Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Ciência da Computação

Caracterização de usuários no Instagram

Rodrigo Ribeiro Oliveira

Feira de Santana

2022

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Ciência da Computação

Rodrigo Ribeiro Oliveira

Caracterização de usuários no Instagram

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: João B. Rocha-Junior

Feira de Santana

2022

Ficha Catalográfica – Biblioteca Central Julieta Carteado

Oliveira, Rodrigo Ribeiro
O51c Caracterização de usuários no Instagram./ Rodrigo Ribeiro Oliveira. Feira de Santana, 2021.
117f.: il.

Orientador: João B. Rocha-Junior
Dissertação (mestrado) – Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Ciência da Computação, 2021.

1.Redes sociais. 2.Inteligência artificial. 3.Caracterização de usuários. 4Aprendizagem de máquina. I.Rocha-Junior, João B., orient. II.Universidade Estadual de Feira de Santana. III. Titulo.

CDU : 004.91

Rodrigo Ribeiro Oliveira

Caracterização de usuários no Instagram

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Feira de Santana, 16 de julho de 2021

BANCA EXAMINADORA

João Batista da Rocha Junior (Orientador(a))
Universidade Estadual de Feira de Santana

Rosalvo Ferreira de Oliveira Neto
Universidade Federal do Vale do São Francisco

Angelo Conrado Loula
Universidade Estadual de Feira de Santana

Abstract

With the popularization of social media, more and more data is created, generating new opportunities of extracting knowledge from it. An example of social media that became popular in recent years is Instagram, whose focus is image sharing. For marketing purposes, the characterization of users is a very important task, because it allows to deliver specific advertisements for each group of users. This approach allows for applications in marketing, pointing to users within an intended demography. This problem is tackled in this work, in particular, the determination of age range and professional area in Instagram users that are native speakers of Portuguese. Two datasets of Instagram profiles were built, one labeled with the age range and another with the professional area of the users. The classifiers Random Forest and Support Vector Machines were used for determining these characteristics, through textual and behavioral attributes. The best results achieved have a accuracy of 60%, performance superior to the baseline for each problem.

Keywords: machine learning, social media, user characterization, artificial intelligence

Resumo

Com o advento e popularização de redes sociais, cada vez mais dados são gerados a partir delas, ensejando oportunidades de obtenção de conhecimento útil. Uma dessas redes é o Instagram, voltada para o compartilhamento de imagens. Um dos ramos de análise que pode ser realizada em redes sociais é a descoberta de características de usuários. Esta abordagem possibilita aplicações na área publicitária, indicando quais os usuários que estejam dentro de uma demografia que se queira alcançar, cobrindo uma área para a qual o Instagram não fornece ferramentas. Este trabalho se volta para este problema, tratando da caracterização de área profissional e faixa etária de perfis do Instagram, cujos usuários são falantes da língua portuguesa, onde há uma carência de trabalhos relacionados. Para isso, dois conjuntos de dados com perfis de usuários do Instagram, que são falantes da língua portuguesa, foram construídos, um deles rotulado com faixa etária e outro com área profissional. Foi realizada a classificação dessas características usando os classificadores *Random Forest* e *Support Vector Machines*, através de atributos textuais e comportamentais. Os resultados alcançados chegam a uma acurácia de cerca de 60%, com desepenho acima do *baseline*.

Palavras-chave: redes sociais, aprendizagem de máquina, inteligência artificial, caracterização de usuário

Prefácio

Esta dissertação de mestrado foi submetida à Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

A dissertação foi desenvolvida no Programa de Pós-Graduação em Ciência da Computação (PGCC), tendo como orientador o Prof. Dr. **João B. Rocha-Junior**.

Agradecimentos

Agradeço a Deus, criador e sustentador de todas as coisas.

À minha mãe, Suely, que me amou e me apoiou na aquisição de minha educação.

À minha família, pelo incentivo ao longo do tempo.

Aos meus amigos que me apoiaram desde quando decidi iniciar o programa de mestrado até a finalização deste trabalho.

Ao professor João, por ter aceitado ter sido meu orientador e auxiliado ao longo deste trabalho.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Os últimos anos trouxeram uma série de mudanças no uso da internet. Cada vez mais pessoas passam grande parte de seu tempo na rede, produzindo e consumindo conteúdo de natureza variada (Kaplan and Haenlein, 2010; Smith and Anderson, 2018; Anderson et al., 2018). Essa atividade se dá principalmente em redes sociais, o que faz esse conteúdo refletir as preferências e características de seus autores (Song et al., 2018). Logo, surge o interesse em obter informações sobre os autores a partir do conteúdo produzido. Essa situação traz novos temas de pesquisa a explorar, tratando da relação entre autoria e conteúdo na internet.

Trabalhos na área de detecção de autoria começaram a ser realizados no fim do século XIX, tratando da atribuição de textos à autores através de uma análise estatística unitária de atributos como tamanho de frases e número de palavras usadas apenas uma única vez nos textos, que se supunha que teria uma curva distinta para cada autor (Mendenhall, 1887). Essa abordagem deu lugar à outra dos anos 60 em diante, em que diversos atributos, como a frequência de certas palavras escolhidas, eram utilizados em métodos estatísticos bayesianos (Mosteller and Wallace, 1963) e análise de componentes principais. As últimas décadas trouxeram o uso de métodos de aprendizagem de máquina para a análise de autoria, levando a avanços significativos na área (Hoorn et al., 1999; Argamon et al., 2009).

Na era digital, os trabalhos passaram a aproveitar novos contextos e elucidar a relação entre conteúdos vindos da internet e seus respectivos autores. Uma das questões levantadas é a caracterização de autoria, que consiste na descoberta das características do autor, como por exemplo sexo, idade, profissão ou localização geográfica (Argamon et al., 2009; Koppel et al., 2009). Esta variedade de dimensões a explorar possibilita maior quantidade de trabalhos a realizar, aproveitando a ampla gama de classes do autor a analisar.

A passagem para o âmbito da internet trouxe o uso de novas fontes de dados: *e-mails* e *blogs* (Estival et al., 2007; Álvarez-Carmona et al., 2016), e posteriormente outros trabalhos foram realizados voltados para redes sociais, como Twitter (Filho

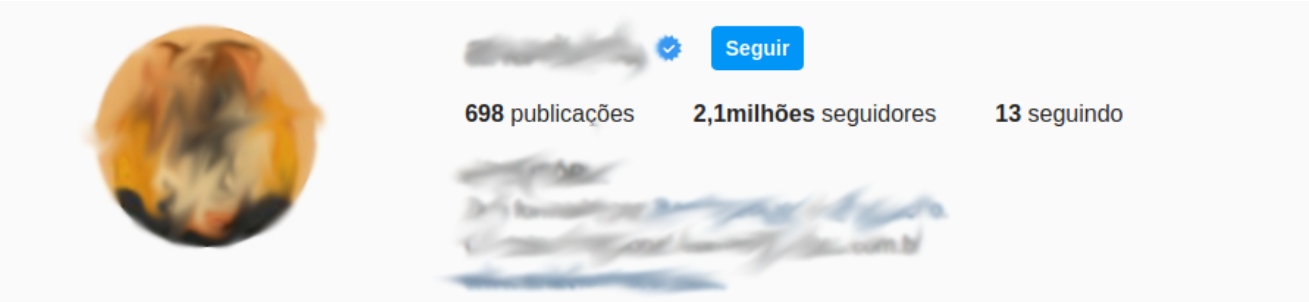


Figura 1.1: Exemplo de perfil do Instagram. Os dados sensíveis foram ocultados.

et al., 2014, 2016). Nestas, dados não são gerados apenas numa grande quantidade, mas também em numa ampla variedade de formatos, com relações entre conteúdo e autoria potencialmente mais complexas para ser investigadas.

Uma dessas redes sociais é o Instagram, lançado em 2010 com o foco no compartilhamento de imagens. Nela, os usuários podem, além de publicar imagens, deixar curtidas e comentários nas publicações, além de seguir uns aos outros para receber atualizações das publicações mais recentes. A Figura 1.1 contém o exemplo de um perfil do Instagram. O perfil é identificado por um nome de usuário único, e o usuário pode escolher uma foto de perfil como identificação adicional. Ficam disponíveis publicamente o número de usuários que o perfil segue e o número de seguidores que ela possui, bem como o número total de publicações. O usuário pode escolher se suas publicações ficarão visíveis para todos ou apenas para os seguidores que ele aprovar. Além disso, ele pode escolher uma biografia contendo uma descrição curta do perfil.

O Instagram alcançou grande número de usuários em alguns anos, sendo usado por 72% dos adolescentes americanos e 35% dos adultos desse mesmo país em 2018 segundo pesquisa do Pew Research Institute (Anderson et al., 2018; Smith and Anderson, 2018). No Brasil, o Instagram possuía 66 milhões de usuários em 2018, apresentando o maior crescimento entre redes sociais naquele ano (Valiati et al., 2020).

Esta popularização chamou a atenção de empresas, publicitários, figuras públicas e comerciantes, que buscam aproveitar o Instagram como meio de se conectar com seu público-alvo de modo mais direto e descontraído, seja através de anúncios comprados na plataforma, seja através de publicações que tenham como objetivo chamar a atenção dos usuários. A efetividade dessas estratégias depende de identificar quais usuários pertençam ao público-alvo de interesse, direcionando os anúncios e publicações para eles. Na esteira disso, logo surgiram trabalhos tratando de autoria nesta rede, explorando novos aspectos específicos dela (Song et al., 2018; Jang et al., 2015). Tais aspectos envolvem, além do texto publicado, as interações sociais entre os usuários e o conteúdo das imagens publicadas.

Trabalhos publicados em língua portuguesa sobre o Instagram reconhecem o poten-

cial de negócios da plataforma e examinam relações entre o conteúdo e os autores, mas na maioria das vezes realizam abordagens qualitativas ou se voltam para outras questões além da caracterização de autoria. Por exemplo, Oliveira (2014) reconhece que os usuários utilizam o Instagram para construir uma identidade própria e que esta é de interesse de esforços publicitários, mas se limita a estudar iniciativas das próprias empresas. Aragão et al. (2016) analisa a relação entre as curtidas e comentários de usuários do Instagram e as suas opções de compras, ressaltando as oportunidades empresariais e afirmando que *"no Instagram, a interação dos consumidores presentes na mídia com as marcas é 58 vezes maior que no Facebook e 120 vezes maior que no Twitter"*. Mas o faz de modo qualitativo, e obtendo os dados sobre o comportamento dos usuários através de *surveys* e não por uma extração direta do Instagram.

Trabalhos anteriores na área realizaram a caracterização de usuários no Instagram quanto a idade (Han et al., 2016; Song et al., 2018) e descoberta de contas comerciais (Campos, 2016). Atributos oriundos de texto, imagens e comportamento dos usuários são utilizados para a tarefa, com bons resultados. Este cenário revela que faltam trabalhos utilizando perfis de falantes de lingua portuguesa na caracterização de usuário no Instagram, bem como avaliando dimensões além da idade e sexo, evidenciando uma lacuna no estado da pesquisa atual.

Este trabalho, logo, vai ao encontro de tais necessidades, realizando uma análise numa base de dados de usuários falantes de língua portuguesa. É feita a caracterização de usuários no Instagram quanto à sua faixa etária e área profissional. Os atributos extraídos são utilizados por métodos de aprendizagem de máquina a fim de determinar a faixa etária e área profissional dos usuários. Além disso, este trabalho explora uma dimensão de interesse pouco analisada na literatura, a saber, a área profissional em que os usuários atuam.

1.1 Justificativa

Com o advento dos métodos de aprendizagem de máquina e produção de grande quantidade de dados na internet, resolver problemas no âmbito de autoria de modo eficiente e preciso tornou-se um campo de pesquisa fértil. Os primeiros trabalhos na área de caracterização de autoria estavam primariamente focados na descoberta de autoria de texto (Estival et al., 2007; Argamon et al., 2009). Com o advento das redes sociais, e o aumento na variedade do tipo de conteúdo, outras fontes (imagens, áudio, vídeos, *geodata*) passaram a ser utilizadas para determinar o perfil de indivíduos em redes sociais, como Twitter (Filho et al., 2016, 2014) e Instagram (Song et al., 2018; Zhang et al., 2016).

Esta preferência é causada pelo aumento no número de usuários de redes sociais ao longo dos anos, conforme indicado pelo gráfico na Figura 1.2, extraída de um levantamento do *Pew Research Center*, que mostra a porcentagem de americanos usuários de diversas redes sociais, de 2012 a 2018. A tendência é de aumento para

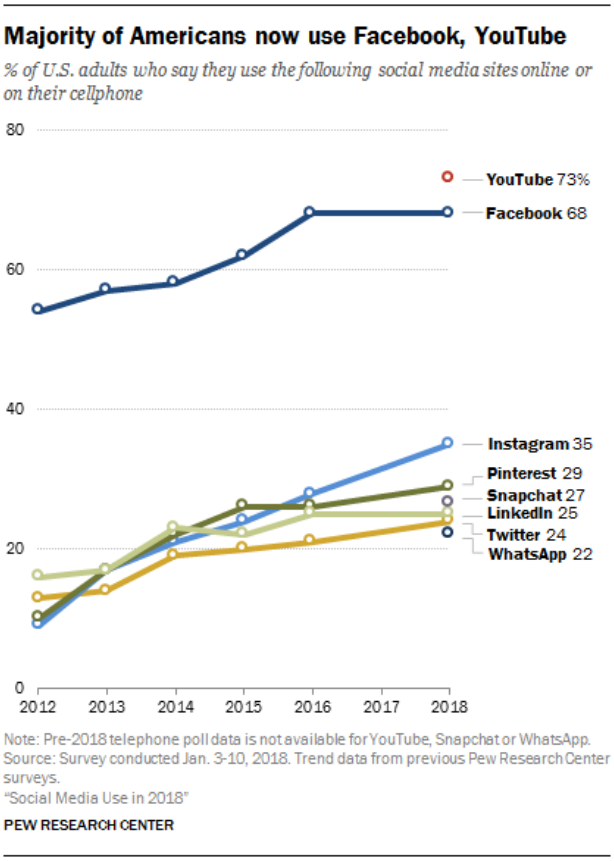


Figura 1.2: Gráfico contendo a evolução da porcentagem de adultos nos EUA presentes em diversas redes sociais.

todas as redes sociais analisadas, com o Instagram tornando-se, em 2018, a segunda rede social mais popular entre os americanos.

A popularização das redes sociais também impulsionou o seu uso para propaganda. Um levantamento feito pelo Influencer Marketing Hub em 2020 [1], composto de 4000 entrevistas feitas com empresas e pessoas na área de *marketing*, mostrou que 91% dos participantes consideravam *marketing* feito através de influenciadores digitais. Além disso, 78% afirmou que em 2020 reservaria um orçamento para esse tipo de *marketing*. Esse uso de redes sociais para *marketing*, assim, é uma aplicação que motiva a criação de trabalhos em caracterização de autoria, com o fim de encontrar usuários dentro do público-alvo a ser alcançado.

Idade e gênero são as características de usuário mais estudadas, nos trabalhos voltados para o Instagram (Souza et al., 2015; Han et al., 2016; Zhang et al., 2016; Song et al., 2018). Outros aspectos dos indivíduos, como profissão, são pouco explorados nesses trabalhos. Além disso, a escassez de pesquisas voltados para conteúdos em língua portuguesa impede saber se os grupos de usuários se comportam de modo

[1] https://influencermarketinghub.com/influencer-marketing-benchmark-report-2020/

diferente para linguagens diferentes, por exemplo, no vocabulário utilizado por cada faixa etária em inglês e português.

1.2 Motivação

A grande maioria dos esforços na caracterização de autoria tem sido realizada na detecção de idade e sexo, em falantes nativos de língua inglesa (Filho et al., 2014; Rodrigues et al., 2017). Este trabalho pretende ampliar o horizonte de pesquisa analisando material proveniente de falantes de língua portuguesa. Assim, um de seus objetivos é construir um conjunto de dados (*dataset*) nesse idioma, contendo usuários coletados do Instagram rotulados com sua faixa etária e área profissional.

Existem diversas aplicações práticas para esse tópico de pesquisa: na área forense, as informações que alguém dá acerca de si podem ser verificadas, de modo a detectar predadores sexuais e falsários (Rodrigues et al., 2017); na publicidade, conhecer melhor alvos de propagandas e ligar perfis a certos comportamentos, identificando potenciais compradores para certos produtos (Argamon et al., 2009); e pesquisas de opinião pública podem utilizar dados oriundos da internet para ampliar seu alcance. Alguns propuseram o uso de caracterização de autor como um auxílio adicional na detecção de plágio, através da busca de discrepâncias no perfil do autor (Mechti et al., 2013). Os resultados aqui alcançados poderão ser usados como substrato para estas aplicações.

1.3 Objetivos

1.3.1 Objetivo geral

Propor e avaliar o uso de métodos de aprendizagem de máquina para determinar a faixa etária e a área profissional de usuários do Instagram.

1.3.2 Objetivos específicos

- Estabelecer um conjunto de dados da rede social Instagram, com perfis de usuários cuja língua nativa é a portuguesa, rotulados com faixa etária e área profissional;
- Apresentar os conjuntos de dados obtidos em detalhes e extrair informações deles;
- Verificar quais os atributos mais discriminativos para a determinação da faixa etária e área profissional dos usuários de Instagram, dentre atributos comportamentais, textuais e de imagem.

1.4 Organização

Esse trabalho está organizado do seguinte modo: após a Introdução, é feita a Fundamentação (Capítulo 2) para a pesquisa a partir da bibliografia, ao que se sucede um capítulo tratando de como a Extração dos dados do Instagram (Capítulo 3) foi realizada. Então é feita uma descrição e análise do conjunto de dados obtido (Capítulo 4). Após isso, os Resultados (Capítulo 5) obtidos são expostos e discutidos. Por fim, são apresentadas as Considerações finais do trabalho (Capítulo 6).

Capítulo 2

Fundamentação

Nesta seção, são expostos temas relevantes para uma melhor compreensão deste trabalho. A Seção 2.1 aborda a aprendizagem de máquina; a Seção 2.2 apresenta o método de aprendizagem *Support Vector Machines*; a Seção 2.3 detalha os componentes do método *Random Forest* e seu funcionamento; a Seção 2.4 trata das medidas para avaliação do desempenho da aprendizagem de máquina; a Seção 2.5 explica o uso do método *cross-validation* na classificação; enquanto a Seção 2.6 explica com mais detalhes a caracterização de autoria. A Seção 2.7 apresenta o processo de extração de dados, suas etapas e as técnicas envolvidas, e por fim, os trabalhos relacionados são apresentados e descritos na Seção 2.8.

2.1 Aprendizagem de máquina

A aprendizagem de máquina (*machine learning*) é um campo dentro da Inteligência Artificial que envolve o estudo e a elaboração de algoritmos que possam, a partir de um conjunto de dados, funcionar automaticamente, sem interferência humana. O aprendizado pode ter como objetivo reconhecer padrões, tomar decisões ou realizar predições (Han et al., 2011). Ele pode ser dividido em dois tipos: supervisionado (Seção 2.1.1) e não-supervisionado (Seção 2.1.2).

2.1.1 Aprendizado supervisionado

O aprendizado supervisionado parte de um conjunto de entradas e saídas associadas entre si, para a criação de uma função que seja uma generalização dos casos desses conjuntos (Han et al., 2011; Skiena, 2017). Quando as saídas são dados discretos, a tarefa de aprendizado é chamada de *classificação*, com as saídas sendo consideradas como classes dos dados. Caso as saídas sejam contínuas, a tarefa é chamada de *predição*.

Esta forma de aprendizado envolve um processo chamado de treinamento. Baseado no conjunto de dados disponível (o conjunto de treinamento), é gerado um modelo,

que possui a capacidade de descobrir a saída de um dado a partir de sua entrada de maneira ótima. O modelo pode ser chamado de *classificador* ou preditor, a depender da tarefa sendo realizada (Han et al., 2011).

De maneira mais formal, havendo um conjunto X composto de N elementos $X_1, X_2, ..., X_N$, onde cada elemento é uma tupla de M atributos $A_1, A_2, ..., A_M$, e outro conjunto Y composto de N elementos $Y_1, Y_2, ..., Y_N$, existe uma relação $Y_i = f(X_i)$ entre cada elemento dos dois conjuntos. O treinamento envolve criar outra função $Y_i = \Phi(X_i)$ que generalize a relação original no conjunto de treino.

Uma vez feito o treinamento, um outro conjunto de dados, independente do de treino, pode ser submetido ao modelo, a fim de avaliar seu desempenho. Essa etapa é chamada de teste e os dados utilizados chamados de conjunto de teste (Han et al., 2011).

Um exemplo de aprendizado supervisionado é o diagnóstico automatizado de pacientes. A partir de um conjunto de exames feitos em diversos pacientes, os correspondentes diagnósticos são extraídos. Esses são usados no treinamento para obter um classificador que, recebendo como atributos os resultados de exames de um paciente, lhe emitirá o diagnóstico para alguma condição.

2.1.2 Aprendizado não-supervisionado

No aprendizado não-supervisionado, os dados utilizados não possuem uma saída "correta" associada à eles. O aprendizado deve buscar possíveis estruturas nos dados por conta própria e a partir disso proceder adequadamente.

Uma das abordagens de aprendizado não-supervisionado é o *clustering*: agrupar os dados em grupos (*clusters*) conforme similaridades e dissimilaridades encontradas entre eles. Outra é a associação, em que os dados são analisados a fim de descobrir regras que estejam presentes neles.

2.2 *Support Vector Machines*

As *Support Vector Machines* (SVM, Máquinas de Vetor Suporte) são um tipo de classificador supervisionado de caráter binário, ou seja, todos os dados são julgados como pertencentes à uma de duas classes possíveis. Uma SVM busca construir um modelo que separe os elementos das duas classes no espaço através de um *hiperplano*, maximizando a distância entre eles e portanto minimizando o erro do classificador (Cortes and Vapnik, 1995; Lorena and de Carvalho, 2007).

Nesta seção, primeiro o modo de funcionamento de uma SVM para o caso em que os dados são linearmente separáveis é descrito (Seção 2.2.1). Após isso, são explicadas os modos de lidar com dados que não são linearmente separáveis (Seção 2.2.2). Por fim, as estratégias para classificar dados com classes não-binárias são apresentados (Seção 2.2.3).

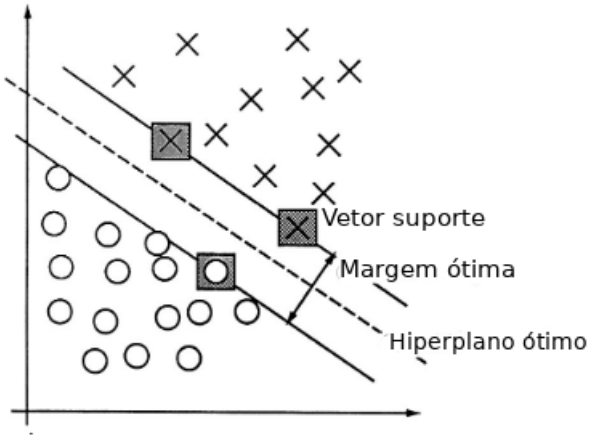


Figura 2.1: Exemplo de SVM para um conjunto de dados em duas dimensões, adaptado de Cortes and Vapnik (1995)

.

2.2.1 SVM para dados linearmente separáveis

A Figura 2.1 contém um exemplo de hiperplano gerado numa SVM para um conjunto de dados de duas dimensões. Um caso em que os dados podem ser separados através de um único hiperplano é chamado de linearmente separável.

Um hiperplano é o conjunto de todos os vetores X de n dimensões que satisfazem a Equação 2.1:

$$W \cdot X + b = 0 \tag{2.1}$$

Em que W é um vetor de pesos que é normal ao hiperplano considerado, e b é um deslocamento (chamado na literatura muitas vezes de *offset*) do hiperplano em direção ao vetor de pesos (Lorena and de Carvalho, 2007). Como o hiperplano divide os dados em duas regiões, é possível construir a seguinte relação (Equação 2.2):

$$g(x) = \begin{array}{l} +1 \text{ se } W \cdot X + b > 0 \\ -1 \text{ se } W \cdot X + b < 0 \end{array} \tag{2.2}$$

Existem infinitos hiperplanos que podem ser obtidos a partir da Equação 2.1. Da Equação 2.1, pode ser obtida a seguinte relação (Equação 2.3):

$$\mid W \cdot x_i + b \mid = 1 \tag{2.3}$$

Em que são considerados n casos x_1, x_2, ..., x_n mais próximos do hiperplano, associados aos rótulos y_1, y_2, ..., y_n correspondentes. Estes casos mais próximos do

hiperplano são os vetores suporte. Essa formulação leva às seguintes inequações (Equação 2.4):

$$\begin{array}{l} \text{se } y_i = 1,\ W \cdot x_i + b \geq 1 \\ \text{se } y_i = -1,\ W \cdot x_i + b \leq -1 \end{array} \tag{2.4}$$

Que podem ser resumidas em (Equação 2.5):

$$y_i(W \cdot x_i + b) - 1 \geq 0,\ \forall y_i, x_i \tag{2.5}$$

Na Figura 2.1, é possível ver elementos de cada classe mais próximos do hiperplano que estão destacados. Eles são os *vetores suporte* (de onde vem o nome do classificador). A distância entre os vetores suporte é chamada de *margem*. O problema da SVM consiste então em determinar o hiperplano que gere a maior margem possível, de modo que ela se torne a distância do hiperplano para o elemento mais próximo de cada classe, satisfazendo a Equação 2.5 (Han et al., 2011; Lorena and de Carvalho, 2007).

Tomando um ponto x_1 que está no lado positivo do hiperplano, e outro x_2 que está do lado negativo, com ambos satisfazendo a Equação 2.3, sob eles podem ser colocados hiperplanos adicionais, representados pelas Equações 2.6 e 2.7:

$$H_1 : W \cdot x_1 + b = 1 \tag{2.6}$$

$$H_2 : W \cdot x_2 + b = -1 \tag{2.7}$$

Projetando o vetor $x_2 - x_1$ no hiperplano ótimo, é obtida a seguinte formulação:

$$(x_1 - x_2)(\frac{w}{\|w\|} \cdot \frac{(x_1 - x_2)}{(\|x_1 - x_2\|)}) \tag{2.8}$$

Que pode ser reduzida a:

$$\frac{2}{\|w\|} \tag{2.9}$$

Correspondendo à distância entre H_1 e H_2. Assim, a distância entre o hiperplano ótimo e H_1 e H_2 é $\frac{1}{\|w\|}$. Essa expressão representa a margem, logo a margem pode ser maximizada através da minimização de $\|w\|$. O problema do treino de uma SVM pode ser expresso da seguinte forma:

$$\underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\| \tag{2.10}$$

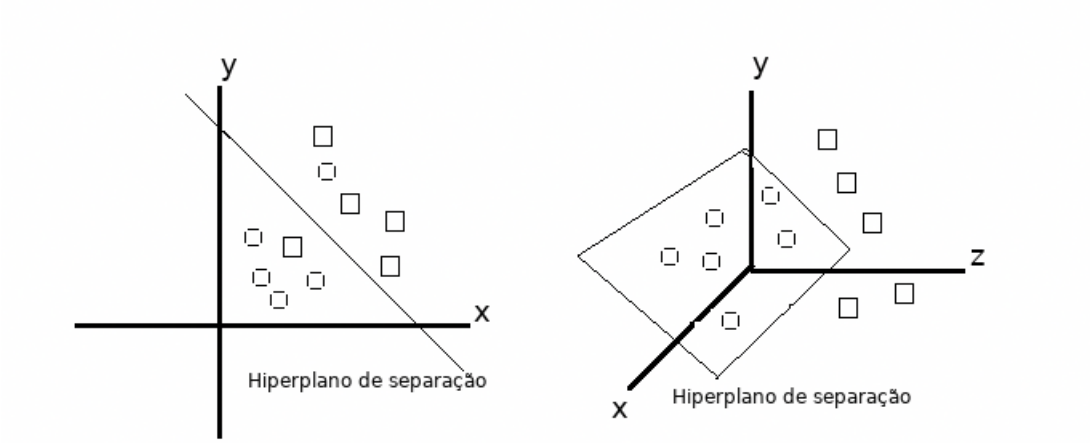


Figura 2.2: Exemplo de dados não linearmente separáveis em duas dimensões, e de uma transformação que os torna linearmente separáveis em três dimensões

.

