI, W. Gender identification via reposting behaviors in social media. **IEEE Access**, v. 6, p. 2879–2888, 2018. Disponível em: <https://doi.org/10.1109/ACCESS.2017.2785813>.

LI, Y.; YANG, L.; XU, B.; WANG, J.; LIN, H. Improving user attribute classification with text and social network attention. **Cogn. Comput.**, v. 11, n. 4, p. 459–468, 2019. Disponível em: <https://doi.org/10.1007/s12559-019-9624-y>.

LIN, J. **Automatic author profiling of online chat logs**. 2007. Tese (Doutorado) — Monterey, California. Naval Postgraduate School.

LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

LIU, H.; COCEA, M. Fuzzy rule based systems for gender classification from blog data. In: **Tenth International Conference on Advanced Computational Intelligence, ICACI 2018, March 29-31, 2018**. Xiamen, China: IEEE, 2018. p. 79–84. Disponível em: <https://doi.org/10.1109/ICACI.2018.8377585>.

MARKOV, I.; GÓMEZ-ADORNO, H.; POSADAS-DURÁN, J.-P.; SIDOROV, G.; GELBUKH, A. Author profiling with doc2vec neural network-based document embeddings. In: PICHARDO-LAGUNAS, O.; MIRANDA-JIMÉNEZ, S. (Ed.). **Advances in Soft Computing**. Cham: Springer International Publishing, 2017. p. 117–131. ISBN 978-3-319-62428-0.

MARQUARDT, J. et al. Age and gender identification in social media. In: CAPPELLATO, L.; FERRO, N.; HALVEY, M.; KRAAIJ, W. (Ed.). **Working Notes for CLEF 2014 Conference, September 15-18, 2014**. Sheffield, UK: CEUR-WS.org, 2014. (CEUR Workshop Proceedings, v. 1180), p. 1129–1136. Disponível em: <http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-MarquardtEt2014.pdf>.

MARTINC, M.; SKRJANEC, I.; ZUPAN, K.; POLLAK, S. PAN 2017: Author profiling - gender and language variety prediction. In: CAPPELLATO, L.; FERRO, N.; GOEURIOT, L.; MANDL, T. (Ed.). **Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, September 11-14, 2017**. Dublin, Ireland: CEUR-WS.org, 2017. (CEUR Workshop Proceedings, v. 1866). Disponível em: <http://ceur-ws.org/Vol-1866/paper_78.pdf>.

MASLENNIKOVA, A.; LABRUNA, P.; CIMINO, A.; DELL'ORLETTA, F. Quanti anni hai? age identification for italian. In: BERNARDI, R.; NAVIGLI, R.; SEMERARO, G. (Ed.). **Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019**. CEUR-WS.org, 2019. (CEUR Workshop Proceedings, v. 2481). Disponível em: <http://ceur-ws.org/Vol-2481/paper43.pdf>.

MODARESI, P.; LIEBECK, M.; CONRAD, S. Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. In: BALOG, K.; CAPPELLATO, L.; FERRO, N.; MACDONALD, C. (Ed.). **Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum**. Évora, Portugal: CEUR-WS.org, 2016. (CEUR Workshop Proceedings, v. 1609), p. 970–977. Disponível em: <http://ceur-ws.org/Vol-1609/16090970.pdf>.

MORGAN-LOPEZ, A. A.; KIM, A. E.; CHEW, R. F.; RUDDLE, P. Predicting age groups of twitter users based on language and metadata features. **PLOS ONE**, Public Library of Science, v. 12, n. 8, p. 1–12, 08 2017. Disponível em: <https://doi.org/10.1371/journal.pone.0183537>.

MUKHERJEE, A.; LIU, B. Improving gender classification of blog authors. In: **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (EMNLP '10), p. 207–217. Disponível em: <http://dl.acm.org/citation.cfm?id=1870658.1870679>.

NGUYEN, D.; GRAVEL, R.; TRIESCHNIGG, D.; MEDER, T. "how old do you think I am?"A study of language and age in twitter. In: KICIMAN, E.; ELLISON, N. B.; HOGAN, B.; RESNICK, P.; SOBOROFF, I. (Ed.). **Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013**. Cambridge, Massachusetts, USA: The AAAI Press, 2013. p. 439–448. ISBN 978-1-57735-610-3. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/5984>.

PANDYA, A.; OUSSALAH, M.; MONACHESI, P.; KOSTAKOS, P.; LOVÉN, L. On the use of urls and hashtags in age prediction of twitter users. In: **2018 IEEE International Conference on Information Reuse and Integration, IRI 2018, Salt Lake City, UT, USA, July 6-9, 2018**. IEEE, 2018. p. 62–69. Disponível em: <https://doi.org/10.1109/IRI.2018.00017>.