O que é um problema de minimização quadrático. Esse tipo de problema pode ser resolvido através de uma função Lagrangiana (Han et al., 2011; Lorena and de Carvalho, 2007). Expor este tipo de técnica está fora do escopo desse trabalho, mas em linhas gerais ela envolve a descoberta de parâmetros α_i, chamados multiplicadores de Lagrange. Definidos esses parâmetros, a SVM pode ser expressa pela Equação 2.11:

$$d(X_T) = \sum_{i=1}^{n} y_i \alpha_i x_i X_T + b_0 \tag{2.11}$$

Em que n é o número de vetores suporte, x_i e y_i o i-ésimo vetor suporte e o rótulo associado, α_i e b_0 parâmetros definidos durante o processo de otimização e X_T uma entrada de teste. Os dados pós-treinamento são classificados inserindo-os nessa equação e verificando o sinal do resultado. Caso ele seja positivo, o dado recebe o rótulo +1, caso contrário, -1 (Han et al., 2011; Lorena and de Carvalho, 2007).

2.2.2 SVM para dados não linearmente separáveis

A formulação da SVM exposta na seção anterior não pode lidar com dados não linearmente separáveis, como os da Figura 2.2. Para resolver este problema, é feito um mapeamento do conjunto de dados, levando-o para um espaço de maior dimensionalidade, em que uma separação linear possa ser feita. Como exemplo, considere-se o aumento de duas para três dimensões através do mapeamento na Equação 2.12:

Kernel	Função
Polinomial	$X_i \cdot X_j + 1^h$
Gaussiano	$e^{\frac{-\|\|X_i - X_j\|\|^2}{2\sigma^2}}$

Tabela 2.1: Exemplos de funções de *kernel* usadas em SVMs.

$$\Phi(x) = \Phi(x1, x2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{2.12}$$

Que transforma o hiperplano na formulação da Equação 2.13:

$$f(x) = W \cdot \Phi(x) + b = w_1x_1^2 + w_2\sqrt{2}x_1x_2 + w_3x_2^2 = 0 \tag{2.13}$$

Assim, a formulação da SVM se torna a Equação 2.14:

$$d(X_T) = \sum_{i=1}^{n} y_i a_i \Phi(x_i)\Phi(X_T) + b_0 \tag{2.14}$$

Esta técnica pode se tornar demasiadamente custosa, caso a dimensão escolhida para a transição seja muito grande (Han et al., 2011; Lorena and de Carvalho, 2007). Para solucionar esse problema, atenta-se que na formulação da SVM, os termos mapeados se relacionam apenas através do cálculo do produto interno. Assim, a operação do mapeamento pode ser resumida apenas na substituição do produto interno de x_i e X_T.

O "truque" feito é a troca do produto interno $\Phi(x_i)\Phi(X_T)$ por uma função, chamada *kernel* (núcleo). Assim, a SVM pode ser treinada na dimensão original, e o uso da função de *kernel* garante que o uso para dados não linearmente separáveis (Han et al., 2011; Lorena and de Carvalho, 2007). A Tabela 2.1 apresenta exemplos de funções de kernel utilizadas em SVMs. No treinamento, o produto interno entre dois pontos no conjunto de dados é substituído pela função de *kernel*, permitindo comparar os dados de modo menos custoso do que fazendo o mapeamento do conjunto para uma dimensão maior.

2.2.3 Utilizando SVMs para casos de classes não binárias

Uma vez que SVMs são classificadores binários, é necessário que alguma técnica seja utilizada para conjuntos de dados em que há mais de duas classes. Existem dois modos de lidar com a situação: treinando um único classificador num esquema um-contra-todos ou vários classificadores combinados de modo um-contra-um. Essas estratégias serão explicadas a seguir.

Um-contra-todos

Na estratégia um-contra-todos, um único modelo é treinado para cada classe, comparando o resultado dele contra o de todas as outras classes. Assim, o número de modelos treinados é igual ao número de classes. A classificação submete o elemento de teste a todos os classificadores, combinando seus resultados para obter a classe final.

Um-contra-um

Nesse modo de classificação, é treinado um classificador para cada par de classes possíveis no conjunto de treino, assim, se existirem n classes, serão treinados $\frac{n(n+1)}{2}$ modelos. No teste, são feitas as classificações em todos esses modelos e a classe de maior frequência termina sendo a classe atribuída ao caso de teste (Han et al., 2011).

2.3 *Random Forest*

Random Forest (RF) é um tipo de classificador formado pela combinação de uma grande quantidade de árvores de decisão, um outro classificador (Liaw et al., 2002). Nesta seção o funcionamento do classificador é explicado, introduzindo os conceitos em que ele se baseia, as Árvores de decisão 2.3.1 e a técnica de *Bagging* 2.3.2, e expondo como eles são utilizados numa *Random Forest* 2.3.3.

2.3.1 Árvores de decisão

As árvores de decisão são classificadores que se baseiam na estrutura de dados árvore. Nelas, as folhas correspondem às classes em que os dados podem ser classificados (ou a probabilidades de pertencerem a essas classes). Os nós internos da árvore representam regras associadas com os atributos do conjunto de dados de treinamento. Frequentemente, as árvores utilizadas são binárias (Han et al., 2011).

A classificação de um dado de teste é realizada testando os valores em cada nó da árvore, e de acordo com o resultado, seguindo para o ramo correspondente até chegar numa folha. A classe representada pela folha é a classe atribuída no teste.

Nas seções seguintes, o funcionamento de árvores de decisão é explicado em detalhes.

Construindo uma árvore de decisão

Implementações de árvores de decisão adotam uma estratégia dividir-e-conquistar na construção da árvore (Han et al., 2011). O conjunto de dados de treinamento é recursivamente dividido ao longo do treino, conforme cada nó interno na árvore é determinado. Na divisão dos nós, busca-se maximizar alguma medida de desempe-

nho escolhida para escolher qual atributo será usado na regra associada ao nó em questão. Esse processo é descrito de maneira mais formal no Algoritmo 1.

Algoritmo 1: Gerar_árvore_de_decisão: Algoritmo para gerar uma árvore de decisão

Input: Conjunto de Dados D, lista de atributos l, função de seleção de atributos F

```
1  if D só possui elementos de uma classe then
2  |  return nó N com a classe dos elementos de D
3  end
4  if l é vazio then
5  |  return nó N com a classe de maior frequência em D
6  end
7  critério_de_divisão, atributo_de_divisão ← F(D,l);
8  lista_atributos ← l - atributo_de_divisão ;
9  foreach divisão C_j criada pelo critério_de_divisão do
10 |  if C_j é vazio then
11 |  |  adicionar um nó folha à N com a classe de maior frequência em D
12 |  else
13 |  |  adicionar Gerar_árvore_de_decisão(C_j, lista_atributos, F)
14 |  end
15 end
```

O Algoritmo 1 tem como argumentos um conjunto de dados, uma lista de atributos e um método de seleção de atributo. Ele conduz a basicamente três casos:

- Todos os casos no conjunto de dados pertencem à mesma classe, então é retornada uma árvore com um único nó folha, rotulado com a classe em questão;
- A lista de atributos é vazia, mas há elementos no conjunto de dados, resultando num nó folha rotulado com a classe de maior frequência no conjunto de dados;
- Quando a lista de atributos não é vazia e há mais de uma classe no conjunto de dados, é feita a seleção de um atributo utilizando a função de seleção de atributos. Esta função escolhe um atributo de modo a separar o conjunto de dados gerando sub-conjuntos que sejam o mais "puro" possíveis, ou seja, que possuam o máximo de elementos de apenas uma classe. A função é chamada recursivamente com esse sub-conjunto e com o atributo escolhido removido da lista de atributos.

Selecionando atributos na divisão de nós

Ao realizar uma divisão dentro da árvore de decisão durante sua construção, uma partição do conjunto de dados usado naquele nó é feita estabelecendo uma condição para um dos atributos considerados. Esta divisão deve ser feita de modo a garantir que cada um dos conjuntos gerados tenha o máximo de elementos de apenas uma

classe. De modo prático, ela envolve a utilização de alguma medida que ordene os atributos, selecionando o atributo que obteve melhor desempenho nessa medida na divisão. O atributo selecionado é utilizado numa regra de divisão do conjunto de dados.

Exemplos de medidas utilizadas na seleção de atributos são o ganho de informação e o índice de gini. O ganho de informação baseia-se na medida chamada entropia ou quantidade de informação, definida por Shannon (1948) no contexto da teoria da informação. A entropia indica o quão incerto é o estado de um certo evento, em bits. Considerando n eventos com probabilidades $p_1, ..., p_n$, a quantidade de informação é dada pela expressão:

$$Info = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.15}$$

No contexto de árvores de decisão, podemos dizer que considerando um nó N de uma árvore, em que o conjunto de dados de treino para N possui m classes C_i, com $i = 1, ...m$, a quantidade de informação é dada por (Han et al., 2011):

$$Info(D) = -\sum_{i=1}^{m} \frac{|C_{i,D}|}{|D|} \log_2 \frac{|C_{i,D}|}{|D|} \tag{2.16}$$

Com $|C_{i,D}|$ o número de elementos da classe C_i em D , e $|D|$ o número de elementos em D. Seja um atributo A com v valores possíveis de seleção para dividir o nó, de modo que D pode ser dividido em $D_1, ..., D_v$ conjuntos. A quantidade de informação de uma dessas partições dá uma noção do grau de impureza dela, sendo determinada por:

$$Info_A(D) = -\sum_{i=1}^{v} \frac{|D_i|}{|D|} Info(D) \tag{2.17}$$

Em que o termo $\frac{|D_i|}{|D|}$ é um peso da entropia de D, fazendo com que $Info_A(D)$ seja o grau de incerteza ou a quantidade de informação necessária na classificação de um elemento de D por A. Logo, o ganho de informação é dado por:

$$Ganho(A) = Info(D) - Info_A(D) \tag{2.18}$$

Essa fórmula indica o quanto de informação seria ganho ao dividir a árvore no nó N usando o atributo A. Logo, quão maior for o ganho de informação para A, mais adequado é ele para servir de critério para a divisão do nó N. Um dos tipos de árvore que utiliza o ganho de informação como medida para seleção de atributos é o algoritmo ID3.

A outra medida de seleção de atributo, o índice de Gini, é definida como (Han et al., 2011):

$$Gini(D) = 1 - \sum_{i=1}^{m} \frac{|C_{i,D}|}{|D|} \tag{2.19}$$

Utilizando a mesma notação introduzida na discussão sobre o ganho de informação: há um conjunto de dados D, cujos elementos pertecem a m classes, indicadas por $C_1, ..., C_m$. No algoritmo CART, o índice de Gini é utilizado para realizar divisões binárias considerando cada atributo. Realizando uma divisão de D em dois sub-conjuntos D_1 e D_2 usando o atributo A, o índice de Gini para essa partição será:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{2.20}$$

Sendo uma soma ponderada do índice de gini para cada sub-conjunto. Esse valor é calculado para cada uma das divisões possíveis em cada um dos atributos considerados. É escolhido o atributo que contenha a divisão que minimiza o valor do índice de Gini (Han et al., 2011).

2.3.2 *Bagging*

Bagging é um método que, através da criação de várias versões de um classificador, reune suas saídas individuais, com o objetivo de melhorar o desempenho original de classificação. Em cada iteração, é selecionada para treino uma fração do conjunto de dados aleatoriamente, havendo reposição entre as seleções (ou seja, instâncias do conjunto de dados podem aparecer em mais de uma iteração) Breiman (1996).

Considerando um conjunto de dados C, o *bagging* realiza N iterações, em que são criados N conjuntos $C_1, ..., C_N$ a partir de C de maneira aleatória, com reposição. Em cada iteração, um classificador é treinado com o conjunto de dados C_i associado, de modo que são gerados N modelos. Cada um dos C_i é obtido com reposição, um elemento de C pode não estar presente nele, e outro estar presente mais de uma vez.

Ao classificar um dado inédito através de *bagging*, todos esses classificadores são invocados num esquema de votação, com a saída de cada um deles considerada como um voto. A classe que receber mais votos é atribuída à instância sendo classificada.

Classificação por *bagging* tende a possuir uma *performance* superior àquela realizada com apenas um classificador (Han et al., 2011), pois a combinação de diversos modelos reduz a variância na classificação. Além disso, ela é mais robusta ao lidar com dados que tenham ruído.

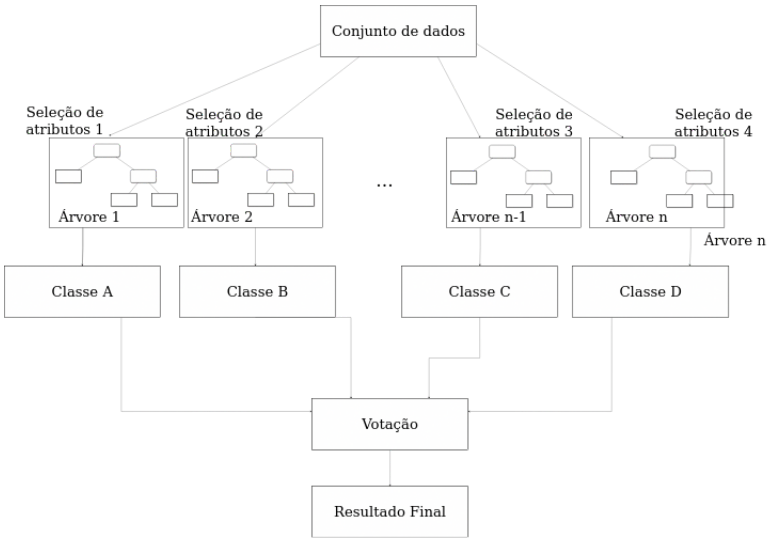


Figura 2.3: Ilustração do funcionamento de uma Random Forest.

2.3.3 Criando uma *Random Forest*

O *Random Forest* consiste basicamente da aplicação de *bagging* com árvores de decisão como o classificador usado. Para cada iteração, uma árvore diferente será treinada, utilizando um conjunto de dados distinto selecionado com reposição.

Contudo, no treinamento da árvore, uma modificação é feita na etapa de divisão dos nós: ao invés de considerar todos os atributos, é feita uma seleção aleatória deles, e dela será escolhido o atributo mais apto para a divisão de cada um dos nós da árvore (Breiman, 2001; Biau, 2012; Friedman et al., 2001). A Figura 2.3 contém uma ilustração do funcionamento de uma *Random Forest*. Os dados são enviados para cada uma das árvores de decisão treinadas, e as saídas de cada uma delas são combinadas para alcançar o resultado final.

Em Breiman (2001) é fornecida uma demonstração matemática de como esse método converge e o classificador resultante possui poder de generalização para outros conjuntos de dados. A randomização na seleção de atributos faz com que os resultados de cada árvore individual estejam menos correlacionados, evitando que o resultado final seja composto por votos concentrados numa única classe.

O número de árvores utilizado, bem como a quantidade de atributos selecionado na randomização, podem ser alterados em cada implementação. Outros modos de combinar a saída das árvores também podem ser utilizados, além de votação simples, como estabelecendo pesos para cada classificador.

2.4 Medidas de avaliação

Para avaliar o poder de predição de um classificador, comparando-o com outros classificadores e com um *baseline* (taxa base de desempenho), torna-se necessário

	C_1 (predito)	C_2 (predito)
C_1 (real)	verdadeiros positivos	falsos negativos
C_2 (real)	falsos positivos	verdadeiros negativos

Tabela 2.2: Matriz de confusão para um caso de classificação com duas classes.

calcular medidas que indiquem sua capacidade de previsão de maneira quantitativa. Quatro medidas são expostas aqui: a acurácia, a precisão, a sensibilidade e a medida F1, descritas nas seções 2.4.2, 2.4.3, 2.4.4 e 2.4.5, respectivamente.

2.4.1 Matriz de confusão

Uma matriz de confusão é uma maneira visual de analizar o desempenho de um classificador. Havendo m classes, ela tem o tamanho m por m, com as colunas indicando as classes preditas no teste, e as linhas as classes reais. Assim, cada elemento da matriz M_{ij} indica o número de casos de uma classe i classificados como da classe j.

A Tabela 2.2 contém uma matriz de confusão para um exemplo de matriz de confusão para duas classes. Pode-se abstrair a matriz em termos de se foi detectada a classe C_1 ou não, do que se fala de resultados positivos e negativos, respectivamente. Nos casos positivos que de fato pertencem à classe C_1, fala-se de *verdadeiros positivos*, se não pertencerem, são chamados de *falsos postitivos*. Analogamente, os casos negativos que não pertecem à classe C_1 são chamados de *verdadeiros negativos*, caso contrário, de *falsos negativos*.

Estes valores podem ser utilizados para medir o desempenho do classificador através de certas medidas. As próximas seções apresentam e detalham essas medidas.

2.4.2 Acurácia

A acurácia é uma das medidas mais simples em classificação, consistindo na taxa de previsões feitas corretamente (Han et al., 2011). Formalmente, pode ser definida pela Equação 2.2:

$$Acurácia = \frac{vp + vn}{vp + vn + fp + fn} \tag{2.21}$$

Em que vp corresponde a verdadeiros positivos, fp a falsos positivos, vn a verdadeiros negativos e fn a falsos negativos. Numa matriz de confusão, um classificador de boa acurácia terá os valores mais altos concentrados na diagonal.

2.4.3 Precisão

A precisão é dada pela Equação 2.3 (Han et al., 2011):

$$Precisão = \frac{vp}{vp + fp} \tag{2.22}$$

A precisão mede qual fração dos dados classificados como positivos eram de fato positivos. Ela é uma boa escolha para determinar a performance de um classificador num contexto em que falsos positivos devem ser evitados, pois uma quantidade excessiva de instâncias classificadas como positivas resultarão numa baixa precisão (Skiena, 2017).

2.4.4 Sensibilidade

A sensibilidade está descrita na Equação 2.4:

$$Sensibilidade = \frac{vp}{vp + fn} \tag{2.23}$$

Ela detecta que fração dos dados realmente positivos foram detectados pelo classificador, de modo que ele é adequado para uma situação em que falsos negativos são custosos (Skiena, 2017).

2.4.5 Medida F1

Medida F1 é uma medida que combina outras duas: precisão e sensibilidade, realizando a média harmônica entre elas.

A partir das Equações 2.3 e 2.4, a Medida F1 é dada, assim, pela Equação 2.5:

$$F = 2 \cdot \frac{\text{precisao} \cdot \text{sensibilidade}}{\text{precisao} + \text{sensibilidade}} \tag{2.24}$$

Essa medida, diferente da acurácia, não leva em conta apenas o quão o classificador acertou, mas fornece uma noção da relevância dos resultados obtidos. A média harmônica é utilizada pois resulta num valor menor que a média aritmética caso alguma das duas medidas seja muito menor que a outra, conforme explicado por Sasaki et al. (2007).

2.5 k-fold Cross-validation

Existe o interesse de obter uma medição mais confiável do desempenho de um classificador do que a fornecida através da utilização de uma única divisão entre conjunto de treino e teste. Para tal, torna-se necessário utilizar em seu treino diferentes combinações de dados, ampliando seu poder de generalização. O método *k-fold cross-validation* (Han et al., 2011; Friedman et al., 2001) é uma técnica que busca

realizar isso dividindo o conjunto de dados em k pedaços de tamanho igual e mutuamente exclusivos, com o processo de treino sendo feito k vezes. Em cada iteração, uma partição dos dados é usada como teste, e todas as outras como dados de treino para gerar o classificador. Uma média das métricas de avaliação em cada iteração é calculada ao final do processo.

Usualmente, os valores escolhidos para k são 5, 10 ou o próprio tamanho do conjunto de dados (N) (Friedman et al., 2001). Neste último caso, os modelos são treinados com todos as instâncias do conjunto de dados, exceto a i-ésima sendo considerada, o que é chamado de *leave-one-out*.

2.6 Investigação de autoria

Análises de autoria têm sido feitas desde o início do século XX, com o foco na determinação do autor de um texto de origem duvidosa ou desconhecida. Mosteller and Wallace (1963) é considerado um marco na utilização de métodos computacionais na investigação de autoria, buscando esclarecer a origem dos *Federalist Papers*, um conjunto de panfletos políticos americanos publicados de forma anônima on século XVIII. A virada do século trouxe uma diversificação das abordagens na área, indo além da simples atribuição de um autor a um texto. Koppel et al. (2009) lista quatro possíveis cenários de investigação de autoria:

- A *atribuição de autoria*, em que existe um conjunto bem determinado de autores candidatos para ser atribuído um dado conteúdo, como texto (tratado em Hoorn et al. (1999));
- A *verificação de autoria*, em que não há um conjunto de autores candidatos, mas um suspeito de ser o autor e a tarefa consiste em determinar se essa suspeita é verdadeira (um trabalho significativo nessa área é Koppel and Winter (2014));
- O problema da *agulha no palheiro* (*needle in the haystack*), em que existem milhares de possíveis autores, para cada um dos quais se tem pouco material (exemplos de trabalhos nessa abordagem são Kapovciute-Dzikiene et al. (2017) e Schwartz et al. (2013));
- A *caracterização de autoria*, em que não existe um conjunto de autores para fazer atribuição, e somente é possível determinar características do autor.

Esta última abordagem é a tratada neste trabalho. Ela pode ser definida da seguinte maneira mais formal: Sejam n indivíduos $i_1, i_2, ..., i_n$, cada um dos quais pode ser rotulado com uma classe pertencente à alguma dimensão, por exemplo, masculino ou feminino para sexo. Cada indivíduo é associado com um conjunto de documentos, que podem ser de diferentes tipos (por exemplo, textos, imagens, *geodata*), dos quais são extraídos m atributos numéricos que formam um vetor de atributos $a_1, a_2, ..., a_m$. Estes vetores, associados com as classes dos indivíduos, são utilizados para treinar

um modelo, que pode ser aplicado em dados inéditos, permitindo a identificação das características de novos indivíduos (Argamon et al., 2009).

Esse método baseia-se na noção que o comportamento de diferentes classes de pessoas irá resultar em particularidades nos conteúdos que elas produzem: cada classe escreve com estilos diferentes e voltados para tópicos distintos; tira fotos de estilos específicos; expoem a si próprios mais ou menos; etc. Por exemplo, considerando a dimensão idade, seria de se esperar que adolescentes tratem de assuntos escolares (Argamon et al., 2009); enquanto pessoas com maior escolaridade escrevem textos mais elaborados e com menos erros gramaticais.

Escolher os atributos mais adequados para representar cada autor é crucial: deve-se buscar aqueles que possuam maior poder discriminador para a(s) característica(s) sendo analisadas. No caso do Instagram, os atributos podem ser agrupados em três tipos, descritos nas seções 2.6.1, 2.6.2 e 2.6.3.

2.6.1 Atributos Textuais

Indivíduos costumam se expressar e descrevr a si próprios através de textos, inclusive na internet. Logo, utilizar a grande quantidade de recursos para a caracterização de autoria voltada para texto descrita na literatura torna-se uma opção atraente (Song et al., 2018).

Certos atributos focam-se no estilo do texto, de modo independente de seu conteúdo. Em Argamon et al. (2009) esse tipo de atributo é utilizado na caracterização de autoria de textos no tocante à: sexo, idade, língua nativa e personalidade. Exemplos desse tipo de atributo são:

- Número de caracteres do texto
- Número de palavras do texto;
- Taxa de pontuação, que compreende a razão entre a quantidade de sinais de pontuação usados no texto e a quantidade de caracteres total;
- Corretude do texto, correspondente à razão entre o número de palavras escritas de modo incorreto e o número total de palavras no texto

Outros atributos textuais podem ser extraídos do conteúdo do texto. Um exemplo é a seleção das N palavras dentro do *corpus* (conjunto de textos geral) que sejam mais frequentes, com N sendo um valor pré-definido conforme trabalhos anteriores ou experimentação. Então, para todo texto, a frequência de cada uma dessas palavras será utilizada como atributo.

Outra abordagem é a seleção das N palavras mais relevantes conforme alguma métrica, por exemplo, *term frequency-inverse document frequency* (frequência de termo-frequência inversa de documento, *tf-idf*). Esta métrica não usa apenas a frequência de uma palavra ao longo do corpus, mas leva em conta também o número de documentos em que ele aparece, de modo a reduzir a influência de palavras que apareçam

igualmente na maioria dos textos e que possuem menor poder discriminativo. A formulação do tf-idf está na Equação 2.6:

$$td - idf = t_f * log(\frac{N}{D_f}) \tag{2.25}$$

Em que t_f é a frequência da palavra num documento, N é o número total de documentos e D_f é o número de documentos em que a palavra aparece no *corpus*. As palavras selecionadas são, então, as que possuam tf-idf mais alto. Robertson (2004) apresenta uma justificação teórica da utilidade do *tf-idf* na determinação de termos relevantes para encontrar documentos num *corpus*.

2.6.2 Atributos Comportamentais

As ações dos indivíduos na internet geram dados, que refletem seu comportamento, os quais se supõe que sigam certos padrões próprios a cada classe de indivíduos. Nas redes sociais, geralmente os usuários podem reagir ao conteúdo postado através de "curtidas", e estabelecer relações entre si através de pedidos de "amizades" ou seguindo uns aos outros para receber atualizações das postagens mais recentes.

Estes atributos podem ser utilizados para estabelecer relações entre o comportamento de um indivíduos e a classe a que ele pertença. Por exemplo, Pfeil et al. (2009) analisa diferenças de comportamento entre adolescentes e idosos na rede social MySpace, com adolescentes fazendo mais uso de recursos multimídia e recebendo mais comentários nos seus perfis, enquanto idosos possuíam amigos numa faixa de idade mais variada. Jang et al. (2015) detecta que adolescentes tendem a postar menos na rede social Instagram, ainda que interajam mais. Na rede social Twitter, o número de seguidores e seguidos é outro exemplo de atributo comportamental utilizado para identificar sexo e idade (Filho et al., 2016).

2.6.3 Atributos Visuais

Com a produção e publicação cada vez maior de imagens na internet, especialmente de cunho pessoal, um abordagem promissora é analisá-las a fim de extrair dados de seus usuários.