PANDYA, A.; OUSSALAH, M.; MONACHESI, P.; KOSTAKOS, P. On the use of distributed semantics of tweet metadata for user age prediction. **Future Generation Computer Systems**, v. 102, p. 437 – 452, 2020. ISSN 0167-739X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X19304509>.

PARDO, F. M. R. et al. Overview of the 3rd author profiling task at PAN 2015. In: CAPPELLATO, L.; FERRO, N.; JONES, G. J. F.; SANJUAN, E. (Ed.). **Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum**. Toulouse, France: CEUR-WS.org, 2015. (CEUR Workshop Proceedings, v. 1391). Disponível em: <http://ceur-ws.org/Vol-1391/inv-pap12-CR.pdf>.

PARDO, F. M. R. et al. Overview of the author profiling task at PAN 2014. In: CAPPELLATO, L.; FERRO, N.; HALVEY, M.; KRAAIJ, W. (Ed.). **Working Notes for CLEF 2014 Conference**. Sheffield, UK: CEUR-WS.org, 2014. (CEUR Workshop Proceedings, v. 1180), p. 898–927. Disponível em: <http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-RangelEt2014.pdf>.

PARDO, F. M. R.; ROSSO, P.; POTTHAST, M.; STEIN, B. Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in twitter. In: CAPPELLATO, L.; FERRO, N.; GOEURIOT, L.; MANDL, T. (Ed.). **Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, 2017**. Dublin, Ireland: CEUR-WS.org, 2017. (CEUR Workshop Proceedings, v. 1866). Disponível em: <http://ceur-ws.org/Vol-1866/invited_paper_11.pdf>.

PARDO, F. M. R. et al. Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In: BALOG, K.; CAPPELLATO, L.; FERRO, N.; MACDONALD, C. (Ed.). **Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum**. Évora, Portugal: CEUR-WS.org, 2016. (CEUR Workshop Proceedings, v. 1609), p. 750–784. Disponível em: <http://ceur-ws.org/Vol-1609/16090750.pdf>.

PARK, G. et al. Women are warmer but no less assertive than men: Gender and language on facebook. **PLOS ONE**, Public Library of Science, v. 11, n. 5, p. 1–26, 05 2016. Disponível em: <https://doi.org/10.1371/journal.pone.0155885>.

PASCUCCI, A.; MASUCCI, V.; MONTI, J. Computational stylometry and machine learning for gender and age detection in cyberbullying texts. In: **8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACII Workshops 2019**. Cambridge, United Kingdom: IEEE, 2019. p. 1–6. Disponível em: <https://doi.org/10.1109/ACIIW.2019.8925101>.

PEARSON, F. K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor & Francis, v. 50, n. 302, p. 157–175, 1900. Disponível em: <https://doi.org/10.1080/14786440009463897>.

PEERSMAN, C.; DAELEMANS, W.; VAERENBERGH, L. V. Predicting age and gender in online social networks. In: **Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents**. New York, NY, USA: ACM, 2011. (SMUC '11), p. 37–44. ISBN 978-1-4503-0949-3. Disponível em: <http://doi.acm.org/10.1145/2065023.2065035>.

PENTEL, A. Effect of different feature types on age based classification of short texts. In: **2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)**. Corfu, Greece: IEEE, 2015. p. 1–7.

PEREIRA, R. B.; PLASTINO, A.; ZADROZNY, B.; MERSCHMANN, L. Correlation analysis of performance measures for multi-label classification. **Information Processing & Management**, v. 54, n. 3, p. 359 – 369, 2018. ISSN 0306-4573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457318300165>.

READ, J.; MARTINO, L.; LUENGO, D. Efficient monte carlo methods for multi-dimensional learning with classifier chains. **Pattern Recogn.**, Elsevier Science Inc., USA, v. 47, n. 3, p. 1535–1546, mar. 2014. ISSN 0031-3203. Disponível em: <https://doi.org/10.1016/j.patcog.2013.10.006>.

RENNIE, J. D. M.; SHIH, L.; TEEVAN, J.; KARGER, D. R. Tackling the poor assumptions of naive bayes text classifiers. In: **Proceedings of the Twentieth International Conference on International Conference on Machine Learning**. AAAI Press, 2003. (ICML'03), p. 616–623. ISBN 1-57735-189-4. Disponível em: <http://dl.acm.org/citation.cfm?id=3041838.3041916>.

RODRÍGUEZ-FDEZ, I.; CANOSA, A.; MUCIENTES, M.; BUGARÍN, A. STAC: a web platform for the comparison of algorithms using statistical tests. In: **2015 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2015**. Istanbul,

Turkey: IEEE, 2015. p. 1–8. ISBN 978-1-4673-7428-6. Disponível em: <https://doi.org/10.1109/FUZZ-IEEE.2015.7337889>.