Uma maneira de fazer isso é utilizar certas ferramentas que permitem que, a partir de uma imagem, descrições do conteúdo da imagem e conceitos relacionados à ela sejam extraídos. Esses dados são usados como atributos para alimentar um classificador (Song et al., 2018). O surgimento das técnicas de *deep learning* tornou mais acessível a análise de imagens, oferecendo a extração de informações de grande quantidade de imagens com precisão.

Como exemplo de atributos visuais, existem a frequência de *selfies* publicadas em redes sociais, que foi utilizada na detecção de gênero e idade, com o uso de ferramen-

tas de análise de imagem para encontrar as *selfies* entre as imagens (Souza et al., 2015; Jang et al., 2015).

2.7 Extração de dados

As seções anteriores estabeleceram elementos fundamentais para esse trabalho: a escolha dos métodos de aprendizagem de máquina a usar na classificaão (Seções 2.1 a 2.3), as medidas para avaliar os resultados da classificação (Seção 2.4) e o problema a ser abordado (Seção 2.6). Resta definir de que modo serão extraídos os dados a ser utilizados no trabalho. Conforme Roh et al. (2019), que realiza um *survey* na área de coleta de dados, esta tarefa pode tornar-se um gargalo, devido a necessidade de conseguir grande quantidade de dados para construção dos modelos de aprendizagem de máquina. Além disso, estes dados devem estar rotulados de modo confiável a fim de que a classificação seja realizada refletindo cenários reais.

Três grandes tópicos dentro da coleta de dados são destacados: aquisição de dados, rotulação de dados e melhoramento de dados, descritos nas seções 2.7.1, 2.7.2 e 2.7.3, respectivamente. Um apanhado de como os dados têm sido coletados na literatura selecionada é feito na seção 2.7.4.

2.7.1 Aquisição de dados

Aquisição de dados (*data acquisition*) é a área voltada para busca e compartilhamento de novos conjuntos de dados. Os conjuntos de dados podem vir de alguma fonte disponível, o que é chamado de descoberta de dados. Essa descoberta pode ser feita a partir de plataformas em que conjuntos de dados já formatados são disponilizados, como o Kaggle[1]. Ou então, os dados, já encontrados de maneira estruturada ou semi-estruturada na internet através de páginas *web* ou APIs são coletados e reunidos (Roh et al., 2019).

Outra abordagem é o incremento de dados, em que a um conjunto de dados já existente são adicionados novos dados, seja preenchendo valores faltantes em atributos dos elementos já presentes no conjunto de dados (Yakout et al., 2012), seja criando novos atributos para todos os elementos a partir dos atributos já presentes (Pennington et al., 2014), ou adicionando novos elementos ao conjunto.

No caso de escassez de fontes de coletas de dados, eles podem ser criados, tarefa denominada geração de dados. Essa criação pode ser manual, envolvendo a técnica de *crowdsourcing*, em que várias pessoas fornecem dados que são incorporados num conjunto de dados, por exemplo, respondendo questionários. Há também a possibilidade de gerar de conjunto de dados por métodos automáticos, como através de *Generative Adversarial Networks* (GANs), que através de redes neurais gera grande quantidade de dados a partir de um pequeno conjunto de elementos. Por exemlo,

[1] https://kaggle.com

a partir de alguns prontuários médicos, uma GAN pode gerar dados de paciente artificiais, aprendendo padrões dos atributos (Choi et al., 2017).

2.7.2 Rotulação de dados

Rotulação de dados (*data labeling*), consiste na atribuição de rótulos para cada dado, existindo diversas técnicas para isso. Os rótulos indicam a que classe cada elemento pertence, logo é necessário que eles reflitam conhecimento verdadeiro para que a classificação seja feita de modo confiável. A mais direta é extrair os rótulos juntamente com os atributos quando da construção do conjunto de dados. Ou então, com uma parte do conjunto de dados já rotulada confiavelmente, utilizar métodos de aprendizagem de máquina para rotular o restante.

No caso em que o conjunto de dados vier totalmente sem rótulos, uma alternativa é utilizar novamente o *crowdsourcing*, com várias pessoas manualmente rotulando cada elemento, com um sistema de votação para determinar o rótulo final. Em Marcus et al. (2011), por exemplo, é apresentado um sistema para elaborar pesquisas que construam conjuntos de dados a partir do serviço da Amazon Mechanical Turk.[2] Snow et al. (2008) utiliza a mesma técnica para diversas tarefas de anotação de texto, como análise de sentimento e similaridade de palavras. Os autores concluem que *crowdsourcing* de rótulos por não-especialistas resulta num conjunto de dados com qualidade similar a um rotulado por um especialista.

Há também a abordagem dos rótulos fracos (*weak labels*), que consiste na obtenção de rótulos com confiabilidade reduzida, mas num contexto em que o conjunto de dados seja tão grande, que a existência de rótulos incorretos tenha pouco impacto. Essa estratégia pode ser útil nas situações que a rotulação correta de todo o conjunto de dados seja demasiadamente custosa. Tambi et al. (2020) utiliza rótulos fracos para identificar a linguagem em consultas da engine de busca Adobe Stock.

2.7.3 Melhoramento de dados

Melhoramento de dados (*data improvement*) é a abordagem que busca melhorar um conjunto de dados já pronto, de modo a aumentar a qualidade da classificação. Roh et al. (2019) dá como exemplo de melhoramento de um conjunto de dados a remoção de elementos indesejados, que representem anomalias e impactem negativamente a classificação. Um tipo de elemento indesejável são os *outliers*, que possuem pelo menos um atributo que esteja extremamente distante dos demais, e possam causar ruído no conjunto de dados. Outro caso é o de atributos fora dos limites possíveis, como idade, altura ou peso com valor negativo. A Tabela 2.3 contém um exemplo de conjunto de dados de altura e idade de indivíduos com anomalias: o indivíduo de id 0003 possui altura negativa, estando fora dos limites válidos para o atributo; enquanto o indivíduo de id 0005 possui 60 anos, contrastando com os outros que estão na faixa dos 20.

[2]https://www.mturk.com/

Id	Altura	Idade
0001	188	22
0002	175	26
0003	-170	23
0004	176	28
0005	180	60

Tabela 2.3: Exemplo de dados com *outliers* e valores fora dos limites

Outro exemplo é o melhoramento dos rótulos dos elementos dos conjunto de dados obtidos, refazendo a rotulação nos casos em que ela esteja imprecisa. Sheng et al. (2008) melhora a qualidade de rótulos em conjuntos de dados através da combinação de múltiplos rotuladores distintos para reduzir o ruído deles. Esses rotuladores podem ser automatizados ou então humanos, num esquema de *crowdsourcing*.

2.7.4 Extração de dados no Instagram

A literatura relacionada contém diversos métodos de coleta de perfis do Instagram. Souza et al. (2015) coleta todos os usuários públicos ativos entre Dezembro de 2011 e Dezembro de 2014, através de um processo de verificação de quais *ids* de usuário são válidas. Jang et al. (2015); Han et al. (2016); Song et al. (2018), por sua vez, escolhem um conjunto de usuários como sementes e a partir dos seguidores deles selecionam novos usuários até atingir certo quantidade. Desses usuários obtidos, outro conjunto é escolhido aleatoriamente e o mesmo processo é repetido para uma quantidade maior. Zhang et al. (2016) coleta *posts* do Instagram através de buscas por uma *hashtag* previamente definida, registrando os seus autores.

Uma vez coletados os perfis, é necessário rotulá-los corretamente com as características estudadas, que neste e na maioria dos trabalhos relacionados inclui a faixa etária, com exceção de Zhang et al. (2016). Análise das fotos publicadas pelos usuários através da ferramenta Face++, que reconhece faces de pessoas em imagens e infere dados demográficos sobre elas, é feita em Souza et al. (2015); Jang et al. (2015); Han et al. (2016); Song et al. (2018) para determinação da idade dos perfis. Enquanto Souza et al. (2015) utiliza uma seleção aleatória das fotos publicadas pelos usuários para a rotulação, Jang et al. (2015); Han et al. (2016); Song et al. (2018) utilizam as fotos de perfil.

Adicionalmente, Jang et al. (2015); Han et al. (2016); Song et al. (2018) procuram na biografia dos usuários descrições explícitas de idade. Han et al. (2016) usa também uma abordagem para criar outro conjunto de dados rotulado com faixa etária: procura entre os usuários coletados aqueles cujos posts contenham *hashtags* que implicam uma determinada idade (como *#sweet16* ou *#my18birthday* para usuários com 16 e 18 anos, respectivamente).

2.8 Trabalhos relacionados

Nesta seção, os trabalhos relacionados ao tema são apresentados e discutidos. A Seção 2.8.1 trata dos primeiros trabalhos acerca da caracterização de autoria, já a Seção 2.8.2 trata dos trabalhos acerca da caracterização de autoria no Instagram. A Tabela 2.4 contém um resumo dos trabalhos apresentados nessa seção.

2.8.1 Trabalhos iniciais

Os primeiros esforços em detecção de características de indivíduos estavam ligados a elementos textuais, como por exemplo Estival et al. (2007), que investiga idade, sexo, origem geográfica, nível educacional e linguagem nativa de *e-mails* utilizando 689 atributos divididos em três grupos: caractere (como tamanho das palavras), léxico (por exemlo, frequência de palavras-função) e estrutural (como uso de *html* no e-mail). Diversos métodos de aprendizagem de máquina foram utilizados: árvores de decisão J48 e *Random Forest* (Breiman, 2001), *lazy learning* (Han et al., 2011), algoritmo JRip de aprendizado de regras (Cohen, 1995), *Support Vector Machines* nas implementações SMO e LibSVM (Chang and Lin, 2011); e os *ensembles Bagging* e *AdaBoostM1* (Han et al., 2011; Schapire, 2013). Em todos os atributos o desempenho foi superior ao *baseline*, com o método SMO sendo o melhor para idade, sexo e país (acurácias de 56,46%; 69,26% e 81,13%, respectivamente); *Random Forest* para linguagem nativa, com acurácia de 84,22%; e *Bagging* o melhor para nível educacional, alcançando acurácia de 79,92%. Em quase todos os casos o conjunto total de atributos resulta nos melhores resultados. Uma limitação deste artigo está na análise exclusiva de e-mails, que foram disponibilizados de maneira voluntária por pessoas contatadas pelos autores.

Argamon et al. (2009) trata da descoberta de sexo, idade, língua nativa e nível de neuroticismo de indivíduos, a partir de textos vindos de blogs escritos em inglês. Foram utilizados dois tipos de atributos: listas de palavras extraídas dos textos, a fim de capturar os temas mais comuns para cada classe de indivíduos; e outro voltado para o estilo de escrita, buscando capturar como cada classe usa a linguagem. A detecção é realizada através de aprendizagem de máquina com o algoritmo Regressão Multinomial Bayesiana. Uma combinação dos grupos de atributos de conteúdo e estilo obtém o melhor desempenho para a detecção de idade e sexo, chegando em acurácias de 76,1% e 77,7%.

Os autores em Álvarez-Carmona et al. (2016) tratam da caracterização de autor em redes sociais ainda no âmbito de texto para a descoberta de sexo e idade, mas utilizando como atributos dois grupos: seleção de 5000 palavras usadas pelos indivíduos nos textos; e os tópicos tratados por eles, ou seja, os temas mais comuns utilizados no *corpus*. Duas estratégias para descobrir tópicos foram utilizadas: um conjunto de tópicos pré-definidos na biblioteca LIWC e tópicos descobertos a partir do *corpus* usando *Latent Semantic Analysis* (LSA). O conjunto de dados veio da edição 2014 da competição PAN@CLEF, que envolvia uma variedade de redes sociais, *blogs*

e comentários. Na classificação, LibLINEAR (Fan et al., 2008) foi utilizado, com resultados que apontaram a superioridade dos tópicos para atributos *bag-of-words*, chegando a 72% de acurácia na classificação por sexo e 48% para idade. Este artigo utiliza apenas atributos textuais, pois esse era o foco da competição, não aproveitando o potencial dos demais atributos possíveis em rede sociais.

Em Filho et al. (2016) a identificação de sexo no Twitter, para indivíduos falantes de língua portuguesa, foi feita usando meta-atributos de quatro tipos: baseados em caracteres ou sintaxe (ex.: número total de caracteres, razão entre o número de exclamações e o número total de caracteres); baseados em palavra (como número total de palavras e tamanho médio de palavra); baseados em estrutura textual (como total de parágrafos e número médio de palavras por parágrafo) e baseados em morfologia (ex.: razão entre o número de pronomes e o número total de palavras). O aprendizado de máquina foi feito com os métodos *BFTree* (Friedman et al., 2000), NBM e SVM. O procedimento foi feito em duas maneiras de lidar com o *corpus*: considerando cada *tweet* de um usuário individualmente ou concatenando todos os publicados pelo mesmo perfil num único texto. O melhor desempenho foi alcançado por *BFTree* usando todos os *tweets* concatenados, com 81,66% de acurácia. Este trabalho também não utiliza os atributos comportamentais que poderiam ser obtidos dos usuários, não aproveitando o potencial dessa abordagem.

Em Rodrigues et al. (2017) foi analisada a detecção de idade de usuários no site Meu Querido Diário, que é voltado para o público adolescente. Os autores mencionam como motivação a detecção de potenciais predadores sexuais. Um conjunto de entradas de usuários (dividido igualmente entre adultos e adolescentes) foi processado com o LIWC, de modo a contabilizar a participação de cada usuário em 64 categorias pré-definidas na biblioteca. Foram utilizados os classificadores ZeroR, *Random Forest*, *Naive Bayes*, *Naive Bayes Multinomial*, SMO e LMT com *10-fold cross-validation*. O método LMT obteve o melhor desempenho, com medida F1 de 0,729.

Os autores em Filho et al. (2014) abordam a identificação de sexo e idade na língua portuguesa, com uma base de dados vinda da rede social Twitter. A identificação de sexo teve duas etapas, uma em que todos os termos do *corpus* eram utilizados como atributos, e outra em que foram selecionados os termos mais relevantes de acordo com InfoGain ou Chi-Quadrado. Esses atributos foram modelados de acordo com sua presença ou ausência nos *tweets* de um indivíduo, considerando uma frequência mínima de três. Como classificadores foram utilizados *Naive Bayes Multinomial* (NBM) (Rennie et al., 2003), *Naive Bayes* (Han et al., 2011), SVM e *Random Forest*. Para encontrar a melhor performance, a seleção de termos com InfoGain e Chi-Quadrado foi realizada iniciando em 10 termos e aumentando a quantidade de termos de 10 em 10. O caso ótimo foi obtido com 20 termos selecionados, com a classificação sendo feita com o método *Naive Bayes Multinomial* levando a uma precisão de 73,2%. Os autores conseguiram ampliar a precisão para mais de 90% incrementando o método com um dicionário de nomes de usuários. No tocante à idade, um dicionário de palavras foi construído contendo termos que seriam dis-

criminativos, como relacionados a entretenimento, ocupações e gírias, conforme a "intuição" dos autores, segundo os próprios, "retratando eventos, lugares e expressoes tıpicas de cada etapa da vida". Além disso, foram calculados certos atributos textuais estruturais: média e mediana de termos positivos e negativos conforme uma análise de sentimento, erros ortográficos, preposições, artigos e emoticons utilizados; e atributos não-textuais: sexo; média e mediana de *hyperlinks*; *hashtags*; número de caracteres; frequência de postagem (dia, semana e mês); número de palavras; divulgação de localização geográfica; preenchimento de cidade; número de seguidos e de seguidores. Os melhores resultados vieram do uso do texto completo dos tweets através do método *Naive Bayes Multinomial*, com acurácia de 81,51% e medida F1 de 0,81.

Ainda que os trabalhos apresentados nesta seção possuam pontos de contato com este, eles diferem em certas maneiras. A maioria não utiliza atributos oriundos de redes sociais, limitando-se a abordagens textuais. Há pouca variação nas características estudadas. Sexo e idade aparecem com frequência, com mais alguns (língua nativa, origem geográfica, nível de neuroticismo e nível educacional) aparecendo em Estival et al. (2007); Argamon et al. (2009). Contudo, estes últimos trabalhos utilizaram dados oriundos apenas de blogs. Surge a oportunidade de detectar novas características no âmbito de redes sociais.

2.8.2 Trabalhos que utilizam o Instagram

Os primeiros trabalhos na área que utilizam bases de dados oriundas do Instagram não realizam a abordagem de aprendizagem de máquina, limitando-se a uma análise estatística. Por exemplo, Jang et al. (2015) investiga a identificação de idade na plataforma, dividindo os usuários entre adultos e adolescentes e verificando a diferença em interações (número de fotos, curtidas, *tags*, comentários, seguidos e seguidores), tópicos utilizados nas *hashtags*, número de *selfies* postadas e complexidade dos textos usados em cada *post*. Os autores concluem que adolescentes publicam menos fotos, mas interagem mais e são mais expressivos sobre si mesmos.

A mesma abordagem é realizada em Zhang et al. (2016), que investiga como o uso de *hashtags* varia de acordo com o sexo no Instagram. Os autores formularam duas hipóteses: mulheres usam *hashtags* mais emocionais, enquanto homens usam *hashtags* de cunho mais positivo, ambos no contexto de posts sobre comida. Através de teste Mann-Whitman, as duas hipóteses foram confirmadas, e os autores buscaram justificar os resultados através da teoria de usos e gratificação. Este é o único trabalho voltado para o Instagram que realiza análises estatísticas para validar os resultados.

Trabalhos posteriores iniciam a análise da caracterização de autoria através da aprendizagem de máquina. Han et al. (2016) trata da descoberta de idade no Instagram, construindo um dataset de usuários rotulados como adolescentes ou adultos, utilizando duas técnicas: análise das faces dos indivíduos através da ferramenta Face++, que estima a faixa etária de alguém baseada numa foto, e busca por *posts* contendo

hashtags indicadoras de idade. Três tipos de atributos foram utilizados: oriundos do conteúdo dos posts (número de *hashtags* utilizadas e *selfies* publicadas), baseados em interações (como comentários e *likes*) e baseados em relações de seguir ou ser seguido. Foram utilizados quatro métodos de aprendizagem de máquina: SVM, Regressão Logística, *Random Forest* e *Adaptive Boosted Decision Trees*. A classificação com SVM conseguiu melhores resultados, com apenas os atributos baseados em conteúdo e interações tendo acurácia consideravelmente acima do *baseline*, em 72,6%.

Linha similar seguiu Song et al. (2018), que construiu um conjunto de dados de usuários do Instagram e tratou da descoberta de sexo e idade. Dois tipos de atributos foram utilizados: *hashtags* extraídas dos textos dos posts publicados pelos usuários, e tópicos extraídos das imagens publicadas através da ferramenta *Microsoft Azure Cognitive Services*, que utiliza métodos de aprendizagem de máquina para identificar objetos, faces e textos inclusos na imagem. Duas estratégias para a representação textual foram testadas: seleção dos termos usando tf-idf e *word2vec*, cada um deles alimentando classificadores de Regressão Logística e *Random Forest*. Utilizar apenas os atributos oriundos das imagens levou à melhor performance tanto para sexo quanto idade (na medida F1 os resultados foram 0,74 com Regressão Logística e 0,88 com *Random Forest*, respectivamente).

O único trabalho tratando de caracterização de usuários no Instagram em língua portuguesa encontrado nesta revisão bibliográfica é Campos (2016), que trata da classificação de perfis como comuns ou de negócio. Foram coletados 1389 perfis distintos, rotulados como comerciais ou comuns de maneira manual. A biografia dos perfis foi examinada para extrair os seguintes atributos: número de palavras ou expressões que caracterizam spam; número de caracteres maiúsculos; número de caracteres númericos; número de URL's no texto; número de endereços de e-mail; número de telefones; número de informações de contato no texto; número de palavras e número de palavras maiúsculas. Outros atributos foram extraídos do comportamento dos perfis, a saber: número de fotos postadas; número de seguidos; número de seguidores; fração do número de seguidores por seguidos; se possui informação na biografia e se informou website. Por fim, um atributo adicional foi extraído da imagem de perfil utilizada, um score de popularidade da imagem calculado por uma ferramenta (descrita por Khosla et al. (2014)) que realiza uma análise das propriedades visuais da imagem em questão. A classificação foi utilizada com o método *Random Forest*, com as métricas acurácia, *recall*, precisão e *F-measure*. O uso de todos os atributos em conjunto alcançou uma precisão de 81,64%.

Trabalho	**Característica de estudo**	**Origem dos dados**	**Atributos**	**Métodos**
Argamon et al (2009)	sexo, idade, língua nativa e nível de neuroticismo	Blogs	Palavras mais comuns no corpus e atributos de estilo textual	Regressão Multinomial Bayesiana
Álvarez-Carmona et al. (2016)	sexo e idade	Redes sociais, blogs e comentários em sites	Seleção de 5000 palavras mais usadas pelos indivíduos nos textos; e os tópicos mais comuns utilizados no *corpus*	LibLINEAR
Filho et al. (2016)	sexo	Twitter	meta-atributos de quatro tipos: baseados em caracteres ou sintaxe; baseados em palavra; baseados em estrutura textual e baseados em morfologia	*BFTree*, NBM e SVM
Rodrigues et al. (2017)	idade	rede social Meu Querido Diário	presença dos textos dos usuários em 64 categorias do LIWC	ZeroR, *Random Forest*, *Naive Bayes*, *Naive Bayes Multinomial*, SMO e LMT com *10-fold cross-validation*
Filho et al. (2014)	sexo e idade	Twitter	termos presentes no texto	*Naive Bayes Multinomial*, *Naive Bayes*, SVM e *Random Forest*
Jang et al. (2015)	idade	Instagram	interações entre usuários, tópicos utilizados nas *hashtags*, número de *selfies* postadas e complexidade dos textos usados em cada *post*	-
Zhang et al. (2016)	sexo	Instagram	*hashtags* utilizadas nos posts	teste Mann-Whitman
Han et al. (2016)	idade	Instagram	conteúdo dos posts, interação entre usuários, relações de seguir-seguidos	SVM, Regressão Logística, *Random Forest* e *Adaptive Boosted Decision Trees*
Song et al. (2018)	sexo e idade	Instagram	*hashtags* extraídas dos textos dos posts e tópicos extraídos das imagens publicadas	Regressão Logística e *Random Forest*
Campos e Costa (2016)	conta de negócio	Instagram	atributos extraídos dos textos pubicados, comportamento dos usuários e imagem de perfil	Random Forest

Tabela 2.4: Referências encontradas na revisão bibliográfica.

Capítulo 3

Extração de dados do Instagram

Devido à ausência de trabalhos de caracterização de autoria em língua portuguesa voltados para o Instagram, não estão disponíveis dados públicos de usuários já rotulados, formatados de modo a serem utilizados para classificação. Logo, é necessário construir os conjuntos de dados necessários para o trabalho, o que envolve diversos passos. Primeiro, realizar a coleta diretamente no Instagram, selecionar quais usuários possuem as características desejadas e decidir quais atributos a coletar. Uma vez coletados os dados, verificar a corretude dos dados obtidos, efetuando limpeza no conjunto de dados caso sejam detectadas anomalias. Este capítulo trata dessa etapa, descrevendo os passos realizados e os percalços encontrados no caminho.

É estimado em Terrizzano et al. (2015) que 70% do tempo em análise de dados está ligado à descoberta, limpeza e integração de informações, devido à natureza fragmentada delas, que cria a necessidade de reuní-las e colocá-las num formato de utilização fácil. Esta etapa do trabalho, portanto, é a que mais demanda tempo e esforço, com potencial de diversos obstáculos aparecerem. Este capítulo contribui apresentando particularidades da coleta de dados no Instagram.

A Figura 3.1 contém um diagrama do processo executado ao longo deste trabalho no tocante aos dados. Esses passos são os seguintes:

- Pesquisa automática por perfis no Instagram através de uma lista de termos previamente definidos alimentando pesquisas por *hashtags* no Instagram através de um *crawler*. Este utiliza as listas de termos para realizar pesquisas por *hashtags* na rede social. O resultado dessas consultas automatizadas são publicações de usuários com boas chances de pertencer aos grupos de profissão (saúde, humanidades e exatas) e idade (adolescente e adulto) propostos (etapa detalhada na Seção 3.1);
- Armazenamento local dos perfis coletados em arquivos .csv, permitindo sua manipulação sem necessidade de requisições adicionais ao Instagram;
- Remoção de perfis cuja biografia está num idioma que não seja português através de uma ferramenta que realiza essa detecção de modo automático.

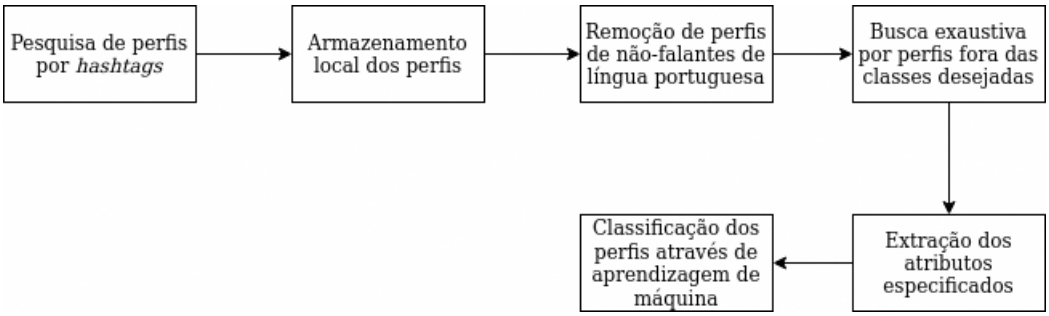


Figura 3.1: Diagrama do processo executado ao longo do trabalho com os dados.

Assim, é garantido que todos os perfis sejam de falantes nativos de língua portuguesa;

- Examinação manual dos conjuntos de dados para encontrar perfis que não pertencem às classes desejadas, e realizar sua remoção (etapa detalhada na Seção 3.3);
- Extração dos atributos especificados através de uma análise da literatura relacionada e suposições sobre os perfis;
- Classificação dos perfis através de métodos de aprendizagem de máquina (etapas detalhadas em capítulos posteriores).

3.1 Obtenção dos perfis no Instagram

O método de coleta de perfis de modo aleatório e depois obter perfis que estejam rotulados de modo confiável usando Face++ ou o texto da biografia, como feito na maioria dos trabalhos relacionados (Seção 2.7.4), acaba levando ao aproveitamento de apenas uma pequena parte dos perfis coletados: em Souza et al. (2015), 5.170.062 perfis foram coletados, e apenas 738.901 foram rotulados; Jang et al. (2015) começou com 2.000.000 e terminou com 26.885 perfis; Han et al. (2016) conseguiu, de 2.000.000, obter 60.000 perfis rotulados com faixa etária usando o método mencionado; Song et al. (2018) iniciou com 500.000 e terminou com 20.000 usuários rotulados.

A fim de reduzir este problema, neste trabalho os perfis são coletados utilizando uma estratégia que combina o que foi visto em Zhang et al. (2016) e Han et al. (2016). Para cada demografia a coletar, são definidas certas *hashtags* que lhe sejam relacionadas, e são efetuadas buscas automáticas para encontrar perfis com probabilidade de pertencerem à classe desejada. Essa abordagem baseia-se no pressuposto de que indivíduos dentro de uma certa demografia publicarão utilizando *hashtags* diretamente relacionadas à ela. Certos atributos considerados úteis para a classificação são extraídos de cada perfil. A lista de atributos e a justificativa para a escolha deles é detalhada na Seção 3.2.

Inicialmente, a biblioteca Instagram-API[1] foi utilizada para obter os dados do Instagram. Entretanto, essa biblioteca foi derrubada devido a um pedido de direitos autorais do Instagram, e parou de receber novas atualizações que acompanhassem o desenvolvimento da rede social. Portanto, passou a ser utilizada outra abordagem, com a biblioteca em python Selenium[2], que permite simular a navegação do usuário em páginas *web* e acessar os dados nela contidos de maneira automatizada.

Os perfis encontrados foram filtrados de modo a retirar os que não possuam biografias em língua portuguesa, através da ferramenta em Python langid[3], que detecta a linguagem em que um texto foi escrito através de aprendizagem de máquina. Em Lui and Baldwin (2011) é feita uma descrição do método, que utiliza o classificador *Naive Bayes* e *n-grams* a nível de caractere. Foi realizada uma comparação de desempenho e tempo de processamento em diversos *corpus* com outros dois detectores de linguagem: TextCat e GoogleAPI. O langid se mostrou superior nos dois quesitos ao TextCat em todos os *corpus*, com precisão superior à 90%. O GoogleAPI se sai levemente melhor que o langid, chegando a uma melhoria na previsão de 1,8% em relação à este, contudo sempre é significantemente mais lento, demorando até 20 vezes mais para realizar a detecção de idioma.

O *username* dos perfis coletados foi armazenado em arquivos csv, juntamente com os atributos descritos na Seção 3.2 para cada um deles.

3.1.1 Hashtags utilizadas na busca por área profissional

Para formular a lista de *hashtags* usada na busca por perfis por área profissional, foi utilizada uma lista das 10 graduações mais populares no Brasil, segundo pesquisa publicada pela revista Guia do Estudante em 2018[4]. Essa seleção de profissões está focada nas mais populares entre estudantes para tentar obter um número grande o suficiente de usuários através das pesquisas. As 10 graduações informadas na pesquisa são: Direito, Administração, Pedagogia, Engenharia Civil, Ciências Contábeis, Enfermagem, Psicologia, Educação Física, Arquitetura e Urbanismo e Engenharia de Produção.

Como duas engenharias figuram na lista (Civil e de Produção), a menos popular foi eliminada, sobrando nove graduações, três para cada área profissional. A Tabela 3.1 contém uma lista dessas *hashtags*. Como *hashtags* não podem ter espaços, as profissões “Engenharia Civil”, “Ciências Contábeis” e “Educação Física” se tornaram as *hashtags* “engenhariacivil”, ‘ “cienciascontabeis” e “educacaofisica”, respectivamente.

[1] https://github.com/mgp25/Instagram-API/wiki

[2] https://selenium-python.readthedocs.io/

[3] https://github.com/saffsd/langid.py

[4] https://guiadoestudante.abril.com.br/blog/pordentrodasprofissoes/os-10-cursos-de-graduacao-mais-procurados-do-brasil/

Exatas	Biológicas	Humanas
arquitetura	psicologia	direito
engenhariacivil	enfermagem	administração
cienciascontabeis	educaçãofisica	pedagogia

Tabela 3.1: Lista de *hashtags* usadas na pesquisa de publicações de Instagram por área de profissão.

Adolescentes (15-19)	Adultos (21-25)
meus15anos	meus21anos
meus16anos	meus22anos
meus17anos	meus23anos
meus18anos	meus24anos
meus19anos	meus25anos

Tabela 3.2: Lista de *hashtags* usadas na pesquisa de publicações de Instagram por idade.

3.1.2 Hashtags utilizadas na busca por faixa etária

Neste trabalho, foram definidas duas faixas etárias: adolescentes, considerados os usuários de 15 a 19 anos, e adultos, consistindo de usuários entre 21 e 25 anos. Conforme Smith and Anderson (2018), a maioria dos adultos que utiliza o Instagram nos EUA têm até 24 anos de idade, com 72% dos integrantes dessa demografia afirmando usar o Instagram. Em contraste, na faixa de 25 a 29 anos o uso cai para 54%, e de 30 a 49 para 40%. Baseado nisso, a faixa etária dos usuários adultos coletados se limita a 25 anos, a fim de garantir um maior número de perfis coletados.

A Tabela 3.2 contém as *hashtags* usadas para pesquisar os usuários em cada faixa etária. Cada faixa etária contém cinco *hashtags* para pesquisa, associadas às faixas etárias dos indivíduos buscados.

3.2 Atributos extraídos

A fim de poder alimentar métodos de aprendizagem de máquina, os dados dos usuários precisam ser transformados em atributos quantitativos, que melhor discriminem as classes consideradas. Cada usuário é representado, assim, por um vetor de valores numéricos. Os atributos são escolhidos conforme seu uso em trabalhos anteriores na área, de forma a selecionar os mais comuns e com possibilidade de fornecerem um melhor desempenho.

As seções 3.2.1, 3.2.2 e 3.2.3 descrevem os atributos agrupados por tipo e as razões por trás de seu uso.

3.2.1 Atributos comportamentais

Os atributos comportamentais mais comuns na literatura (Seção 2.6.2) referem-se a frequência de publicações, curtidas, comentários, número de seguidos e seguidores (Jang et al., 2015; Filho et al., 2014) e divulgação de localização geográfica (Filho et al., 2014).

Um mecanismo que o Instagram oferece como oportunidade para capturar o comportamento dos usuários não explorado na literatura é o recurso das contas de negócio[5], para aqueles que queiram impulsionar seus negócios através da rede social. Uma conta de negócio tem acesso a recursos adicionais que permitem entender melhor qual seu público. Além disso, a categoria da empresa e as informações de contato (como telefone e endereço) podem ser exibidas no perfil. É de se esperar que adultos, por estarem numa proporção maior no mercado de trabalho, utilizem o Instagram para fins profissionais numa proporção maior e tenham uma quantidade mais alta de contas de negócio. Além disso, em profissões em que boa parte dos profissionais atuam de forma autônoma, como direito e psicologia, também têm a possibilidade de conter maior quantidade de contas de negócio.

O número de fotos publicadas no Instagram é identificado como um diferencial entre adolescentes e adultos por Jang et al. (2015), que através de um estudo ANOVA identifica que adolescentes publicam menos. O mesmo estudo não identificou diferenças significativas entre adultos e adolescentes nos números de seguidores e seguidos. Han et al. (2016), por sua vez, identifica que adultos tendem a possuir mais seguidores. Visto que o material utilizado é de falantes de língua inglesa, resta verificar se tais relações se mantém para falantes de língua portuguesa.

Baseado nesse levantamento, os atributos comportamentais escolhidos são: se a conta é de negócio ou não; o número de publicações realizadas pelo usuário; o número de seguidores que o usuário possui e o número de usuários que ele segue. Este grupo de atributos está na primeira coluna da Tabela 3.3.

3.2.2 Atributos textuais

Atributos textuais (Seção 2.6.1) são utilizados em quase todos os trabalhos relacionados consultados. Portanto, eles também são usados neste trabalho, sendo extraídos da biografia dos usuários. Certos trabalhos extraem os atributos a partir do conteúdo dos textos: utilizando tópicos (Álvarez-Carmona et al., 2016; Rodrigues et al., 2017); palavras (Filho et al., 2014, 2016; Álvarez-Carmona et al., 2016) e *hashtags* (Song et al., 2018; Zhang et al., 2016; Jang et al., 2015). Outros analisam a estru-

[5]https://help.instagram.com/502981923235522

Comportamentais	**Textuais (biografia)**	**Imagem**
Identificador de conta de negócio	100 hashtags mais frequentes da biografia	Quantidade de selfies
Identificador de localização	Taxa de pontuação utilizada	Quantidade de imagens de paisagem
Número de seguidores	Taxa de emojis utilizados	Quantidade de imagens com texto
Número de seguidos	Riqueza de vocabulário	Quantidade de imagens com usuários marcados
	Número de caracteres	Quantidade de imagens com crianças
	200 palavras mais relevantes da biografia	

Tabela 3.3: Lista de atributos coletados de cada usuário.

tura dos textos, medindo o uso de pontuação (Filho et al., 2014, 2016), emoticons e emojis (Filho et al., 2014) e a complexidade dos textos (Jang et al., 2015).

Os atributos textuais escolhidos são: a frequência das 100 *hashtags* mais frequentes em cada classe (área profissional e faixa etária); a frequência das 200 palavras mais relevantes para cada classe; taxa de pontuação utilizada (razão entre número de sinais de pontuação e total de caracteres); taxa de emojis utilizados (razão entre emojis e total de caracteres); riqueza de vocabulário (razão entre o número de palavras únicas e o número total de palavras) e número de caracteres utilizados. A Tabela 3.3 contém esses atributos na segunda coluna.

A fim de selecionar quais palavras sejam mais relevantes, foi utilizada a métrica *term frequency-inverse document frequency*. Nessa seleção, foi feito um pré-processamento das biografias, consistindo na remoção de todos os caracteres não-alfanuméricos (como sinais de pontuação e emojis) e de *stopwords*, e colocando todas as palavras em minúsculas.

Na extração das *hashtags*, verificou-se que diversos usuários escreviam várias delas sem espaços entre si, no seguinte formato: *#hashtag1#hashtag2#3hashtag3...* Foi realizado um pré-processamento também nessa etapa para lidar com isso, separando as *hashtags* nesse formato. A seleção das *hashtags* e termos foi realizada com o conjunto de dados obtido após a limpeza, passo detalhado na Seção 3.3.

3.2.3 Atributos de imagem

Quanto aos atributos de imagem (Seção 2.6.3), as bibliotecas de processamento de imagem usadas nos trabalhos relacionados são proprietárias e pagas (em Jang et al.

Figura 3.2: Imagem publicada no Instagram, que recebeu o texto alternativo *May be an image of 1 person, waterfall and nature.*

(2015); Souza et al. (2015) a ferramenta Face++ é usada, e em Song et al. (2018) a Microsoft Azure Cognitive Services). A solução escolhida nesse trabalho foi usar o texto alternativo automático das imagens, que passou a ser fornecido pelo próprio Instagram a partir de Novembro de 2018[6] e não estava disponível para os trabalhos citados. Conforme descrito pela plataforma, "o texto alternativo automático usa uma tecnologia de reconhecimento de objetos a fim de criar descrições de fotos para pessoas com deficiências visuais"[7]. Assim, pode-se reduzir o custo envolvido tanto no processamento das imagens, quanto no *benchmarking* que seria feito para escolher uma dentre diversas bibliotecas de processamento de imagem.

Os txtos alternativos das imagens obtidos estão em língua inglesa. A Figura 3.2 contém uma imagem publicada no Instagram. A plataforma atribuiu a essa imagem o texto alternativo *May be an image of 1 person, waterfall and nature.* Vale notar que nem todas as fotos recebem um texto alternativo pelo Instagram, e o usuário pode alterar o texto alternativo fornecido pela plataforma caso deseje. Além do mais, o recurso só está disponível para fotos e não para vídeos, que também podem ser publicados na rede social. No caso de fotos que contenham texto, uma transcrição dele é inserida no texto alternativo. Caso a foto contenha pessoas, algumas características delas podem aparecer, como se estão em pé, se possuem barba, o tipo de roupa usada, acessórios como óculos de sol ou se são crianças.

A ideia é analisar o texto alternativo das imagens publicadas pelos usuários a fim de detectar quais publicações são *selfies* e quais são fotos da natureza, baseando-se na ideia que esses tipos de imagem são indicativos de adolescentes e adultos, respectivamente, conforme indicado em Han et al. (2016) e Song et al. (2018). Como *selfies*, são consideradas as imagens com texto alternativo incluindo as palavras

[6]https://about.instagram.com/blog/announcements/improved-accessibility-through-alternative-text-support

[7]https://help.instagram.com/503708446705527

"selfie", *"person"* ou *"closeup"*. As publicações de natureza são considerados os cujo texto alternativo incluem os termos *"nature"* ou *"outdoors"*.

Outra identificação feita nos textos automáticos das imagens foi se existe alguma transcrição de texto neles, sob a hipótese de que adultos publiquem mais fotos com texto, seja porque usem o Instagram como modo de comunicação, seja porque adolescentes estão concentrados em *selfies*. Além do mais, classes profissionais, ao usarem o Instagram para anúncios de seus serviços, também poderiam publicar imagens contendo texto. O texto automático que caracteriza essa situação é contém a frase *"May be a picture of text"*.

Também foi formulada a hipótese que dada a tendência de adolescentes de estabelecer mais relações através do Instagram (Han et al., 2016), eles poderiam marcar outros usuários em suas fotos. Assim, foi feita a detecção de se o símbolo de menção, @, é utilizado no texto alternativo. Por fim, foi considerado que adultos têm mais chances de ter filhos, por isso procurou-se pelo termo *"child"* no texto alternativo das imagens. Foram coletados os dados das 12 publicações mais recentes de cada perfil, devido às limitações de tempo deste trabalho.

3.3 Limpeza do conjunto de dados rotulado por área profissional

A coleta dos perfis dos usuários do Instagram apresentou problemas na corretude das classes dos usuários. Após a remoção daqueles cuja biografia não estava em português, uma inspeção inicial nos conjuntos de dados resultantes revelou que restaram diversos usuários que estavam fora das classes procuradas. No *dataset* rotulado de acordo com área profissional, esses tipos de usuários pertenciam a dois tipos de perfis, que colocavam as hashtags pesquisadas em suas publicações com os seguintes intuitos:

- Perfis buscando vender cursos ou materiais relacionados à área, frequentemente lojas especializadas em artigos voltados para à área profissional, bem como cursos pré-vestibulares. Por exemplo, a Tabela 3.4 contém o início de um arquivo CSV contendo os perfis encontradas na busca pela *hashtag* "enfermagem". O primeiro perfil encontrado pertence à um curso, enquanto o terceiro é voltado para venda de jalecos, o segundo e o quarto são perfis pessoais da área de enfermagem.
- Perfis de áreas profissionais adjacentes, que eventualmente publicavam algo sobre a área sendo pesquisada. Continuando com o exemplo da pesquisa pela *hashtag* enfermagem, perfis de pessoas de fisioterapia, medicina e química, entre outras, foram encontrados. As duas últimas linhas da Tabela 3.4 contém alguns perfis que se enquadram nesse caso.

Assim, foi realizada uma busca manual exaustiva dos conjuntos de dados, a fim de identificar perfis que se encaixassem em algum desses casos. Eles foram então

seguidores	seguindo	negócio	publicações	bio
2492	2035	sim	585	Cursos de excelência em ...
1203	529	não	277	Acad. Enfermagem 7/10
87	83	sim	84	Ó seu melhor LOOK!´ Inhumas-GO (Em breve. Aguardem!)•𝒥⊣↕⌉⌋≀∫⇔ ↙↙
296	550	não	62	Estudante de Universidade Pública ...
2547	3874	sim	555	Atendemos a todo o Brasil
134	106	não	70	Fisioterapeuta: Pós em ...
215	3	não	12	Professora e fisioterapeuta ...

Tabela 3.4: Resultados da busca por enfermagem, contendo perfis voltados para vendas juntamente com outros que de fato pertencem à área.

removidos do conjunto de dados final, e todos os demais mantidos no rótulo proposto. Destes perfis, parte possui uma declaração explícita de pertencimento à profissão desejada no nome de usuário ou na biografia, do que se depreende um grau de confiança considerável que os usuários de fato pertencem às profissões buscadas.

3.4 Limpeza do conjunto de dados rotulado por faixa etária

Na construção do conjunto de dados rotulado por faixa etária, os perfis fora do padrão desejado se enquadravam em dois casos: estabelecimentos de festa que utilizavam as hashtags de aniversário para anunciar seus serviços (especialmente na hashtag meus15anos) e usuários que fizeram as publicações há muito tempo, de modo que atualmente estão fora da faixa etária proposta. A Tabela 3.5 contém um exemplo de resultados da pesquisa de perfis para a *hashtag meus16anos*. Pode-se notar que o segundo perfil listado é um perfil voltado para vendas, e interessantemente os outros perfis contém diretamente na biografia suas idades.

O método de coleta de usuários através das hashtags indicativas de faixa etária foi alterado de modo a considerar apenas publicações feitas há no máximo um ano antes, resolvendo o problema dos usuários cuja faixa etária mudou desde a data da publicação. Para os outros casos de usuários fora das classes nos conjuntos de dados mencionados acima, foram realizadas inspeções manuais como descrito na seção anterior, com a remoção dos casos encontrados.

seguidores	**seguindo**	**negócio**	**publicações**	**bio**
574	182	não	não	Viva pelo que te faz sorrir... 16y
1183	150	sim	316	Loja N7 Acessórios, Vestuário...
254	557	não	27	Deus a cima de tudo Católica 16 anos...
1997	1651	sim	52	| 16y | IEADPE | Violinista...
441	240	não	5	Beba água e espalhe positividade! Ji-Paraná Ro 16 primaveras

Tabela 3.5: Exemplos de resultados da busca pela *hashtag* meus16anos, com perfis da faixa etária em meio a perfis voltados pra vendas.

Capítulo 4

Conjunto de dados obtido

Esta seção trata do conjunto de dados obtido com os métodos descritos no Capítulo 3. Primeiro, uma descrição do conjunto de dados é feita na Seção 4.1. Uma análise das biografias dos perfis coletados é realizada na Seção 4.2. O objetivo é verificar se os atributos extraídos dos perfis exibem diferenças o bastante para discriminar entre as classes. Análises similares são feitas na Seção 4.3 para o número de seguidores e seguidos e na Seção 4.5.

4.1 Descrição dos conjuntos de dados

O conjunto de dados rotulado por área profissional gerado possui 1.535 perfis marcados como pertencentes às ciências biológicas, 1.157 perfis marcados com ciências humanas, e 716 perfis de ciências exatas, totalizando 3408 elementos. Esta quantidade foi alcançada após a eliminação dos perfis fora das classes desejadas, sendo que o conjunto de dados coletado inicialmente possuía 9.397 perfis. Ou seja, apenas 36,3% dos perfis coletados foram aproveitados para o trabalho, apesar do objetivo

	Mínimo	**Máximo**	**Média**	**Desvio Padrão**
Seguidores	0	1.720.000	6.476,45	1.720,25
Seguindo	0	7.587	1.364,34	1.632.33
Publicações	0	34.935	539,41	1.321,81
Tamanho Biografia	0	152	117,41	37,40
Emojis	0	32	4,18	3,73
Pontuação	0	23	1,59	1,88
Vocabulário	0	1	0,89	0,18

Tabela 4.1: Descrição do conjunto de dados rotulado por área profissional.

	Mínimo	**Máximo**	**Média**	**Desvio Padrão**
Seguidores	7	288.722	3.297,04	9.960,68
Seguindo	1	7.506	1.742,06	1.636,90
Publicações	0	19.741	425,87	877,16
Tamanho Biografia	0	150	97,91	41,84
Emojis	0	32	5,34	4,07
Pontuação	0	15	1,30	1,72
Vocabulário	0	1	0,90	0,18

Tabela 4.2: Descrição do conjunto de dados rotulado por faixa etária.

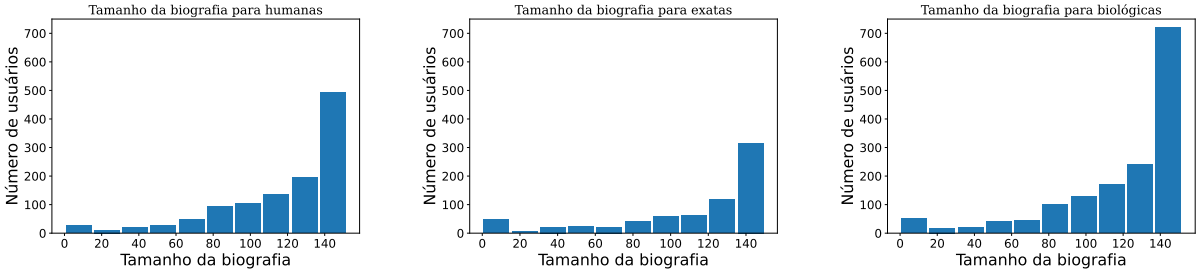
do método de coleta de dados utilizado ser reduzir a perca de dados na rotulação. A Tabela 4.1 contém o resumo desse conjunto de dados, contendo as informações descritivas de cada um dos atributos (mínimo, máximo, média e desvio padrão).

O total de perfis rotulados por faixa etária gerado possui 1.122 perfis marcados como adolescentes, e 1.377 perfis de adultos, resultando em 2.499 perfis. Originalmente, foram coletados 3.779 perfis, o que significa que 66,1% dos perfis coletados foram aproveitados. Considerando ambos os conjuntos de dados, houve um aproveitamento de 44,8% dos perfis coletados. A Tabela 4.2 contém o resumo desse conjunto de dados, com a descrição de cada um de seus atributos (máximo, mínimo, média e desvio padrão).

Nas descrições de ambos os conjuntos de dados, surge a questão de como podem haver perfis com 0 publicações, visto que para a coleta era necessário haver pelo menos um. Supõe-se que no intervalo entre a detecção dos perfis e a coleta dos dados os posts de alguns perfis foram deletados. Em área profissional o máximo de seguidores é de mais de um milhão, estando em pouco mais de 200.000 na faixa etária, indicando que os perfis coletados por área profissional possuem maior popularidade. A média no conjunto rotulado por área profissional também é maior para esse atributo, contudo é para faixa etária que há um maior desvio padrão.

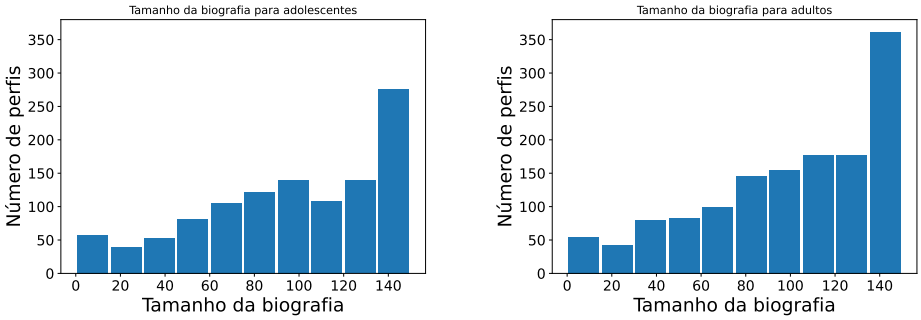
O máximo de perfis seguidos e do tamanho da biografia em ambos os conjuntos de dados têm valores similares, próximos a 7.000 e 150, respectivamente. Estes valores estão próximos dos limites que o Instagram oferece para o número de perfis que alguém pode seguir e o número de caracteres na biografia. Nos dois conjuntos de dados os desvios padrão também estao próximos. A taxa de emojis e pontuação, bem como a variedade de vocabulário, apresentam valores similares nos dois conjuntos de dados.

As seções seguintes contém análises dos atributos tanto para faixa etária quanto para área profissional, investigando a maneira como os atributos estão distribuídos em cada classe. Diferenças significativas podem apontar para maior potencial do



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.1: Distribuições dos tamanhos das biografias para o conjunto de dados rotulado por área profissional



((a)) Adolescentes ((b)) Adultos

Figura 4.2: Histograma dos tamanhos das biografias para o conjunto de dados rotulado por faixa etária.

atributo de discriminar entre as classes analisadas.

A fim de verificar com maior precisão a similaridade da distribuição desse atributo, foram feitos testes de correlação ponto-bisserial. Este teste mede a correlação entre uma variável contínua e outra binária, resultando numa escala entre -1 e 1, indicando correlação inversa e direta, respectivamente. Logo, quão mais distante de 0, menor é a relação da variável contínua com a binária. Nos testes realizados neste trabalho, as variáveis contínuas são os atributos coletados, e as binárias a classe atribuída ao perfil. Como o conjunto de dados rotulado por área profissional possui três classes, ele não pode ser dividido naturalmente numa variável binária. Para resolver isso, foi usado um esquema um-contra-todos, em que os perfis de uma classe recebiam o valor 1, e os de todas as outras o valor 0, resultando em três testes de correlação para cada atributo.

Os resultados dos testes de correlação ponto-bisserial estão nas Figuras 4.3 e 4.4 para área profissional e faixa etária, respectivamente. Para faixa etária, o número de emojis e a quantidade de caracteres na biografia são os atributos com valores de r mais distantes de 0: ambos por volta de $0,04$. Em área profissional, as correlações

que mais se destacam são as de número de emojis para humanas e biológicas e o vocabulário para exatas, com valores em torno de $0,10$.

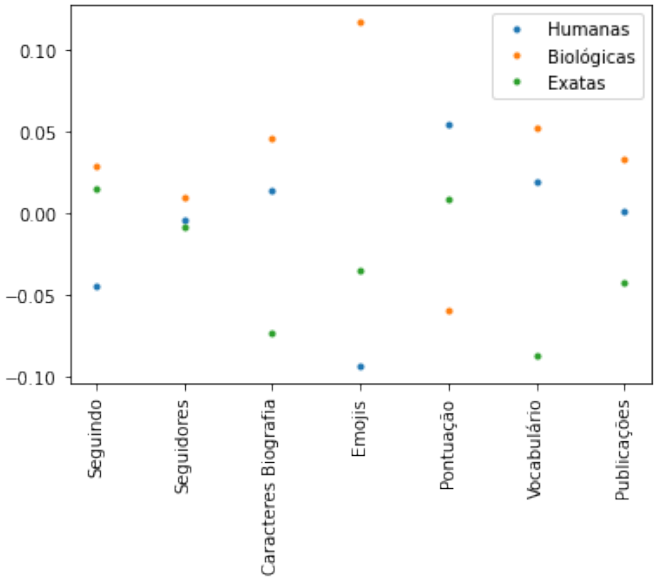


Figura 4.3: Resultados dos testes de correlação ponto-bisserial para o conjunto de dados rotulado por área profissional

4.2 Análise das biografias

Nesta seção, são apresentadas descrições sobre os dados extraídos das biografias dos perfis coletados. A Seção 4.2.1 apresenta uma análise das distribuições dos valores extraídos das biografias de cada perfil. As Seções 4.2.2 e 4.2.3 descrevem os termos e as hashtags, respectivamente, selecionados como atributos a partir de cada classe.

4.2.1 Análise das distribuições dos valores extraídos das biografias

Foi realizada uma análise das distribuições dos valores nas biografias coletadas em cada conjunto de dados através de uma inspeção visual da representação gráfica dos dados, a fim de detectar potenciais padrões contidos nelas. A Figura 4.1 contém a distribuição do número de caracteres das biografias no conjunto de dados rotulado por área profissional para cada classe. Para as três áreas profissionais, a maioria das biografias estão próximas de 150 caracteres, o máximo permitido pelo Instagram, com os usuários usando a maioria do espaço dado a eles.

A Figura 4.2 contém a distribuição dos tamanhos das biografias para o conjunto de dados rotulado por faixa etária em cada classe. Novamente ambas as classes apresentam a maioria dos perfis usando biografias com cerca de 150 caracteres, sem nenhuma diferença significativa visível entre os gráficos.