SCHLER, J.; KOPPEL, M.; ARGAMON, S.; PENNEBAKER, J. W. Effects of age and gender on blogging. In: **AAAI spring symposium: Computational approaches to analyzing weblogs**. Stanford, California, USA: AAAI, 2006. v. 6, p. 199–205.

SCHWARTZ, H. A. et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. **PLOS ONE**, Public Library of Science, v. 8, n. 9, p. 1–16, 09 2013. Disponível em: <https://doi.org/10.1371/journal.pone.0073791>.

SILESSI, S.; VAROL, C.; KARABATAK, M. Identifying gender from sms text messages. In: **2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)**. Anaheim, CA, USA: IEEE Computer Society, 2016. p. 488–491.

SIMAKI, V.; ARAVANTINOU, C.; MPORAS, I.; MEGALOOIKONOMOU, V. Using sociolinguistic inspired features for gender classification of web authors. In: KRÁL, P.; MATOUŠEK, V. (Ed.). **Text, Speech, and Dialogue**. Cham: Springer International Publishing, 2015. p. 587–594. ISBN 978-3-319-24033-6.

SINHA, A.; SINHA, A. K. An empirical model of gender based human trait identification from blogs. In: **2015 IEEE International Conference on Computational Intelligence & Communication Technology**. Ghaziabad, India: IEEE, 2015. p. 694–698.

SOLER-COMPANY, J.; WANNER, L. On the role of syntactic dependencies and discourse relations for author and gender identification. **Pattern Recognition Letters**, v. 105, p. 87 – 95, 2018. ISSN 0167-8655. Machine Learning and Applications in Artificial Intelligence. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167865517304440>.

SURENDRAN, K.; HARILAL, O. P.; HRUDYA, P.; POORNACHANDRAN, P.; SUCHETHA, N. K. Stylometry detection using deep learning. In: BEHERA, H. S.; MOHAPATRA, D. P. (Ed.). **Computational Intelligence in Data Mining**. Singapore: Springer Singapore, 2017. p. 749–757. ISBN 978-981-10-3874-7.

TAM, J.; MARTELL, C. H. Age detection in chat. In: **Proceedings of the 2009 IEEE International Conference on Semantic Computing**. USA: IEEE Computer Society, 2009. p. 33–39. ISBN 9780769538006.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining, (First Edition)**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. Amsterdam: Morgan Kaufmann, 2011. (Morgan Kaufmann Series in Data Management Systems). ISBN 978-0-12-374856-0. Disponível em: <http://www.sciencedirect.com/science/book/9780123748560>.

WU, C. et al. Neural gender prediction in microblogging with emotion-aware user representation. In: **Proceedings of the 28th ACM International Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2019. (CIKM '19), p. 2401–2404. ISBN 9781450369763. Disponível em: <https://doi.org/10.1145/3357384.3358077>.

ZHANG, M.-L.; ZHOU, Z.-H. A review on multi-label learning algorithms. **IEEE Transactions on Knowledge and Data Engineering**, v. 26, n. 8, p. 1819–1837, Aug 2014. ISSN 1041-4347.

ZHANG, Y.; DANG, Y.; CHEN, H. Gender classification for web forums. **IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans**, v. 41, n. 4, p. 668–677, July 2011. ISSN 1083-4427.

ZHENG, Y.; LI, L.; ZHANG, J.; XIE, Q.; ZHONG, L. Using sentiment representation learning to enhance gender classification for user profiling. In: SHAO, J. et al. (Ed.). **Web and Big Data - Third International Joint Conference, APWeb-WAIM 2019, Chengdu, China, August 1-3, 2019, Proceedings, Part II**. Springer, 2019. (Lecture Notes in Computer Science, v. 11642), p. 3–11. Disponível em: <https://doi.org/10.1007/978-3-030-26075-0_1>.

ZHONG, X. et al. An emotion classification algorithm based on spt-capsnet. **Neural Computing and Applications**, v. 32, n. 7, p. 1823–1837, Nov. 2020. Disponível em: <https://doi.org/10.1007/s00521-019-04621-y>.

APÊNDICE A – Tempo de execução do experimento na fase de treinamento

Tabela A.1 – Tempo de execução do experimento de cada modelo em segundos - Fase de treinamento

k-value	CC-DTC	LP-DTC	CC-RFC	LP-RFC	CC-XGB	LP-XGB	CC-SVC	LP-SVC	CC-CNB	LP-CNB
9	1	2	32	19	6	5	55	20	1	1
8	1	1	28	4	3	93	43	44	1	1
7	1	1	5	4	3	5	58	46	1	1
6	1	1	22	13	5	71	32	61	1	1
5	1	3	5	17	5	459	40	112	1	3
4	1	3	6	3	3	2	19	31	1	3
3	1	3	1	2	4	63	18	30	1	3
2	1	3	1	2	1	110	16	29	1	3
1	1	3	2	2	3	2	39	12	2	3

Fonte: Autor (2020)