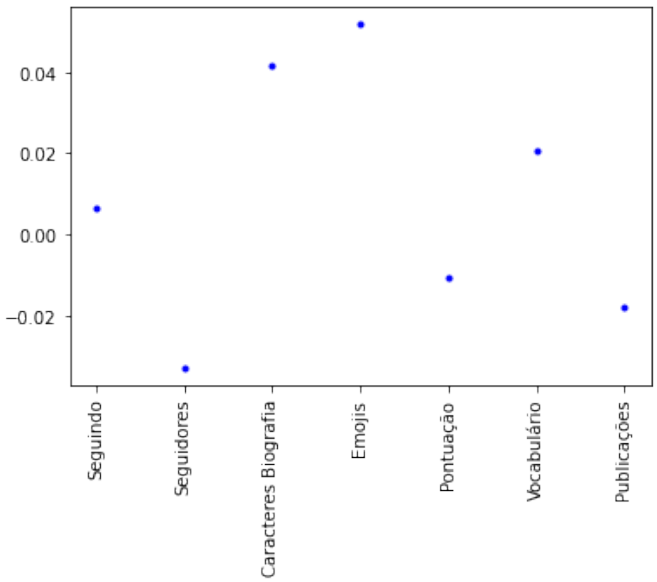


Figura 4.4: Resultados dos testes de correlação ponto-bisserial para o conjunto de dados rotulado por faixa etária

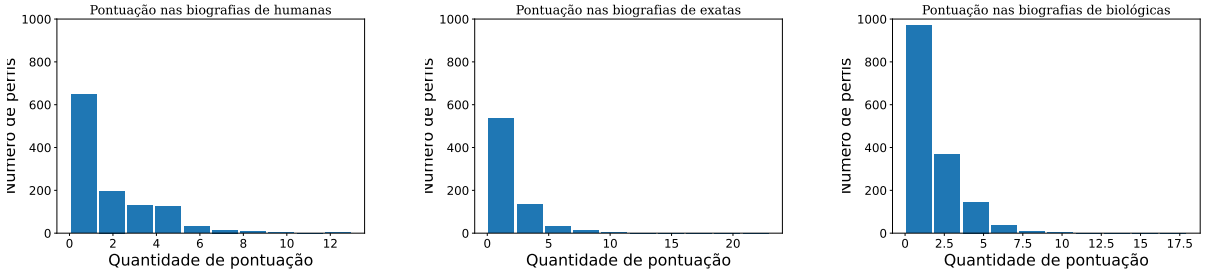
A Figura 4.5 contém os dados para a distribuição da quantidade de sinais de pontuação utilizados nas biografias nos conjuntos de dados rotulados por área profissional. Grande parte das contas não utiliza sinais de pontuação nas biografias, com menos de 15 sinais são utilizados em todas as classes. A Figura 4.5(b) mostra que os perfis de Exatas chegam a alcançar 20 sinais de pontuação utilizados, um limite maior do que as outras classes: 12 para Humanas e 17 de Biológicas, como se vê nas Figuras 4.5(a) e 4.5(c).

A Figura 4.6 mostra as distribuições dos sinais de pontuação nas biografias para faixa etária, em que não aparentam existir diferenças significativas entre as duas classes. As Figura 4.6(a) e 4.6(b) mostram que para adolescentes e adultos, respectivamente, a maioria dos perfis usa menos que dois sinais de pontuação.

As Figuras 4.7 contém os dados para a distribuição da quantidade de emojis utilizados nas biografias nos conjuntos de dados rotulados por área profissional. Em todas as classes, a situação mais comum é não existirem emojis nas biografias. Os perfis de Exatas, conforme a Figura 4.7(b), alcançam um valor maior que os das outras classes, chegando a cada de 30 emojis utilizados. A Figura 4.7(a) classe Humanas a maior quantidade de emojis na biografia encontrada foi 20, significativamente menor.

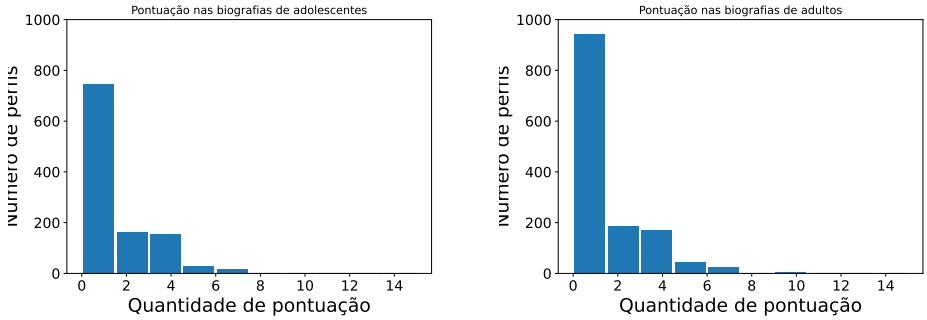
A Figura 4.8 mostra a distribuição da quantidade de emojis usados nas biografias pelos perfis rotulados com faixa etária. Na Figura 4.8(b) é possível perceber que nas contas de adultos o pico de usuários está no uso de cinco emojis, enquanto na Figura 4.8(a) o mais comum é que não sejam usados emojis. De cinco emojis em diante, ambas as classes apresentam uma tendência decrescente no número de perfis.

A Figura 4.9 mostra a distribuição da riqueza de vocabulário no conjunto de dados



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.5: Distribuições da quantidade de sinais de pontuação nas biografias para o conjunto de dados rotulado por área profissional



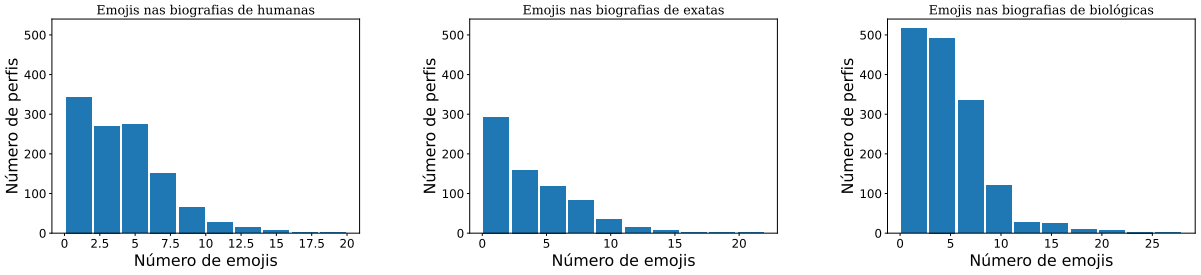
((a)) Adolescentes ((b)) Adultos

Figura 4.6: Histograma da quantidade de sinais de pontuação nas biografias para o conjunto de dados rotulado por faixa etária.

rotulado por área profissional para cada classe. A maioria dos perfis possui uma riqueza de vocabulário acima de $0,6$ para todas as classes, indicando biografias com baixa repetição de palavras independente da área profissional. Um cenário similar se repete na Figura 4.10.

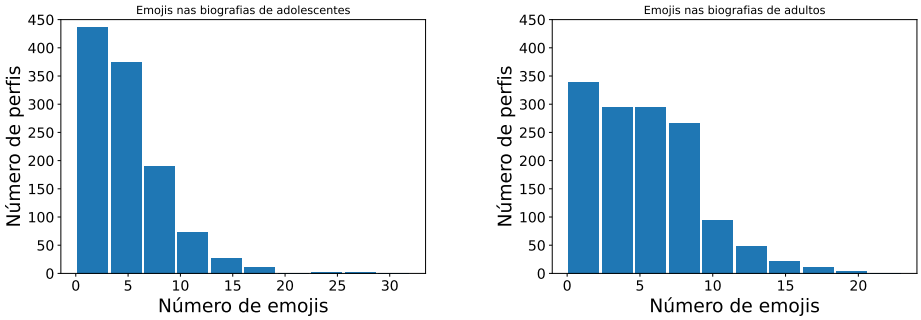
O teste de correlação ponto-bisserial para o número de caracteres na biografia para área profissional aponta baixo poder discriminativo em Exatas e Biológicas, com $r0,0044$ e $r = 0,0098$, respectivamente. Para a correlação em Humanas, o atributo mostra uma correlação próxima de $0,075$, apontando maior capacidade de distinguir essa classe das outras. Para a taxa de emojis, a correlação nos testes com Humanas e Exatas com as demais classes se mostram significativas; enquanto para a taxa de pontuação é em Exatas que o maior mostra uma maior correlação no teste.

Para faixa etária, o número de caracteres apresentou uma correlação de valor $r = 0,04$, correspondendo a uma fraca correlação positiva. Um valor um pouco maior, ainda na faixa de $0,04$ foi alcançado no atributo da taxa de emojis utilizados, contundo ambas são significativamente maiores do que a correlação entre a taxa de pontuação e a faixa etária.



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.7: Distribuições da quantidade de emojis nas biografias para o conjunto de dados rotulado por área profissional



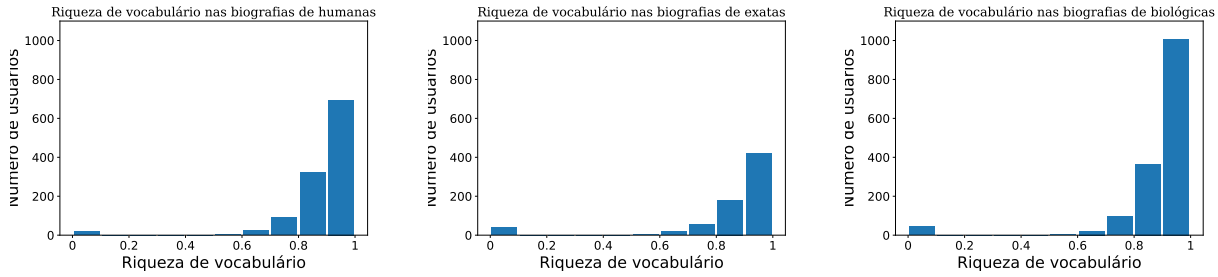
((a)) Adolescentes ((b)) Adultos

Figura 4.8: Histograma da quantidade de emojis nas biografias para o conjunto de dados rotulado por faixa etária.

4.2.2 Análise das *hashtags* selecionadas

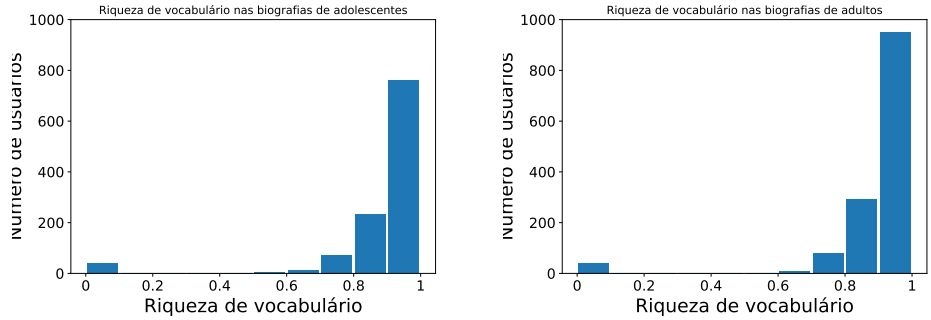
O Apêndice A contém as *hashtags* selecionadas por área profissional separados por classe. Observa-se que para as três áreas profissionais as profissões utilizadas aparecem como *hashtags*, bem como profissões relacionadas. Para a área de humanas, *hashtags* voltados à educação (como "educação", "educaçãoinfantil", "graduação" e "inclusão") aparecem, oriundos dos perfis vindos de pedagogia. Certas *hashtags* femininas, como "professora", "crista", "donadecasa" e "mãe" indicam que este conjunto de dados é composto por uma quantidade significativa de mulheres.

No caso da área de exatas, aparecem *hashtags* referentes à engenharia, como "obra", "engenheiro", "construçao". Outras dizem respeito à arquitetura, como "arqgram", "urbanista", "apecompacto"; o mesmo ocorrendo em ciências contábeis com *hashtags* como "contabilidade", "contadora", "finanças". Além dessas *hashtags* fortemente relacionados, outros aparecem relacionados ao oferecimento de serviços: "consultora", "projetosonline", "projetosautorais", "auditoria". Ao mesmo tempo, existe uma variedade de *hashtags* não relacionadas à profissão, como "winelover", "amoesporte" e "atletacristã". Isso indica que enquanto parte dos perfis dessa classe usam o Insta-



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.9: Histograma da riqueza de vocabulário nas biografias para o conjunto de dados rotulado por área profissional



((a)) Adolescentes ((b)) Adultos

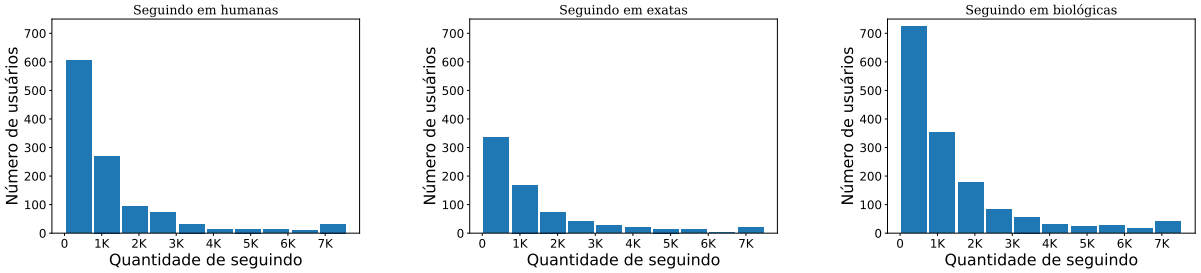
Figura 4.10: Histograma da riqueza de vocabulário nas biografias para o conjunto de dados rotulado por faixa etária.

gram para fins comerciais, outros mantém perfis mais voltados para o pessoal.

Nas ciências biológicas, praticamente todos as *hashtags* são relacionados diretamente às profissões de algum modo: para enfermagem, “emergencia”, “cruzvermelha” e “uti”; para educação física, “bodybuild”, “exerciciofisico” e “personaltrainer”, para psicologia, “autocuidado” e “terapia” são exemplos. Diversas *hashtags* no feminino aparecem, indicando que este conjunto de dados também possui quantidade significativa de mulheres.

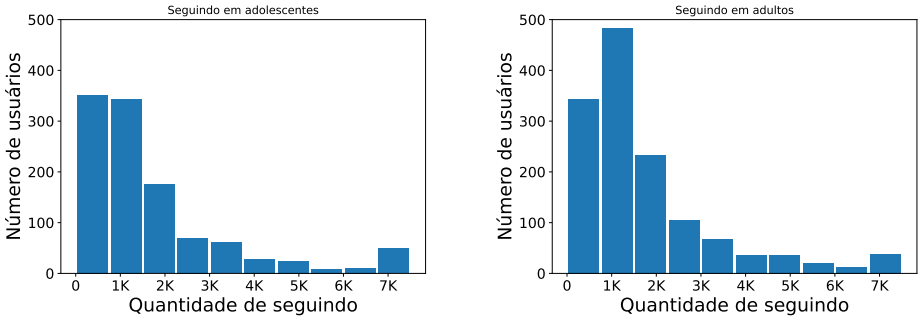
O Apêndice B contém as *hashtags* selecionadas por faixa etária separados por classe. Ambas as classes apresentam *hashtags* com menções de idade, ainda que em formas modificadas como “22primaveras”. Também nas duas há uma grande quantidade de termos femininos, indicando uma representação excessiva de mulheres em todo o conjunto de dados.

O que salta aos olhos imediatamente é como para adultos, *hashtags* citando profissões, ou relacionadas à elas, aparecem, como “administraçãodeempresas”, “educaçãofísica”, “empreendedora” e “pedagogia”. Além do mais, termos ligados à maternidade



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.11: Histograma da quantidade de contas seguidas para o conjunto de dados rotulado por área profissional



((a)) Adolescentes ((b)) Adultos

Figura 4.12: Histograma da quantidade de contas seguidas para o conjunto de dados rotulado por faixa etária.

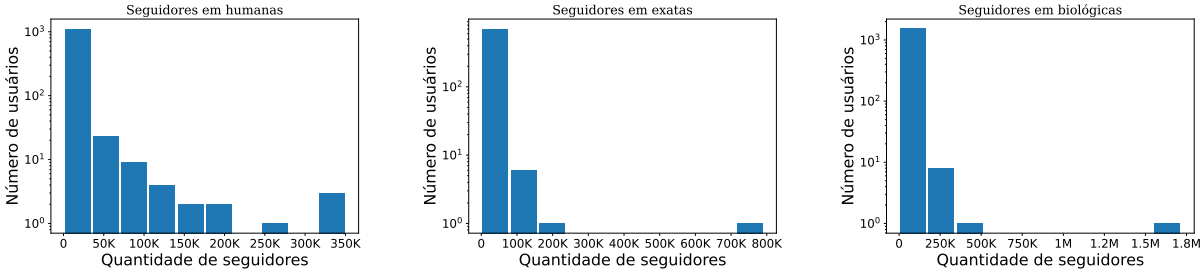
estão presentes nessa classe. Isso está de acordo com o tipo de atividade mais madura que adultos exporiam na rede.

Para adolescentes, as *hashtags* envolvem diversos temas de entretenimento: “ator”, “atriz”, “cantor”, “dj”, “sertanejo” e “teatro”. Além disso, *hashtags* em inglês aparecem, como “act”, “singer” e “theater”. Isso aponta para biografias com tom mais descontraído por parte dos adolescentes.

4.2.3 Análise dos termos selecionados

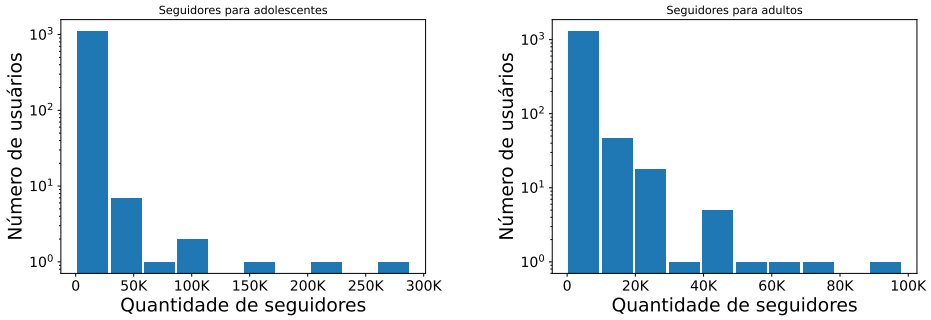
O Apêndice C contém os termos selecionados por área profissional separados por classe. Assim como nas *hashtags*, as profissões utilizadas nas buscas aparecem várias vezes nos termos selecionados, assim como termos relacionados. Alguns termos são comuns a todas as classes, como “via”, “whatsapp”, “acadêmica” e “ajudo”. Os dois primeiros termos apontam para perfis disponibilizando contato através do aplicativo Whatsapp, e o último para o oferecimento de serviços.

Os termos selecionados para os perfis de humanas deixam de ter o domínio da



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.13: Histograma da quantidade de seguidores para o conjunto de dados rotulado por área profissional. O eixo y está em escala logaritmica.



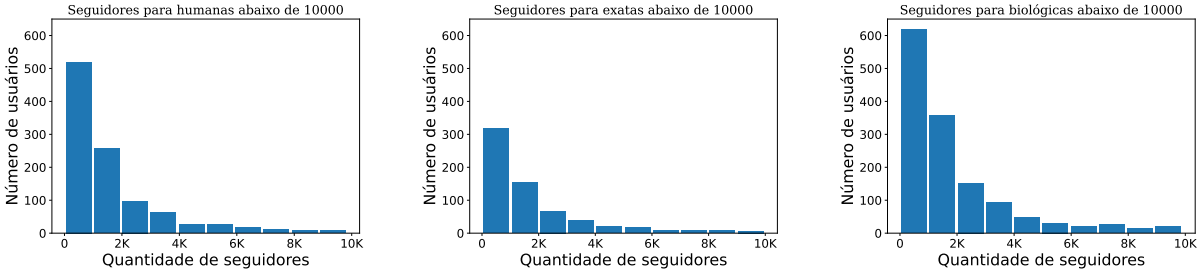
((a)) Adolescentes ((b)) Adultos

Figura 4.14: Histograma da quantidade de seguidores para o conjunto de dados rotulado por faixa etária. O eixo y está em escala logaritmica.

pedagogia nas *hashtags*, e termos relacionados à direito e adminstração também estão presentes. Novamente há uma quantidade considerável de termos no feminino (“apaixonada”, “acadêmica”, “formada”, “graduada”, “mãe”, “psicopedagoga”, são exemplos) reforçando que a classe seja composta por mulheres em sua maioria.

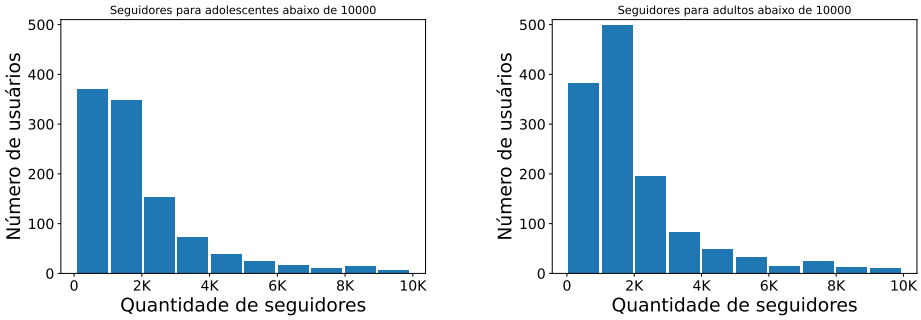
No caso das exatas, existem diversas menções a estados, como “mg”, “rio”, “rs”, “sc” e “sp” evidenciam que os perfis querem evidenciar sua localização. Além disso, os termos “acesse”, “clique” e “http” apontam para a apresentação de links para o leitor. Unindo isso ao uso de termos como “cursos”, “orçamento”, “projetos” e “whatsapp”, os termos selecionados apontam para a classe exatas contendo diversos profissionais anunciando seus serviços através do Instagram.

A classe biológicas é a única em que os termos “empreededor”/“empreendedora” não aparecem. Termos como “coach”, “palestrante”, “terapeuta” e “crp” estão ligados a psicológos descrevendo sua atuação, e possivelmente anunciando serviços. A proeminência de termos da enfermagem é reduzida, em comparação com as *hashtags*, e termos das outras duas profissões aparecem em maior quantidade.



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.15: Histograma da quantidade de seguidores para o conjunto de dados rotulado por área profissional contendo perfis com até 10.000 seguidores



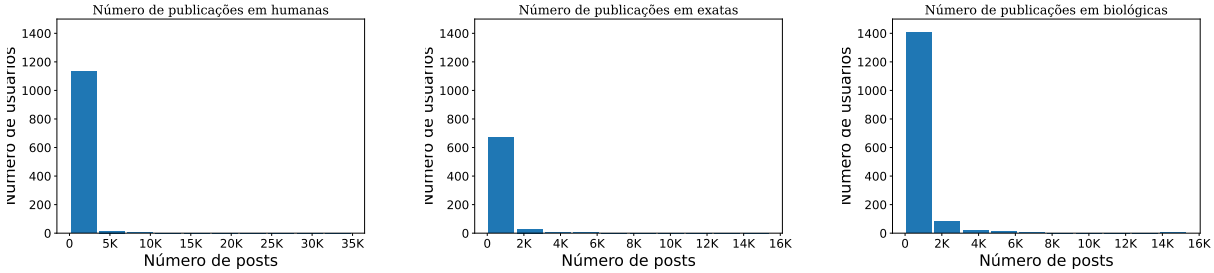
((a)) Adolescentes ((b)) Adultos

Figura 4.16: Histograma da quantidade de seguidores para o conjunto de dados rotulado por faixa etária, contendo perfis com até 10.000 seguidores

O Apêndice D contém os termos selecionados por faixa etária separados por classe. As idades pesquisadas aparecem como termos significativos para ambas as classes. No caso dos adolescentes, o ano de nascimento de uma das idades, 2003 (correspondente a 18 anos quando da conclusão deste trabalho) está presente.

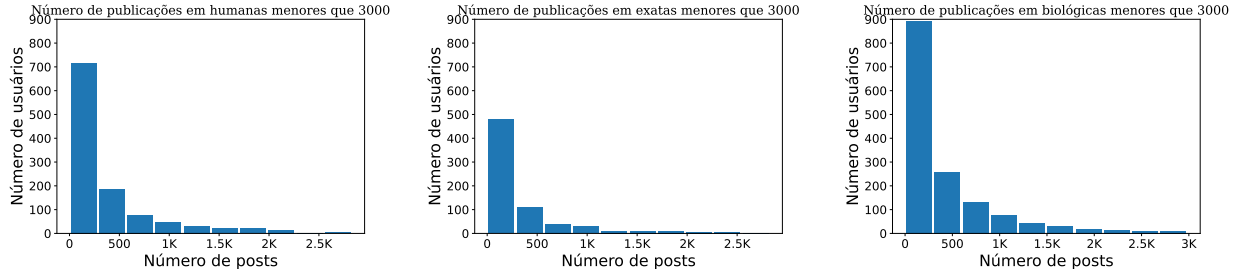
Para adultos, termos envolvendo relacionamentos mais sólidos, como “mãe” e “casada” e menções à filhos ocorrem, bem como profissões ausentes nos termos selecionados (“designer”, “engenheira”, “fisioterapeuta”, “pedagoga”). Os termos selecionados confirmam a impressão de que grande parte da base de dados é composta por perfis de mulheres, visto que há diversos termos femininos.

Dentre os termos selecionados para adolescentes, menções à outras redes sociais, como Facebook e Tiktok, estão presentes, mostrando um interesse nessa faixa etária de expandir suas conexões virtuais.



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.17: Histograma do número de publicações para o conjunto de dados rotulado por área profissional



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.18: Histograma do número de publicações para o conjunto de dados rotulado por área profissional para perfis com até 3000 publicações

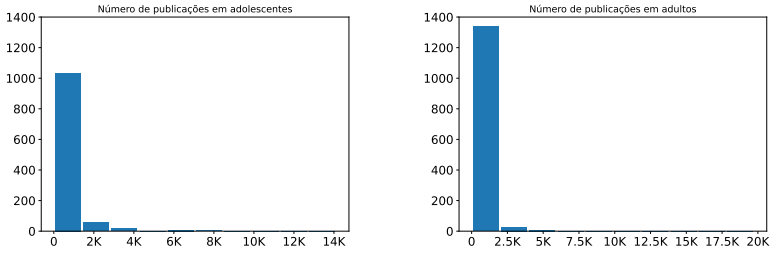
4.3 Análise dos padrões de seguir

As análises feitas na Seção 4.2 serão repetidas nesta Seção para o número de seguidores e de contas seguidas em cada conjunto de dados: uma inspeção dos gráficos de cada uma das classes e testes de correlação ponto-bisserial a fim de verificar se as classes seguem distribuições similares ou não para cada atributo.

A Figura 4.11 exibe as distribuições da quantidade de contas seguidas para área profissional. As distribuições não parecem possuir diferenças significativas, com a maioria das contas seguindo menos de 1.000 usuários.

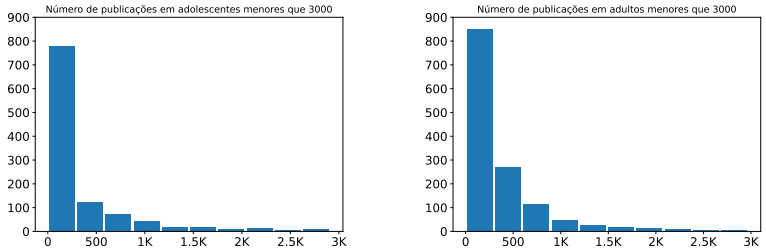
Na Figura 4.12 estão os gráficos da distribuição de contas seguidas para faixa etária. Para adultos, na Figura 4.12(b) pode-se ver que o pico da distribuição está entre 1.000 e 2.000 contas seguidas, enquanto para adolecentes, na Figura 4.12(a), em até 1.000.

A Figura 4.13 exibe as distribuições da quantidade de seguidores para área profissional. Todas as classes exibem perfis com uma quantidade muito grande de seguidores, chegando a mais de um milhão na Figura 4.13(c) para Biológicas, seguido de 800.000 na Figura 4.13(b) para Exatas e 350.000 na Figura 4.15(a) para Humanas. Por essa



((a)) Adolescentes ((b)) Adultos

Figura 4.19: Histograma do número de publicações para o conjunto de dados rotulado por faixa etária



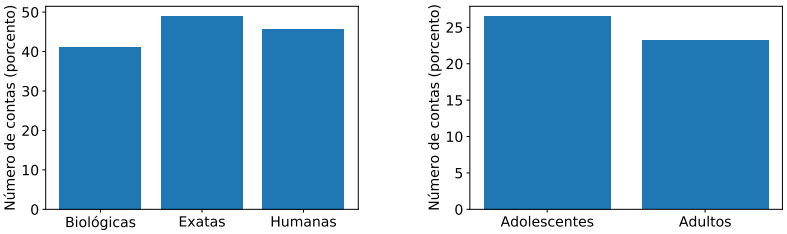
((a)) Adolescentes ((b)) Adultos

Figura 4.20: Histograma do número de publicações para o conjunto de dados rotulado por faixa etária para perfis com até 3000 publicações

razão os gráficos acabam por estar distorcidos, com uma concentração de resultados no início. Para resolver esse problema, os gráficos foram refeitos, considerando até 10.000 seguidores.

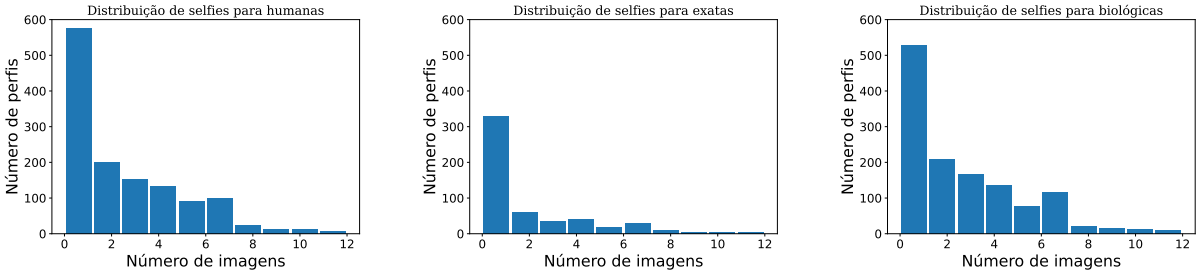
Na Figura 4.14 estão os gráficos da distribuição de seguidores para faixa etária. Novamente valores muito grandes, na casa dos centenas de milhares, acabam por gerar gráficos distorcidos: para adolescentes existem perfis com mais de 250.000 seguidores (Figura 4.14(a)), enquanto pra adultos o máximo está na casa de 100.000 (Figura 4.14(b)). Novamente foram criados gráficos com perfis com até 10.000 seguidores, na Figura 4.16. Pode-se observar, na Figura 4.16(b), que para adultos, há um pico na gráfico entre 1.000 e 2.000 seguidores, alcançando 500 perfis, diferente de para adolescentes, que na Figura 4.16(a) apresentam um pico de perfis em até 1.000 seguidores.

Os testes de correlação para área profissional envolvendo seguidores mostram como nas três classes o resultado está bem próximo de 0, revelando baixo potencial de usar esse atributo para determinar área profissional. Os testes de correlação para o número de perfis seguidos revelam um quadro parecido, com apenas o teste para Exatas estando um pouco mais distante do centro.



((a)) Área profissional ((b)) Faixa etária

Figura 4.21: Porcentagem das contas de negócio para o conjunto de dados rotulado por faixa etária e área profissional.



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.22: Histograma do número de imagens apontadas como *selfies* para área profissional

A correlação obtida entre a faixa etária e o número de seguidores esteve perto de $r = -0,04$, enquanto para seguidos r é bem próximo de 0. Isso aponta para o número de seguidores como um melhor discriminador de faixa etária.

4.4 Análise do número de publicações

A distribuição do número de publicações feitas pelos perfis em cada área profissional está na Figura 4.17. Nota-se que a maioria dos perfis possui menos de 1000 publicações, com poucos casos estando nas dezenas de milhares. Os gráficos foram refeitos considerando apenas os perfis com até 3000 publicações, conforme a Figura 4.18. Quando os *outliers* são removidos, nota-se uma distribuição similar entre as classes, com a maioria tendo menos de 500 publicações. Os testes de correlação ponto-bisserial para Biológicas e Exatas ficaram próximos a 0, apenas para Humanas há uma correlação de $r = 0,051$ mais distante da origem. Ao todo, o atributo parece ter um poder discriminativo baixo nesse cenário.

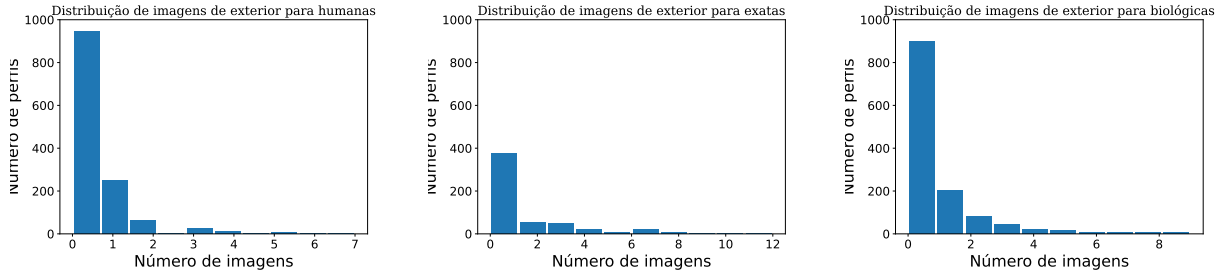
A distribuição do número de publicações feitas pelos perfis por faixa etária está na Figura 4.19. Novamente existem *outliers* na casa nos dezenas de milhares, então

o gráfico foi refeito considerando os perfis com no máximo 3000 publicações. O resultado está na Figura 4.20, onde pode-se perceber que ainda que para ambas as classes haja uma quantidade significativa de perfis com menos de 500 publicações, para adultos a quantidade de perfis com publicações acima desse número é um pouco maior. Isso indica uma tendência à adultos postarem mais, conforme identificado em trabalhos anteriores. O teste de correlação ponto-bisserial nesse cenário resultou num valor por volta de $-0,02$, indicando baixo potencial discriminativo do atributo.

4.5 Análise do atributo de negócio

A Figura 4.21(a) contém a porcentagem dos perfis marcados como de negócio para as classes de área profissional. Os perfis de exatas têm ampla maioria sobre as outras duas classes, e os de biológicas possuem a menor porcentagem. Isso segue a tendência vista nos termos selecionados de que os perfis de exatas estão voltados para oferecer serviços através do Instagram.

Na Figura 4.21(b) estão as porcentagens para os perfis de negócio no conjunto de dados rotulado por faixa etária. Os valores são bem similares para as duas classes, pouco acima dos 25%. Isso contraria a expectativa que contas de adultos usariam o Instagram para fins comerciais e estariam marcadas como de negócio num grau maior.

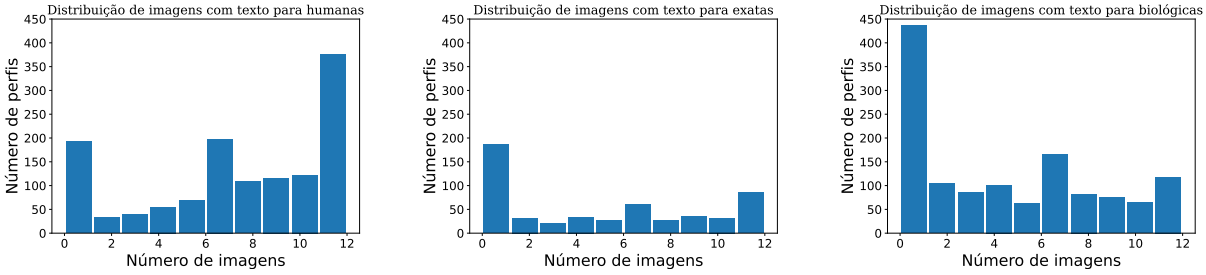


((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.23: Histograma do número de imagens apontadas como de exterior para área profissional

4.6 Atributos de imagem

As distribuições dos atributos oriundos das imagens para o conjunto de dados rotulado por área profissional estão nas Figuras 4.22 a 4.26. A Figura 4.24 exibe uma quantidade maior de imagens que contenham texto por perfis de Humanas, contrário às outras duas classes, que apresentam a maioria dos perfis com poucas imagens. A partir desses dados é possível verificar que profissões como direito e pedagogia façam publicações informativas de seu campo de atuação que consistam unicamente de texto.

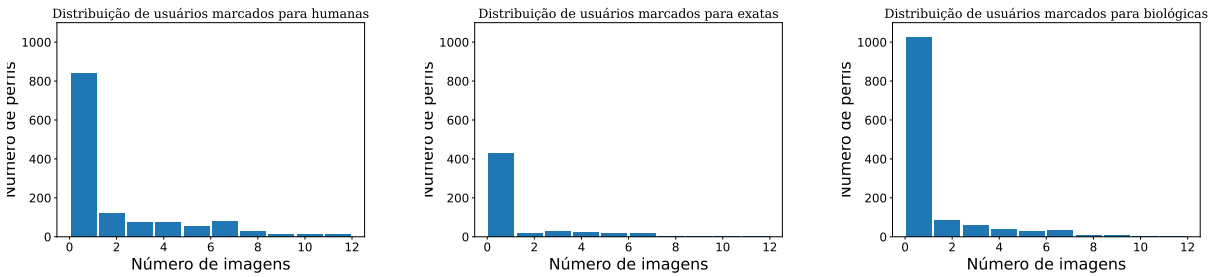


((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.24: Histograma do número de imagens contendo texto para área profissional

A Figura 4.23 mostra que a única classe que possui perfis com um número de imagens de exterior maior que oito é a de Exatas, ainda que em pequena quantidade. Isso indica uma maior propensão da outras classes estarem focadas em apresentar suas profissões e os pertencentes à Exatas publicarem fotos em atividades mais variadas, ao ar livre.

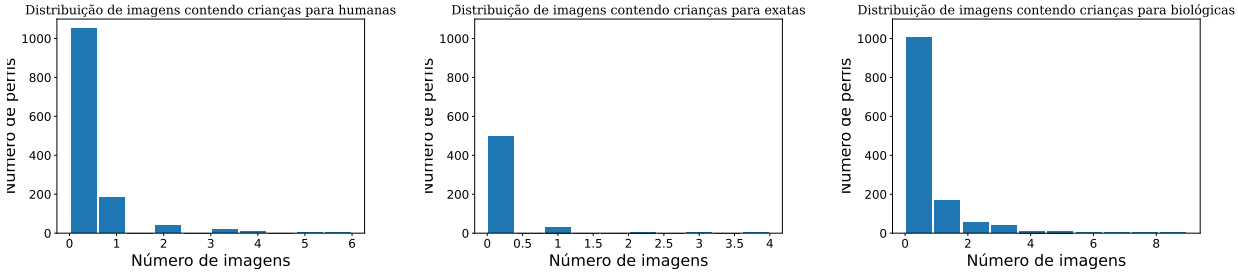
Os demais atributos de imagem, *selfies* (Figura 4.22); imagens com usuários marcados (Figura 4.25) e imagens contendo crianças (Figura 4.26) não aparentam ter diferenças significativas entre eles.



((a)) Humanas ((b)) Exatas ((c)) Biológicas

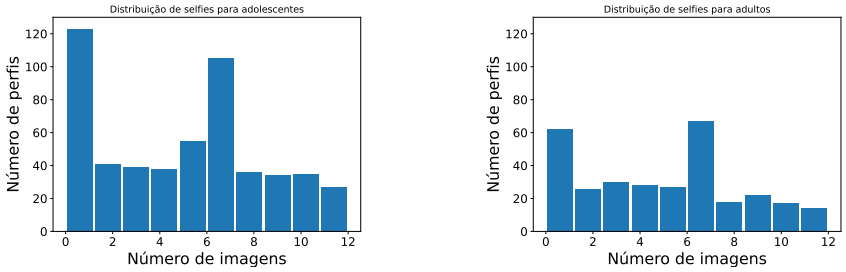
Figura 4.25: Histograma do número de imagens com usuários marcados para área profissional

Para a faixa etária, por motivos de tempo e devido às restrições de coleta de dados recebidas pelo Instagram, não foi possível fazer a coleta dos atributos de imagem para todo o conjunto de dados. Assim, foi feita a coleta para uma parcela do conjunto, permitindo a avaliação da faixa etária por meio deste tipo de atributo, ainda que em escala menor. Foram coletadas os dados oriundos das imagens de 844 perfis, sendo 533 de adolescentes e 311 de adultos. Os resultados estão nas Figuras 4.27 a 4.31. Há pouca diferença nas distribuições para as duas classes. Adolescentes só chegam a publicar até oito imagens de exterior, enquanto adultos chegam até doze, mas numa quantidade pequena (Figura 4.28). Há uma boa quantidade de perfis nas



((a)) Humanas ((b)) Exatas ((c)) Biológicas

Figura 4.26: Histograma do número de imagens apontadas contendo crianças para área profissional



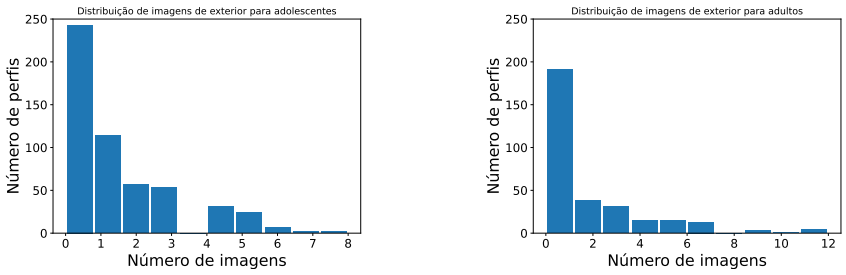
((a)) Adolescentes ((b)) Adultos

Figura 4.27: Histograma do número de imagens apontadas como *selfies* para faixa etária

duas faixas etárias publicando poucas selfies (menos que 6), ao contrário do esperado pelo descrito na literatura.

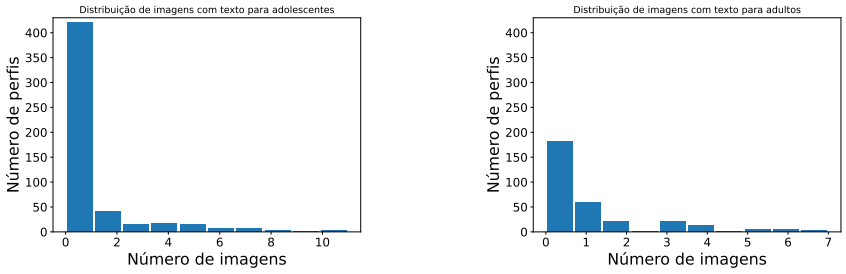
4.7 Relação dos termos usados nos textos alternativos das imagens

Os textos alternativos das imagens obtidos foram analisados com o fim de obter o vocabulário que o Instagram utiliza para descrever as imagens. A análise tem como objetivo uma melhor compreensão dos conjuntos de dados e não a utilização dos termos identificados na classificação, mas pode servir como base para trabalhos futuros o fazerem. Uma série de passos de pré-processamento foi aplicada nesses textos: primeiro, como as descrições obtidas vieram na língua inglesa, foram removidas as palavras em português. Segundo, foram removidos os termos usados na introdução do conteúdo das imagens: *May be an image of* e *May be a*, bem como o termo *text* que introduz as descrições de texto. Terceiro, as indicações de local e horário da postagem foram removidas; e por último os caracteres não-alfabéticos foram eliminados.



((a)) Adolescentes ((b)) Adultos

Figura 4.28: Histograma do número de imagens apontadas como de exterior para faixa etária

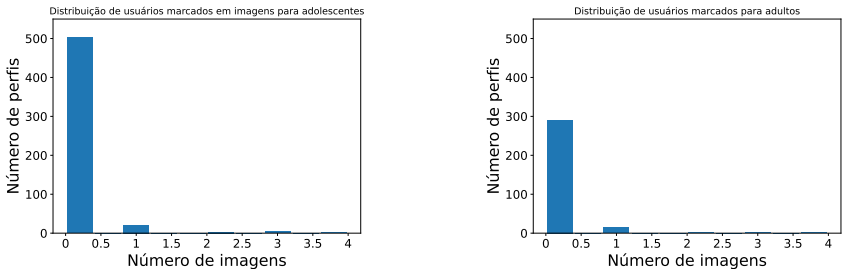


((a)) Adolescentes ((b)) Adultos

Figura 4.29: Histograma do número de imagens contendo texto para faixa etária

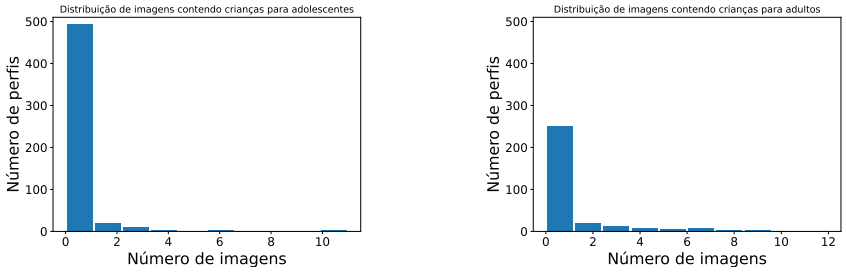
O Apêndice E contém o vocabulário extraído por esse método a partir do conjunto de dados rotulado por área profissional, totalizando 176 termos. Certos termos indicam elementos dentro das imagens, como *bicycle*, *cat*, *person* e *flower*. Outras representam o ambiente em que a imagem se passa, como *beach*, *office* e *outdoors*. Existem também os que apontam para propriedades da imagem, como *closeup*, *black-and-white*, e *meme*. Dos 176 termos, 59 estão presentes em todas as classes, 62 estão presentes apenas em Humanas, 19 apenas em Biológicas e 6 apenas em Exatas.

O vacabulário extraído do conjunto de dados rotulado por faixa etária está no Apêndice F, num total de 123 ternos. Deles, 53 estão presentes nas duas faixas etárias, 6 presentes apenas em Adolescentes, e 64 são oriundos apenas de Adultos. O fato que poucos termos sejam exclusivos de adolescentes, a despeito dessa classe possuir mais perfis com imagens, indica que existe pouco nas imagens publicadas por adolescentes que lhes diferenciem das dos adultos. A saber, os termos são: *heart*, *fries*, *bottle*, *parrot*, *french* e *car*, sendo evidente que dois deles originalmente apareciam juntos: *french fries* (batatas fritas). Todos apontam para elementos nas imagens e não para atividades ou cenários em que elas se passam. Os termos exclusivos aos adultos, por sua vez, não apenas contém elementos das imagens (como *person*, *horse*, *wall*) mas também cenários em que elas se passam (*dune*, *bedroom*, *mountain*), atividades (*fishing*, *kissing*, *sitting*). Também há termos como *twitter* e *screenshot* que apontam



((a)) Adolescentes ((b)) Adultos

Figura 4.30: Histograma do número de imagens com usuários marcados para faixa etária



((a)) Adolescentes ((b)) Adultos

Figura 4.31: Histograma do número de imagens apontadas contendo crianças para faixa etária

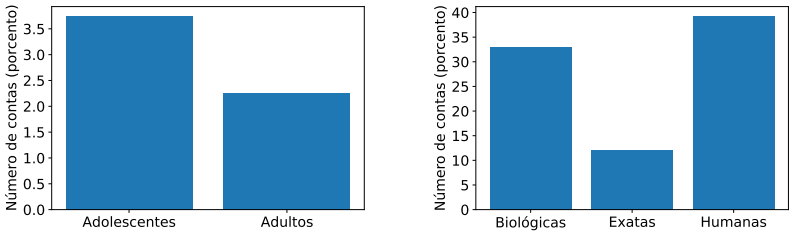
para publicação de conteúdo de outras redes sociais.

4.8 Marcadores de classe nos *usernames*

Song et al. (2018) notou a tendência de usuários informarem o ano de nascimento nos *usernames*, construindo com essa informação um segundo conjunto de dados com 200 perfis para realização de testes adicionais. Baseado nisso, foi calculada a porcentagem de indivíduos no conjunto de dados rotulados por faixa etária cujos *usernames* continham anos potencialmente de nascimento, seja com quatro ou dois digitos (considerando as faixas etárias abordadas, os anos vão de 1995 a 2005). Diferentemente do trabalho mencionado, o objetivo aqui não é criar outro conjunto de dados nem fazer outros testes, mas apenas verificar a viabilidade desse indicador para coleta de usuários no futuro.

O resultado está na Figura 4.32(a). Cerca de 3% dos perfis de cada faixa etária possuem tais termos, o que a primeira vista parece indicar baixa capacidade de discriminação deles como indicadores de idade. Contudo Song et al. (2018) através deste método obtém apenas 200 perfis, com o conjunto obtido anteriormente

consistindo de 12.500 perfis, numa proporção de 0,016, similar à obtida aqui.



((a)) Faixa etária ((b)) Área profissional

Figura 4.32: Porcentagem das contas cujos *usernames* contém termos relevantes para o conjunto de dados rotulado por faixa etária e área profissional.

Para o conjunto de dados rotulado por área profissional, algo similar foi feito: considerando que os perfis teriam *usernames* contendo termos relacionados à profissão do perfil, foi contabilizada a porcentagem dos *usernames* contendo pelo menos a primeira sílaba dos termos utilizados na pesquisa dos perfis. O resultado está na Figura 4.32(b). Para todas as classes, cerca de 30% dos perfis usam os termos selecionados. Isso parece indicar que perfis que pertençam à uma profissão tenham certa tendência à expô-la em seu *username* e que este método possa ser um adicional na coleta de perfis por área profissional.

Capítulo 5

Classificação

Esta seção trata do processo de classificação. Primeiro, na Seção 5.1 é feita uma descrição da metodologia da classificação realizada, tratando dos métodos e das medidas de desempenho utilizados. Após isso, nas seções 5.2 e 5.3, os resultados da classificação para área profissional e idade são expostos.

Os resultados expostos aqui apontam para certos atributos como os mais discriminadores para área profissional e faixa etária. A seleção dos termos por tf-idf e das *hashtags* mais frequentes usadas nas biografias dos perfis aparecem como atributos com performances boas, mas é a combinação de todos os atributos que na maioria dos casos leva ao melhor resultado.

5.1 Metodologia

5.1.1 Métodos de aprendizagem de máquina

Os dados de cada usuário foram transformados em vetores e inseridos em métodos de aprendizagem de máquina. Dois métodos foram considerados: *Random Forest* e *Support Vector Machines*, nas implementações da biblioteca *Scikit-learn*[1] em python Pedregosa et al. (2011).

Estes dois métodos são utilizados por muitos dos trabalhos relacionados: *Random Forest* por Filho et al. (2014); Rodrigues et al. (2017); Song et al. (2018) e *Support Vector Machines* por Filho et al. (2014, 2016); Rodrigues et al. (2017). Ambos os métodos podem lidar com grande quantidades de dados e são resistentes a *overfitting* (Han et al., 2011; Breiman, 2001).

Os parâmetros utilizados nos dois métodos são os valores padrão do *Scikit-learn*. Para o SVM, é utilizada a função de *kernel* função radial de base, representado pela seguinte formulação (Chang and Lin, 2011):

[1] https://scikit-learn.org

$$K(x, x^{'}) = e^{-\gamma||x-x^{'}||} \tag{5.1}$$

Em que γ é um parâmetro livre a ser definido, cujo valor padrão na biblioteca é:

$$\gamma(\text{padrão}) = \frac{1}{\text{número de atributos} * \text{variância do conjunto de treino}} \tag{5.2}$$

Conforme a documentação do *Scikit-learn*[2] o parâmetro γ define o grau de influência de cada dado de treino, numa relação inversa. Quão maior for γ, mais próximos outros dados devem estar dos vetores suporte para ser afetados por ele.

Além deste parâmetro, o SVM possui o parâmetro de regularização C. Ele determina o tamanho da margem criada pelo hiperplano de separação: quão maior for o valor do C, menor será a margem obtida. A depender do tamanho da margem, *outliers* no conjunto de dados podem acabar sendo ignorados. O *Scikit-learn* utiliza um valor padrão para C igual a 1. A implementação do SVM utilizada lida com classificação não-binária através da estratégia um-contra-um.

Para *Random Forest*, existem os seguintes parâmetros pertinentes para a *performance*:

- Número de árvores: 100;
- Métrica utilizada para dividir os nós: índice de Gini;
- Número mínimo de elementos para dividir um nó: 2;
- Número máximo de atributos considerados na divisão de um nó: raíz quadrada do total de atributos.

5.1.2 Setup experimental utilizado

Uma vez definidos os métodos de aprendizagem de máquina a usar para classificação, foi elaborado o *setup* experimental em que eles seriam aplicados. Primeiro, os atributos descritos no capítulo 4 (textuais, comportamentais e visuais) foram reunidos em sub-grupos conforme sua similaridade para serem utilizados em conjunto na classificação. A relação dos sub-grupos e os atributos pertencentes a cada um estão na Tabela 5.1. Na primeira coluna estão os agrupamentos formados: Seguir, Comportamento (atributos comportamentais), Biografia, *Hashtags*, Termos (atributos textuais) e Imagem (atributo de imagem). Na segunda, os atributos presentes em cada grupo.

Cada sub-grupo passou pelos dois métodos separadamente. A classificação foi feita através de *5-fold cross validation*. Devido ao problema de desbalanceamento das

[2]https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

Categoria	Atributo
Biografia	Taxa de pontuação utilizada (Seção 3.2.2) Taxa de emojis utilizados (Seção 3.2.2) Riqueza de vocabulário (Seção 3.2.2) Número de caracteres (Seção 3.2.2)
Seguir	Número de seguidores (Seção 3.2.1) Número de seguidos (Seção 3.2.1)
Comportamento	Número de posts (Seção 3.2.1) Identificador de conta de negócio (Seção 3.2.1)
Termos	200 termos da biografia mais relevantes (Seção 3.2.2)
Hashtags	100 *hashtags* da biografia mais frequentes (Seção 3.2.2)
Imagens	Quantidade de *selfies* Quantidade de imagens de paisagem (Seção 3.2.3) Quantidade de imagens com texto (Seção 3.2.3) Quantidade de imagens com usuários marcados (Seção 3.2.3) Quantidade de imagens com crianças (Seção 3.2.3)

Tabela 5.1: Lista de atributos utilizados na classificação, separados pelos agrupamentos em que foram utilizados.

classes nos conjuntos de dados, a classificação foi feita utilizando a opção de balanceamento das classes oferecida pelo *Scikit-learn* através do parâmetro *class_weight*. Nela, os valores das classes recebem pesos inversamente proporcionais às frequências das classes, conforme a expressão

$$\frac{\text{número de amostras}}{\text{número de classes} \ * \ \text{frequência de cada classe}} \tag{5.3}$$

Os atributos foram normalizados antes da execução dos classificadores, através do normalizador *StandardScaler* do *Scikit-learn*.[3] Ele transforma os dados de modo que possuam média zero e variância unitária, conforme a Equação 5.4:

$$z = \frac{x - u}{s} \tag{5.4}$$

Nessa equação u é a média dos dados, s a variância dos dados, x o valor original do atributo e z o valor normalizado. Esta normalização é importante para o uso de classificadores como SVM, em que alguns *kernels* (como o RBF utilizado neste trabalho) exigem que os dados possuam média nula e variância unitária, podendo apresentar desempenho ruim caso contrário.

[3]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

5.1.3 Medidas de avaliação utilizadas

As performances de cada execução para cada um dos grupos de atributos foram comparadas entre si, usando as medidas acurácia (Seção 2.4.2) e Medida F1 (Seção 2.4.5). Logo, foram criados seis modelos, cada um deles com duas medidas de avaliação.

Essas medidas são utilizadas na maioria dos trabalhos relacionados: acurácia em Argamon et al. (2009), Filho et al. (2014), Álvarez-Carmona et al. (2016), Han et al. (2016), (Campos, 2016) e Medida F1 em Filho et al. (2014), Rodrigues et al. (2017), Song et al. (2018), Campos (2016).

5.2 Classificação de perfis por área profissional

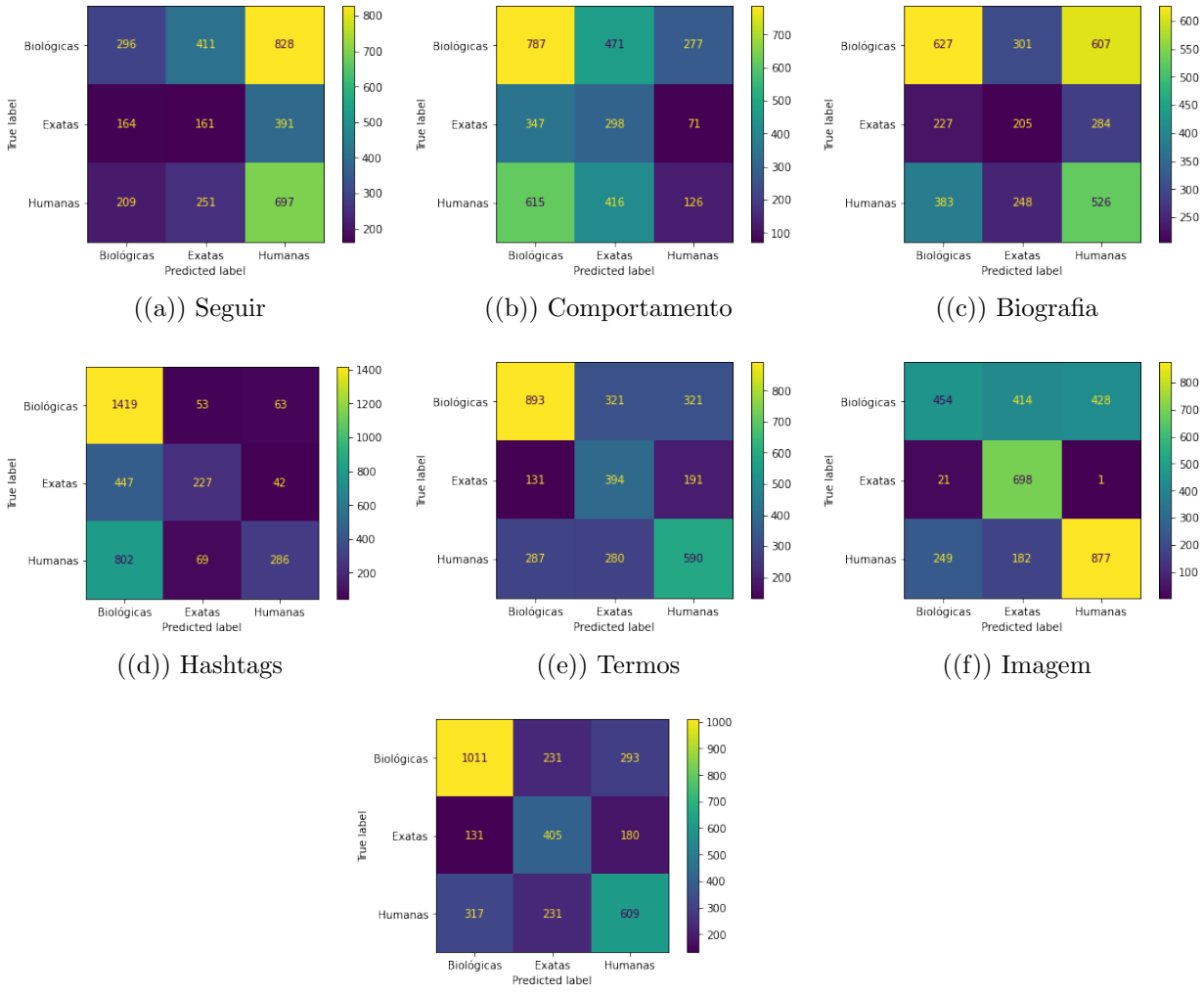
Os perfis foram agrupados por área profissional, levando à três classes: exatas, humanas e biológicas. A classificação foi realizada para cada um dos grupos de atributos descritos na Seção 2.1, para os classificadores *Random Forest* e SVM. Por fim uma classificação foi feita usando todos os atributos reunidos.

As matrizes de confusão para a classificação e os grupos de atributos para o SVM estão na Figura 5.1. A Figura 5.1(a) se refere à classificação pelos grupo de atributos Seguir e exibe uma concentração de predições na classe Humanas, com 1.916 elementos sendo colocados nessa classe pelo modelo. As Figuras 5.1(b) e 5.1(d) mostram que a matriz de confusão para os atributos de Comportamento e *Hashtags* as predições se concentram na classe Biológicas, com 1.749 e 2.668 elementos colocados nelas, respectivamente. A matriz de confusão para os atributos de Biografia, na Figura 5.1(c), mostra que as predições estão principalmente balanceadas entre as classes Biológicas e Humanas, com 1.237 e 1.417 elementos atribuídos a elas, respectivamente. Isso indica que estes quatro grupos de atributos, assim, sejam pouco úteis para a descoberta de área profissional.

Na Figura 5.1(e) há um padrão diferente, em que apesar de existirem muitas predições em Biológicas, o modelo consegue prever mais casos nas outras duas classes, com a diagonal contendo 2.025 elementos. A Figura 5.1(g) mostra que combinar todos os atributos leva igualmente a uma concentração de elementos na diagonal, totalizando 2.160 perfis corretamente classificados.

A Figura 5.1(f) mostra que a matriz de confusão no caso dos atributos de Imagem classifica apenas 22 perfis de Exatas incorretamente. O maior problema nesse modelo está nos perfis de Biológicas, cujas predições estão distribuídas de maneira aproximadamente igual entre as três classes.

A Figura 5.2 contém a matriz de confusão para a classificação *Random Forest* em cada agrupamento de atributos. Fica evidente que os resultados são mais balanceados em comparação com a classificação com SVM. Nas Figuras 5.2(a), 5.2(b) e 5.2(c), nota-se que cada célula das matrizes de confusão contém valores similares en-



((a)) Seguir ((b)) Comportamento ((c)) Biografia

((d)) Hashtags ((e)) Termos ((f)) Imagem

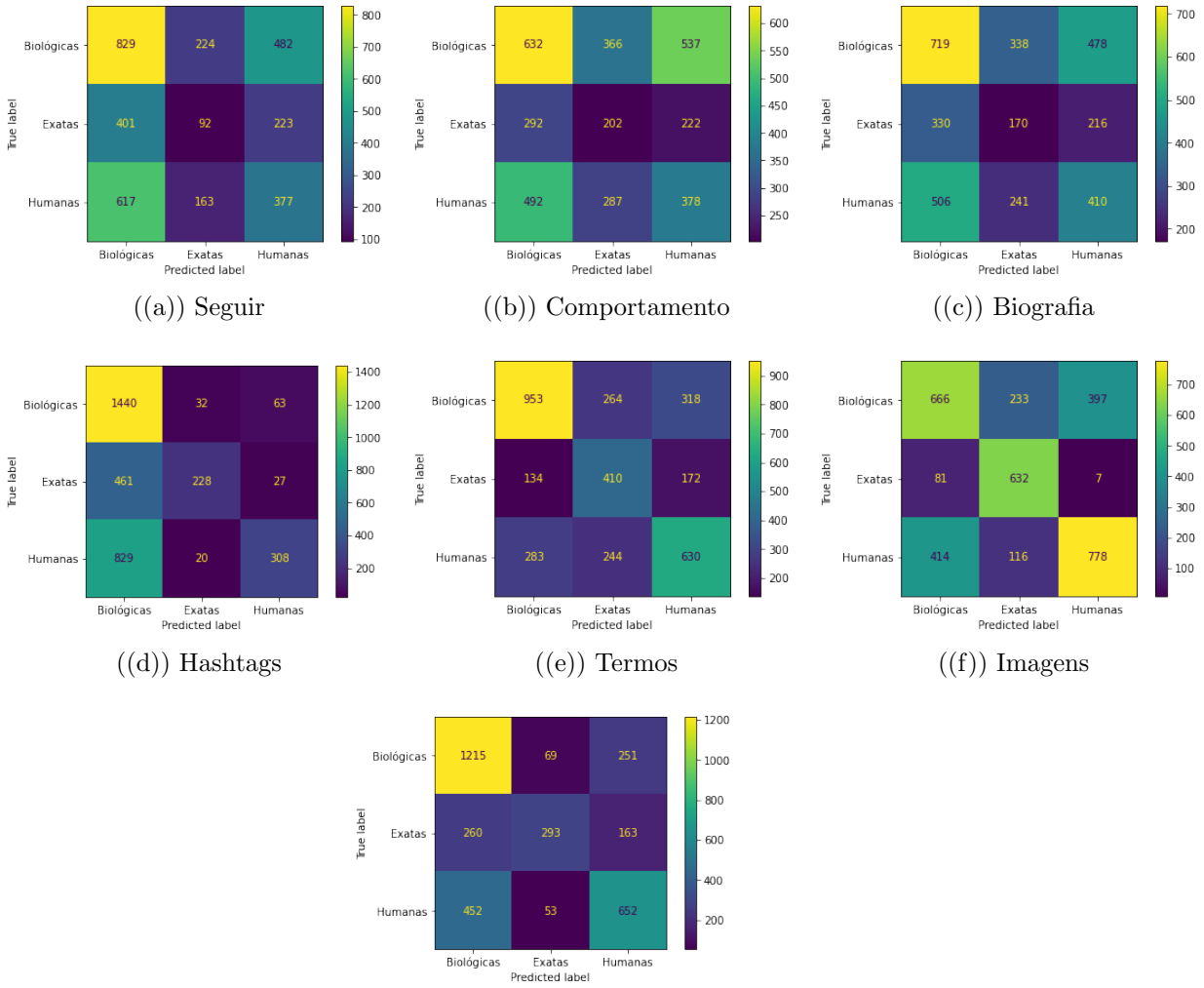
((g)) Todos

Figura 5.1: Matrizes de confusão para a classificação por área profissional usando SVM em cada grupo de atributos

tre si. Isso parece indicar uma classificação próxima do aleatório para esses atributos e um baixo poder discriminador deles.

A Figura 5.2(d) mostra que para *Random Forest* a mesma concentração de elementos na classe Biológicas observada com SVM continua. Este comportamento salta aos olhos, pois foram selecionadas *hashtags* em igual quantidade para as três classes. Uma possível causa disso é que as *hashtags* selecionadas apareçam em maior quantidade em Exatas, gerando o enviesamento do modelo.

As matrizes de confusão para Termos, Imagens e o uso de todos os atributos, presentes nas Figuras 5.2(e), 5.2(f) e 5.2(g), respectivamente, possuem grande quantidade de valores na diagonal principal, mostrando maior potencial para a detecção da área profissional nesses atributos.



((a)) Seguir ((b)) Comportamento ((c)) Biografia

((d)) Hashtags ((e)) Termos ((f)) Imagens

((g)) Todos

Figura 5.2: Matrizes de confusão para a classificação por área profissional usando RF em cada grupo de atributos

Os resultados para a medida acurácia para SVM e *Random Forest* estão na Tabela 5.2. A melhor performance nesta medida usando o método SVM está na utilização dos atributos de Imagens, com 61,04%. Para o método *Random Forest*, o melhor resultado é alcançado na combinação de todos os atributos: 63,38%. Assim, ambos os modelos alcançam melhores resultados similares.

A combinação de todos os atributos e as *Hashtags* são o segundo e terceiro colocados por acurácia para SVM. Para *Random Forest*, Imagens e Termos estão no segundo e terceiro lugar. Os termos e *hashtags* selecionados da biografia parecem ser distintos o bastante para discriminar entre as áreas profissionais, bem como as tendências nas publicações de imagens que foram notadas no Capítulo 4.

Os grupos de atributos Seguir, Biografia e Comportamento, apesar de algumas di-

	Random Forest	**SVM**
Seguir	38,09	33,86
Biografia	38,12	39,85
Comportamento	35,56	35,53
Hashtags	57,98	56,69
Termos	58,48	55,08
Imagens	62,45	**61,04**
Todos	**63,38**	59,42

Tabela 5.2: Acurácia de classificação, em porcento, de acordo com as grupos de atributos para área profissional

ferenças notadas no Capítulo 4 que as classes possuem neles, não apresentaram boa acurácia em nenhum dos dois métodos. A acurácia alcançada está um pouco acima da esperada com uma classificação aleatória (ou seja, dividindo o número total de elementos pela quantidade de classes) com três classes (33,33%). Estes atributos, assim, mostram-se de pouca utilidade sozinhos, contudo podem afetar positivamente no desenpenho alcançado pelos outros grupos.

Para SVM, a combinação de todos os grupos consegue uma acurácia mais baixa, comparando com a do *Random Forest*. Isso indica que o baixo valor obtido nesse cenário com SVM pode ser devido à influência das variáveis de baixo desempenhooo.

Na Tabela 5.3 estão os resultados da classificação de cada grupo de atributos para a Medida F1. Tanto para *Random Forest* quanto para SVM, combinar todos os atributos leva ao melhor resultado. O grupo de atributos que mais se sobressai é o de Imagens, para ambos os métodos. A superioridade de performance dos atributos extraídos das descrições automáticas das imagens sobre os termos usados pelos usuários nas biografias reflete o descrito em Song et al. (2018), que ressalta como as *tags* extraídas automaticamente de imagens do Instagram resultam numa performance na classificação dos perfis superior à feita com o uso de termos produzidos pelos próprios usuários.

Após isso, a classificação foi feita utilizando certas combinações dos grupos de atributos, a fim de determinar a influência de cada um deles na determinação da área profissional. Os atributos de palavras foram combinados com os demais. A Figura 5.3 contém as matrizes de confusão para a classificação feita com SVM. As Figuras 5.3(a), 5.3(b) e 5.3(c) mostram que para as combinações com os grupos Seguir, Comportamento e Biografia existem concentração nas predições para a classe Biológicas, em detrimento das outras: em Seguir, 1.328 perfis; em Comportamento, 1.376 perfis; e para Biografia, 1.366 perfis. Estes modelos, assim, acabam por ser enviesados para alguma das classes e pode-se concluir que essas combinações possuem baixo poder

	Random Forest	**SVM**
Seguir	36,75	31,89
Biografia	38,15	40,11
Comportamento	35,88	33,42
Hashtags	53,48	51,91
Termos	58,88	55,67
Imagens	61,82	58,94
Todos	**62,41**	**59,64**

Tabela 5.3: Medida F1 da classificação, em porcento, de acordo com as grupos de atributos para área profissional

discriminativo.

Para RF, conforme a Figura 5.4, existe um aumento na quantidade de predições corretas para todos os modelos, contudo esse aumento é apenas na ordem de dezenas, apontando para um desempenho similar entre os dois métodos nas combinações dos grupos de atributos.

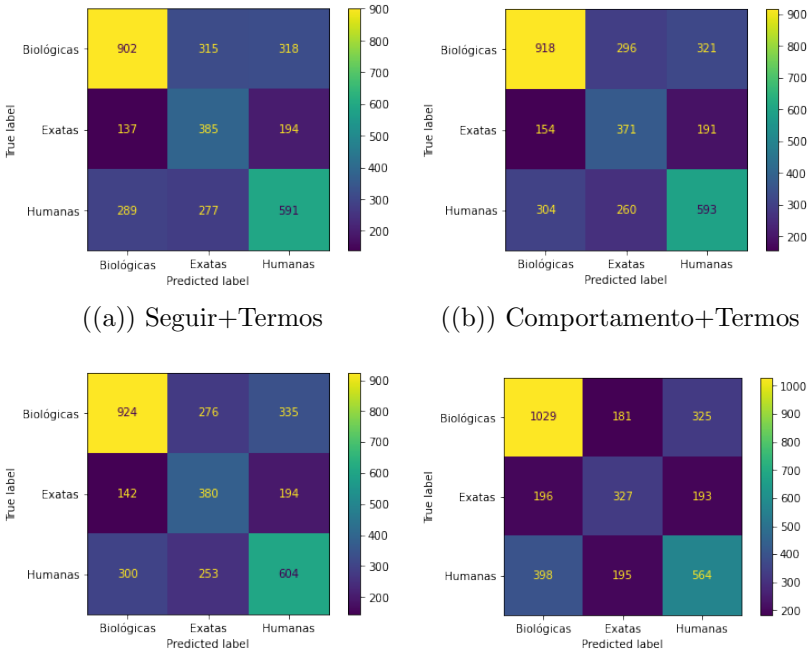
Os resultados para o atributo acurácia para a combinação dos grupos de atributos estão na Tabela 5.4. Nota-se que, apesar da combinação com os termos melhorar todos os resultados, para SVM os valores não ficam muito acima de 55%. Os melhores resultados para *Random Forest* estão na combinação dos Termos com o grupo Biografia, chegando a 61,80%. Isso mostra que as pequenas diferenças nos padrões de seguir expostas na Seção 4.3, ainda que de pouco poder discriminativo por si só, podem ajudar a aumentar o poder discriminativo dos Termos selecionados da Biografia. As outras combinações, apesar de terem acurácias menores, ficaram a apenas cerca de 2% de distância do melhor resultado.

A Tabela 5.5 contém os resultados das combinações de atributos para o atributo Medida F1. Em ambos os métodos, o melhor resultado foi obtido combinando os Termos com os atributos de Biografia.

5.3 Classificação de perfis por faixa etária

Os perfis foram agrupados por faixa etária (adolescentes e adultos) e a classificação foi realizada do mesmo modo que descrito na seção 5.2. As matrizes de confusão para a classificação usando cada grupo de atributos estão nas Figuras 5.5 para SVM e na Figura 5.6 para *Random Forest*.

Para SVM, a classificação utilizando *hashtags* acabou tendo uma concentração muito grande em adolescentes, conforme pode ser visto na matriz de confusão dessa classe,



((a)) Seguir+Termos ((b)) Comportamento+Termos

((c)) Biografia+Termos ((d)) Hashtags+Termos

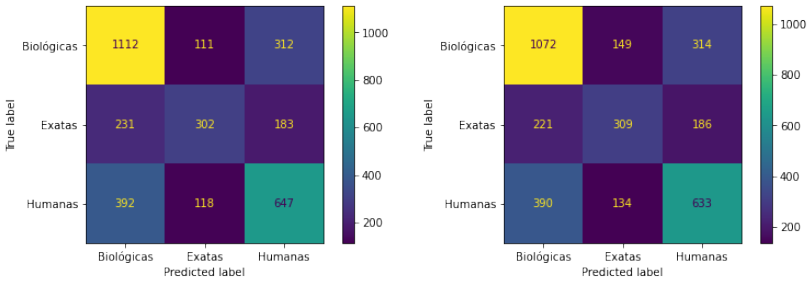
Figura 5.3: Matrizes de confusão para a classificação por área profissional usando SVM em combinações de grupos de atributos

na Figura 5.5(d), em que 2.091 registros são atribuídos à essa classe pelo classificador, contra 408 atribuídos à classe adultos. Este comportamento foi inesperado, pois o treino foi feito com balanceamento entre as classes e há uma grande diferença nas *hashtags* selecionadas para cada uma delas. Uma possível causa disso é que adultos utilizem menos *hashtags* em suas biografias, assim o modelo associa a presença delas à classe adolescentes.

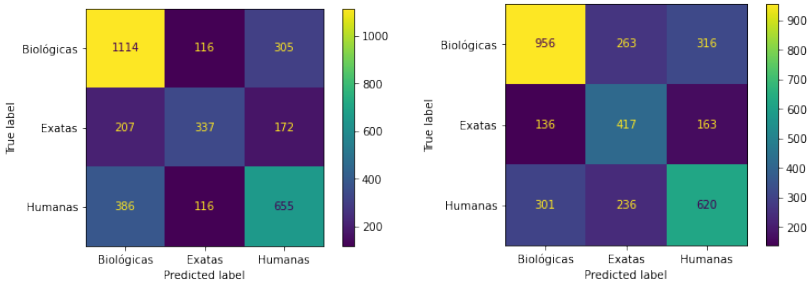
Uma concentração de predições na classe adolescentes também foi observada na Figura 5.5(b), totalizando 1.655 elementos preditos nessa classe. Aqui o número de perfis classificados como adolescentes, ainda que alto, foi um pouco menor do que

	Random Forest	**SVM**
Seguir + Termos	60,48	55,11
Biografia + Termos	**61,80**	55,99
Comportamento + Termos	59,10	55,22
Hashtags + Termos	58,48	**56,34**

Tabela 5.4: Acurácia de classificação, em porcento, de acordo com as combinações de grupos de atributos para área profissional



((a)) Seguir+Termos ((b)) Comportamento+Termos



((c)) Biografia+Termos ((d)) Hashtags+Termos

Figura 5.4: Matrizes de confusão para a classificação por área profissional usando RF em combinações de grupos de atributos

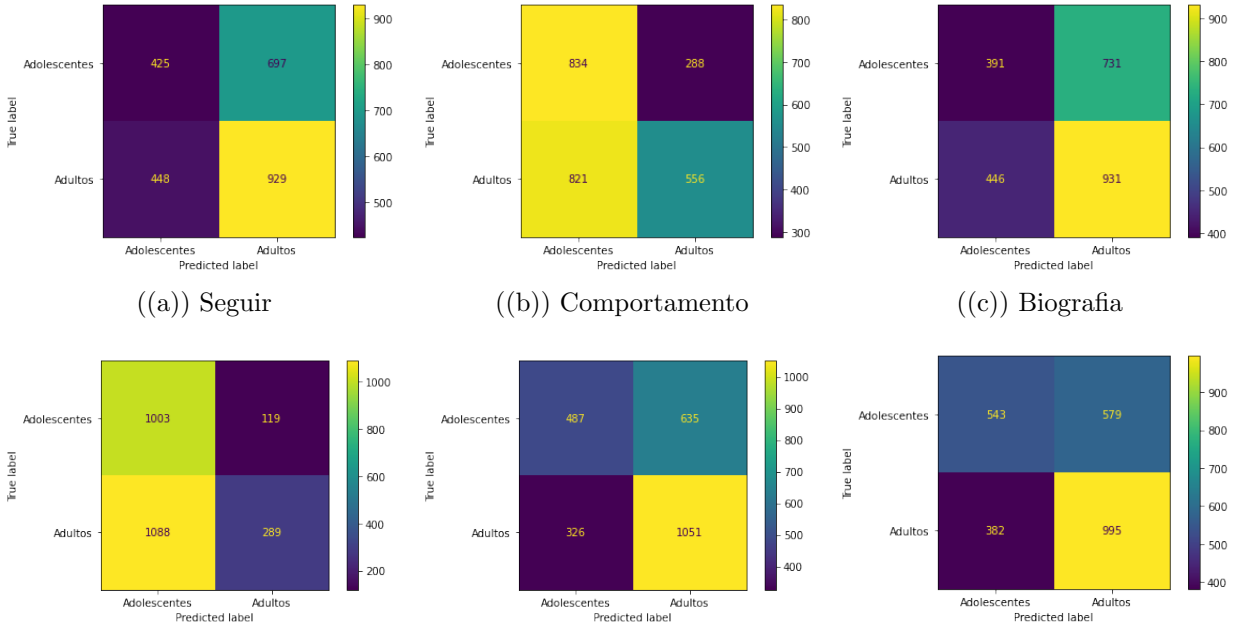
com o grupo de atributos anterior. Isso se deve provavelmente à percentagem maior de contas de negócio entre adolescentes, que passa a tornar o atributo característico dessa classe para o modelo.

As Figuras 5.5(c) e 5.5(e), por sua vez, mostram matrizes de confusão para os grupos Biografia e Termos em que os valores de predições para adultos são significativamente mais altos que para adolescentes: 1.662 e 1.686.

As Figuras 5.5(a) e 5.5(f) referentes às matrizes de confusão dos grupos Seguir e à combinação de todas as classes, mostram um cenário similar entre si, com uma concentração um pouco menor de previsões para adultos: 1.626 perfis categoriza-

	Random Forest	**SVM**
Seguir + Termos	0,5992	0,5569
Biografia + Termos	**0,6140**	**0,5643**
Comportamento + Termos	0,5870	0,5570
Hashtags + Termos	0,5881	0,5614

Tabela 5.5: Medida F1 da classificação, de acordo com as combinações de grupos de atributos para área profissional



((a)) Seguir ((b)) Comportamento ((c)) Biografia

((d)) Hashtags ((e)) Termos ((f)) Todos

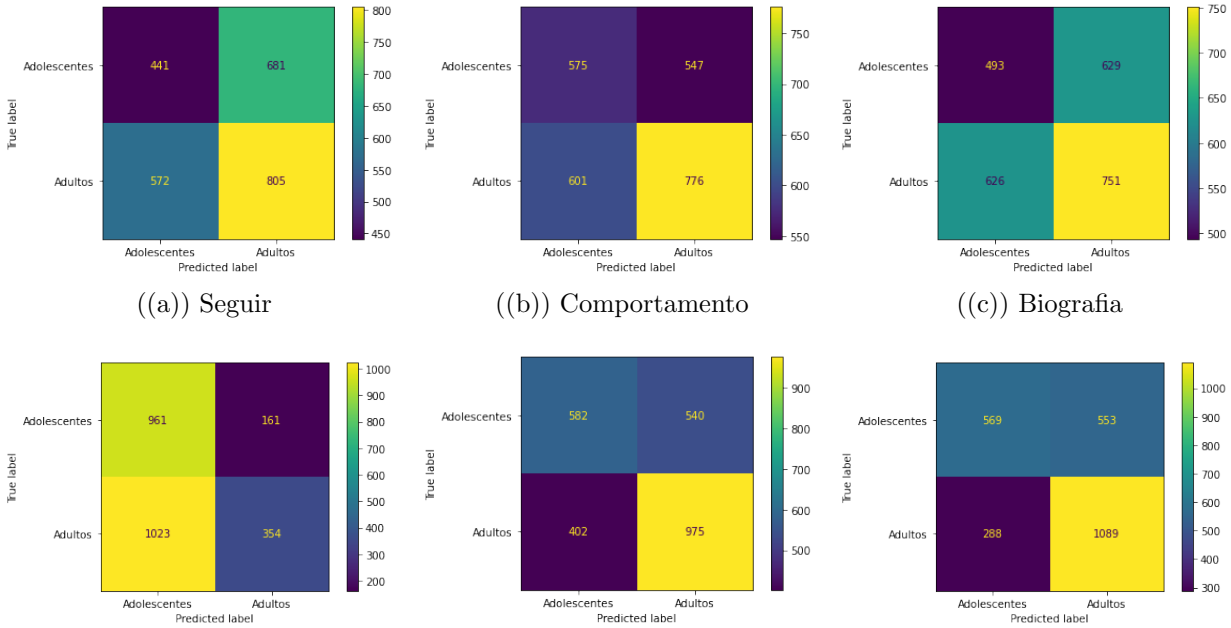
Figura 5.5: Matrizes de confusão para a classificação por faixa etária usando SVM em cada grupo de atributos

dos como adultos no atributo Seguir, enquanto o uso de todos os atributos leva a classificação de 1.574 perfis nessa classe.

A Figura 5.5(e), correspondente ao grupo Termos, respectivamente, exibem maiores valores na diagonal, ou seja, mais predições corretas, com 1.538 perfis. Isso aponta que este grupo possui alto poder discriminativo de faixa etária, devido à diferença nos termos usados nas biografias de cada uma delas. Assim, esta é a matriz de confusão o mais próxima do ideal.

Para *Random Forest*, a Figura 5.9(b) mostra matrizes de confusão com valores mais balanceados, com exceção dos resultados utilizando *hashtags*, que novamente se concentraram em adolescentes: uma quantidade ainda maior de predições estavam nesta classe, 1.984 ao todo. Isso confirma a baixa capacidade de *hashtags* presentes na biografia serem um meio de discriminar faixa etária.

A classe de atributos Termos (Figura 5.6(e)) e a combinação de todos os atributos (Figura 5.6(f)) acabaram por resultar em matrizes de confusão com muitos casos na diagonal (1.557 e 1.658, respectivamente). Isso não apenas confirma a utilidade dos Termos para classificação de faixa etária, como também aponta para uma melhora na classificação usando todos os atributos em relação ao SVM. Isso indica que a queda de desempenho ao combinar todos os atributos usando SVM provavelmente vem de falta de ajuste nos parâmetros ou poderia ser resolvido usando outra função de *kernel*.



((a)) Seguir ((b)) Comportamento ((c)) Biografia

((d)) Hashtags ((e)) Termos ((f)) Todos

Figura 5.6: Matrizes de confusão para a classificação por faixa etária usando RF em cada grupo de atributos

Já a Figura 5.6(a), a Figura 5.6(b) e a Figura 5.6(c), correspondentes aos resultados usando os atributos de Seguir, Comportameto e Biografia, respectivamente, mostram nas células das matrizes valores similares, oscilando por volta de 500-700 perfis. Isso aponta para uma classificação basicamente aleatória, seguindo a mesma tendência da classificação usando SVM.

Os resultados para os grupos de atributos usando a medida acurácia estão na Tabela 5.6. O grupo Termos consegue o melhor desempenho para SVM, com 61,54%, empatando com a combinação de todos os atributos, enquanto é o segundo melhor resultado em RF. Isso corrobora a hipótese que cada faixa etária utiliza um vocabulário diferente.

Para SVM, todos os outros modelos, incluindo a combinação de todos os atributos, ficam com uma acurácia em torno de 50%. Como a classificação é binária, esta acurácia está próxima da que seria alcançada com uma classificação aleatória dos dados. Assim, apenas o uso dos Termos fornece para este método um desempenho promissor.

A classificação com *Random Forest* teve como melhor resultado 66,35% de acurácia, com todos os atributos alimentando o modelo. Logo após vieram os grupos de atributos Termos, com 62,30% e Comportamento, com 54,06%. Em ambos os métodos, assim, os mesmos atributos se sobressaem: os termos mais relevantes das biografias, o identificador de conta de negócio e o número de posts.

	Random Forest	**SVM**
Seguir	49,86	54,18
Biografia	49,78	52,90
Comportamento	54,06	55,62
Hashtags	52,62	51,70
Termos	62,30	**61,54**
Todos	**66,35**	**61,54**

Tabela 5.6: Acurácia de classificação, em porcento, de acordo com os grupos de atributos para faixa etária

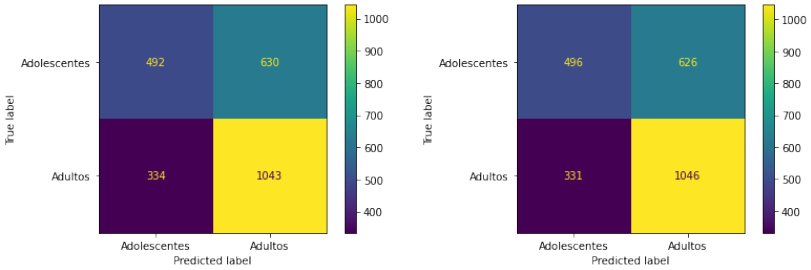
	Random Forest	**SVM**
Seguir	0,4953	0,5322
Biografia	0,4977	0,5268
Comportamento	0,5414	0,5456
Hashtags	0,4840	0,4587
Termos	0,6154	0,6041
Todos	**0,6557**	**0,6098**

Tabela 5.7: Medida F1 de classificação, de acordo com os grupos de atributos para faixa etária

Na Tabela 5.7 estão os resultados para os grupos de atributos individuais com a Medida F1. Os termos continuam sendo o grupo de atributo com melhor performance para SVM, bem como a combinação de todos os atributos para *Random Forest*, e em ambos os métodos o grupo Comportamento se sobressai. As *hashtags* nesse atributo estão na penúltima colocação, devido à concentração excessiva de elementos numa única classe que a Medida F1 captura.

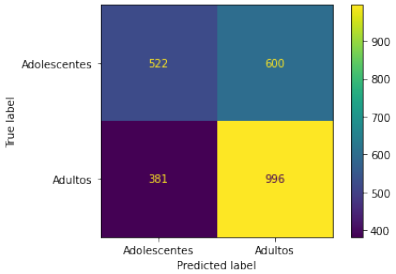
A classificação também foi feita utilizando combinando o grupo de atributos Termos com os demais para faixa etária. A Figura 5.7 contém as matrizes de confusão para a classificação feita com SVM. A combinação com Seguir (Figura 5.7(a)) levou a uma concentração das predições em adultos, com 1.594 elementos; para Comportamento (Figura 5.7(c)) isso também aconteceu, com 1.672.

Para a combinação com Biografia (Figura 5.7(b)), houve pouca mudança do uso do grupo sozinho, com uma concentração considerável de elementos classificados como adultos, totalizando 1.596. Na combinação dos termos com as *hashtags*, cuja matriz de confusão está na Figura 5.7(d) o número de casos corretamente classificados subiu para 1.554 e houve um maior balanceamento nas predições entre as duas classes,

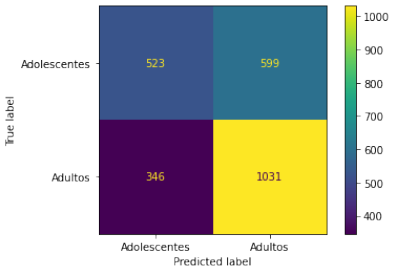


((a)) Seguir+Termos

((b)) Comportamento+Termos



((c)) Biografia+Termos



((d)) Hashtags+Termos

Figura 5.7: Matrizes de confusão para a classificação por faixa etária usando SVM em combinações de grupos de atributos

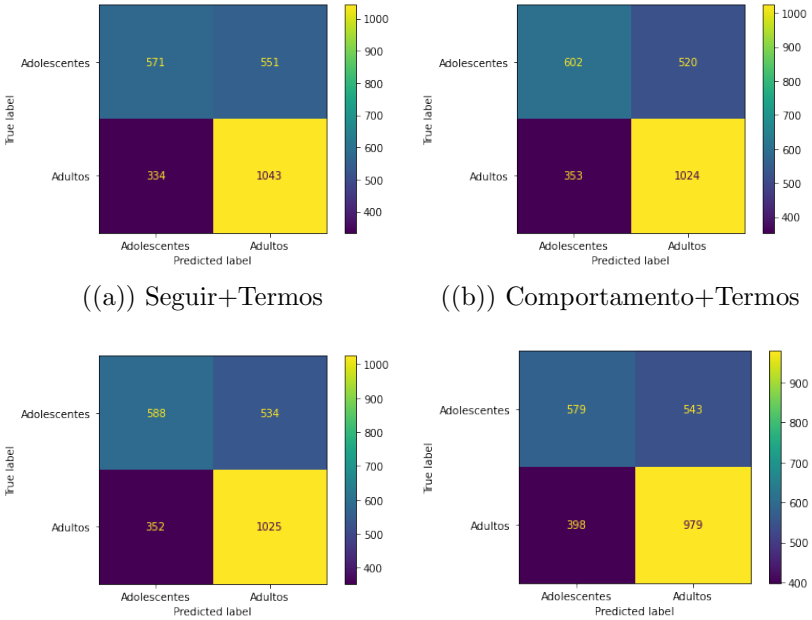
exibindo uma melhora quando comparado com o uso isolado das *hashtags*.

Para RF, conforme a Figura 5.8, todas as combinações levam a resultados similares entre si, com uma concentração de previsões na diagonal. Os resultados estão bem balanceados entre as duas classes, e assim como no SVM, houve uma melhora no grupo *Hashtag*, com 1.558 instâncias classificadas corretamente.

Os resultados para as combinações de atributos usando a medida acurácia estão na Tabela 5.8. O melhor resultado usando SVM é obtido combinando as *hashtags* e as palavras: 62,18%. Isso é levemente superior ao melhor resultado obtido com o uso isolado dos grupos de atributos. Estes obtém uma melhora leve, que não muda muito sua utilidade na classificação de faixa etária, com a combinação com o grupo Comportamento tendo o segundo melhor desempenho. Isso parece indicar que o baixo desempenho ao utilizar todos os atributos com esse método não é oriundo das *hashtags*, apesar de isoladamente elas irem mal.

Para *Random Forest*, a combinação dos grupos Comportamento e Termos supera o uso das demais individualmente, chegando em 65,07%, mas ainda é inferior à melhor classificação com a combinação de todas as classes. Todos os grupos obteram uma melhora significativa na combinação com os termos e ficaram num patamar similar.

Por fim, a Tabela 5.9 contém os resultados das combinações do grupo de atributos Termos com os outros grupos, com a Medida F1 sendo usada na classificação. Para *Random Forest*, o melhor resultado envolveu a combinação com o grupo de atributos



((a)) Seguir+Termos

((b)) Comportamento+Termos

((c)) Biografia+Termos

((d)) Hashtags+Termos

Figura 5.8: Matrizes de confusão para a classificação por faixa etária usando RF em combinações de grupos de atributos

de Comportamento, enquanto para SVM a combinação com as *hashtags* alcançou o topo.

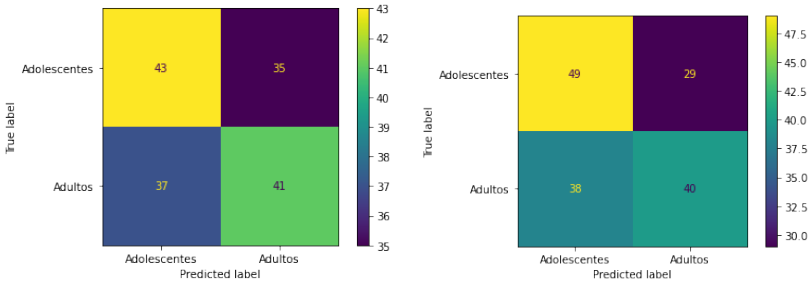
Estes resultados mostram que adolescentes e adultos se distinguem na quantidade de publicações que fazem e no uso da função de conta de negócio, visto o alto desempenho na classificação usando esses atributos. Existia a expectativa que adultos apresentariam marcadores de escrita mais complexa do que adolescentes em suas biografias e assim esse tipo de atributo teria bom potencial discriminativo. Aparentemente os valores altos das correlações ponto-bisseriais nos atributos de biografia davam suporte à isso. Contudo, diante do baixo desempenho deles em comparação com outros grupos de atributos, essa hipótese não se confirmou.

	Random Forest	**SVM**
Seguir + Termos	64,59	61,42
Biografia + Termos	64,55	60,74
Comportamento + Termos	**65,07**	61,70
Hashtags + Termos	62,34	**62,18**

Tabela 5.8: Acurácia de classificação, em porcento, de acordo com as combinações dos grupos de atributos para faixa etária

	Random Forest	**SVM**
Seguir + Termos	0,6398	0,6037
Biografia + Termos	0,6408	0,6007
Comportamento + Termos	**0,6466**	0,6066
Hashtags + Termos	0,6199	**0,6137**

Tabela 5.9: Medida F1 de classificação, em porcento, de acordo com as combinações dos grupos de atributos para faixa etária



((a)) SVM ((b)) RF

Figura 5.9: Matrizes de confusão para a classificação por faixa etária usando os atributos de imagem para SVM e RF

Como os atributos de imagem para faixa etária foram coletados usando um subconjunto do conjunto de dados, sua classificação e avaliação foi feita separadamente. Para garantir o balanceamento na classificação, foram utilizados 311 elementos para cada faixa etária.

As matrizes de confusão estão na Figura 5.9. Para ambos os métodos, apesar da maioria dos elementos estarem na diagonal principal (84 elementos para SVM e 89 para RF), todas as células apresentam valores similares. Os resultados da classificação estão na Tabela 5.10. Os melhores resultados foram obtidos com RF para as duas métricas, com a acurácia apontando para um resultado apenas um pouco melhor de uma classificação aleatória.

	Random Forest	**SVM**
Acurácia	**53,78**	56,54
Medida F1	**0,5360**	0,5527

Tabela 5.10: Acurácia, em porcento, e Medida F1 de classificação, para RF e SVM, usando os atributos de imagem para faixa etária

Capítulo 6

Considerações Finais

Este trabalho definiu como objetivo determinar a faixa etária e a área profissional de usuários do Instagram, reunindo perfis desta rede social falantes de língua portuguesa e verificando quais atributos possuem maior poder discriminativo dessas classes. Assim, foram construídos dois conjuntos de dados a partir de perfis de Instagram, um rotulado com faixa etária e outro com área profissional. A partir desses conjuntos de dados, foi realizada a classificação nesses dois quesitos, usando uma série de atributos selecionados conforme trabalhos anteriores com os métodos *Random Forest* e *Support Vector Machines*. Ambos os conjuntos de dados foram construídos de modo a conter apenas perfis de falantes de língua portuguesa. A classificação foi realizada separadamente com diferentes grupos de atributos de modo a comparar o poder discriminativo deles.

Na análise dos conjuntos de dados encontrados, a comparação dos atributos entre as classes revelou diferenças significativas apenas entre algumas delas. Para área profissional, o número de emojis e pontuação utilizados nas biografias, quantidade de contas seguidas e seguidores e taxa de perfis de negócio apresenta diferença entre certas classes; enquanto para faixa etária a quantidade de seguidores e a quantidade de posts se revelaram como discriminativos. Os termos e as *hashtags* selecionadas como mais relevantes para cada classe também exibiram particularidades entre si.

Os termos e as *hashtags* selecionados a partir da biografia dos perfis permitem fazer certas inferências sobre os conjuntos de dados. Nas ciências exatas, termos voltados à fornecimento de serviços aparecem com frequência, apontando que profissões dessa classe tendem a utilizar o Instagram para anunciar seus serviços. Outra evidência à favor disso é a alta quantidade de contas marcadas como de negócio nessa classe. Isso vai ao encontro de uma dos pontos relevantes desse trabalho, que é o uso do Instagram para alcançar públicos-alvo para produtos e serviços.

Já em todo o conjunto de dados rotulado por faixa etária há uma grande quantidade de termos e *hashtags* femininos. Isso indica uma maior propensão de mulheres realizarem publicações em que revelam sua idade, num indicador possível de sexo a

investigar. Contudo, essa situação também indica que o conjunto de dados obtido possui certo viés, prejudicando a capacidade de generalizar as conclusões obtidas para usuários masculinos.

Os resultados na classificação para área profissional mostraram uma superioridade dos termos mais relevantes extraídos das biografias e do texto alternativo das imagens sobre os outros atributos. A utilização de todos os atributos conseguiu a melhor performance utilizando ambos os métodos, chegando ao máximo de 63,38% de acurácia com *Random Forest*. Este valor é aproximadamente o dobro do que seria obtido numa classificação aleatória (33,33%).

A classificação de faixa etária revelou que, novamente, os termos extraídos da biografia são os atributos de maior poder discriminativo. O número de posts e o identificador de conta de negócio se revelaram atributos promissores na detecção de faixa etária, conforme previsto em parte da literatura. Combinar todos os atributos conseguiu o melhor desempenho com o método *Random Forest*, com 66,35% de acurácia. Ainda que superior ao desempenho de uma classificação totalmente aleatória, este valor é inferior ao obtido em trabalhos feitos no Instagram com perfis falantes de língua inglesa, que conseguem alcançar acurácia de mais de 89% (Han et al., 2016).

O trabalho conseguiu estabelecer um conjunto de dados composto de perfis do Instagram, como se pretendia. A coleta desses dados tomou uma parte significativa do tempo, sendo usada uma técnica nova para isso (Seção 3.1), a partir de uma adaptação de técnicas apresentadas na literatura: a pesquisa de publicações feitas por usuários a partir de *hashtags* relacionadas com a classe sendo coletada, com o objetivo de aumentar a quantidade de perfis de fato pertencentes à classe. Ptetende-se investigar formas de disponibilizar esses dads para pesquisa futura, sem infringir os termos de uso definidos pelo Instagram.

Os resultados obtidos apontam para o potencial de diferenciar perfis no Instagram quanto à área profissional. As maiores contribuições, contudo, estão no processo de coleta de dados, através da apresentação dos passos e das ferramentas usados na obtenção dos perfis e os modos de verificar se eles pertenciam às classes desejadas; bem como na análise dos dados, detectando tendências nos atributos em cada uma das classes.

6.1 Limitações encontradas

Apesar do método de coleta de dados ter como objetivo maximizar a quantidade de perfis dentro das classes desejadas, perfis sem relação com as classes usando as *hashtags* em suas publicações, muitas vezes com fins de publicidade, foram obtidos. Um processo de inspeção manual do conjunto de dados teve de ser realizado para remover estes perfis.

Outro problema foi o encerramento da ferramenta utilizada na coleta dos dados devido a um pedido de direitos autorais do Instagram, levando a uma interrupção

do processo de coleta de dados. A isso se juntaram os diversos banimentos de contas e de IP realizados pela plataforma, obrigando a reiniciar o processo. Disso tudo se depreende que o processo de coleta de dados, incluindo os passos de rotulação e limpeza, é o maior gargalo num trabalho como este, merecendo maior atenção.

Por causa desses problemas, não foi possível obter os atributos de imagem para todos os perfis no conjunto de dados rotulado por faixa etária. Assim, foi realizada a classificação desse atributo num subconjunto dos perfis coletados.

6.2 Trabalhos Futuros

Técnicas diferentes de coleta de dados podem ser utilizadas para obter perfis do Instagram, comparando a qualidade dos rótulos e da classificação realizada com os deste trabalho. A utilização de rotuladores humanos e *crowdsourcing* pode ser considerada para isso.

Métodos de coleta de dados que sejam mais efetivos em localizar perfis do Instagram dentro das classes devem ser encontrados, a fim de conseguir mais dados. Com um conjunto de dados maior, poderá ser feita uma classificação de maior qualidade, que esteja num desempenho comparável com os trabalhos nas mesmas classes em língua inglesa.

Quanto aos atributos visuais, a qualidade da classificação utilizando os dados vindos do texto alternativo fornecido pelo Instagram pode ser comparada com a abordagem de utilizar modelos de aprendizagem de máquina para extrair tópicos das imagens. Outras características dos perfis, como sexo e localização geográfica, podem ser exploradas no futuro, bem como outros atributos para descrevê-los.

Apêndice A

Hashtags selecionadas para área profissional

A Tabela A.1 contém as *hashtags* selecionadas a partir do conjunto de dados rotulado por área profissional.

Humanas	**Exatas**	**Biológicas**
3d	3d	acidentesdomésticos
administração	ambientepequenopor	autocuidado
alfabetizarcomamor	amoesporte	bióloga
alynepsicopedagogajundiai	amominhafamília	bodybuild
aprenderaqui	apartamentopequeno	bolsonaristas
artesa	apecompacto	bscellstore
atividadespedagogicas	arqgram	capricorniano
autismo	arquiteto	ciênciaesaúde
brinquedoseducativos	arquitetura	concurseira
cacheada	artista	corengoiasdúvida
cibelepsicopedagoga	associação	cruzvermelha
compartilhandoideias	atletacristã	cruzvermelhaportuguesa
compartilhar	auditoria	cursosonline
concurseira	b3	desenho
conexãopsicológica	bim	deusacimadetudo
crista	bioconstructoresideias urbanas	dicas
cruzeirodosul	brasilquando	direitoasaude
desenvolvimentoinfantil	claudiabaldassoarquitetura	direitoesaude
dicaseducacionais	construcao	direitomedico

divulgaçãocientíficanovidades	construção	diskbrejacariri
donadecasa	construcaocivil	doadordeorgaos
eadpreparamos	construçãocivil	ebooksacesse
educacao	consultora	educacaofisica
educação	consultoriacontábil	emagrecimento
educacaoinfantil	contabeis	emergencia
educaçãoinfantil	contabilidade	emergencista
educaçãoinfantileterna	contabilidadepublica	emergency
educadorasince	contabilidadetributaria	empoderamentoda enfermagem
educandopequenos	contadora	enf360
emocional	contandopracontador	enfermagem
ensinofundamental	crcce	enfermagemalta performance
espiritual	csc	enfermagemporamor
familia	curiosa	enfermeira
financeira	departamentopessoal	enfermeiro
fiqueemcasa	designer	enfermeiros
ginásticaparaocerebro	dicascontabeisufsjpassa tempo:costura	escolhisersocorrista
graduação	dicasdobarsaosbarsaarquitetos	estomas
ideias	direito	estudaqueavidamuda
inclusao	elainebeloni	exerciciofisico
inclusãoescolar	empreenderreierainhamcz	feridas
institutomaenatureza	eng	feridasecurativos
libras	engenharia	ficaemcasa
lúdicas	engenhariacivil	flamenguista
ludicidade	engenhariacivilbr	forabolsonaro
mae	engenhariacivilplanilha	fotografia
mãe	engenhariacompropositobetelmoveissobmedidacursos	gestores
maedecasal	engenheiraflaviaarruda	hipnose
mamae	engenheiro	hobby
materiaisadaptados	engenheirocivil	icu
métodosupera	engenheiros	ilustracoes
mimosvanessab	engeplificando	libras

minasgerais
modelodenver
muitoalémdoensino
03125661150
natalrn
neurociência
neuroeducação
neuropsicopedagogia
obamamaeensina
oficinas
palestrante
papelaria
pedagoga
pedagogaesp
pedagogia
pedagogiacomamor
pedagogiacomamorgosto
pedagogiadadepressão
pedagogiaefetiva
pedagogiainfantil
pedagogiaporamor
pedagógico
professora
professoracoruja
professorapedagogialetras
professoras
professoraserva
profleandrovieira
projetoseducacionais
psicologiainfantil
psicopedagoga
psicopedagogaensinando
psicopedagogariogrande
psicopedagogia
psicopedagogoinstitucional
estagiaria
estruturas
estudantedecontabeis
execucao
finanças
futuracontadora
glutenfree
graduaçãoead
igrejas
imóvel
interiores
isacurado
itanhomi
itsmylife
kopkefire
lexcontabilidadevem
lightsteelframe
minhasandançaso
móveisplanejados
negocios
obra
obracivil
obras
ongs
pedraescondida
pequenosespaços
pericia
periciatecnica
peritos
permaculturagyn
pro
professor
professora
projetos
libriana
lideranças
lifestyle
livredevasinhosdúvidas
loucosporobstetriciaeneo
lovedogsbh
maededog
maedemenina
mamae
marilupaivanutri
ministeriodasaude
mulher
musculação
neurociência
notivago
nurse
nutrimarilupaiva
obstetricia
papelaria
parceria
parcerias
peim
personaltrainer
photo
pl2564
primeirossocorros
profissionaisdesaude
propaganda
psicologa
psicologia
psm
qualidade
queimadura
rn

psicoteca	projetosautorais	rotina
roboticaeducacional	projetosonline	saude
rotinaescolar	prouni	saúde
saladerecursos	reforma	sofiasgondimenfermeira
servadedeus	semlactose	studygram
soberanaimponente	simonevieccelliarquiteturaarq	studygramsprotina
sobremim	sintaforteoquetemove	sus
sousagrado	solobrasil	terapia
stem	study	terapialarval
tentandoserfitnessos	técnicoemedificações	tradiçãoecompetência
terapiaaba	unidoctumyoutube	transiçãocapilar
trabalho	universidades	uci
vegetariana	urbanista	userainhaateliê
vilacatavento	vemcomigoamplaengenharia	uti
vng	winelover	vestibulanda

Tabela A.1: *Hashtags* mais comuns selecionadas para cada área profissional.

Apêndice B

Hashtags selecionadas para faixa etária

A Tabela B.1 contém as *hashtags* selecionadas a partir do conjunto de dados rotulado por faixa etária.

Adolescentes	**Adultos**
:	mãe
18primaverasfutura	13
act	25
adesivos	22primaveras
allyfwyllyam	23anos
amorporviolao	2out
amuhhvaquejada	30anos
apresentadoradetv	acapulco
artesanatoaprenda	acima
atendimentoaltopadrao	acrediteconfiepersevere
ator	administraçãodeempresas
atriz	adp
banners	amor
bemestar	artesanatoaprenda
bemestargaúcha	ascom
biscuitpersonalizado	bemvindos
blacklivesmatter	biscuitpersonalizado
blogueira	canal
bodypositive	capricorniano
bolsasdeestudo	carioca

cachiada
cachos
cantor
cartãovisita
cases
catoledorocha
cinema
consultoramarykay
corpolivregabiinaturavendas0nline
corredor
cristão
dancarinosdefunk
discotecagem
dj
djrickmenezes
douradoraçõesjusticeiro
drawing
explorar
explore
falabaixo
feedorganizado
feministabelieve
física
foryou
fotografia
geminiano
gibranssa
girlpower
humorista
imunizado
instablogger
intagram
intalove
juazeirodonortetéc
libriana
coreografo
cosmeticostatypaiva
crochetera
cursando
dancer
davi
de
deboista
deus
dicasdebeleza
dicasdecasa
do
donadecasa
educaçãofísica
empreendedora
encomendesuamascara
esmeralda
euamoenfermagemtécnica
familia
fé
filhadoreicristã
fitnessmotivationjamais
fluenstudent
foconoobjetivo
formada
goiania
goleiro
hairtransition
hopihari
influencer
influencerstyle
jadsoneduardo
jesus
kaeldamamãe
maedemenina

longevidade
madeirartdiy
makeupete
marketing
matemática
monstrotrinkado
music
musicas
nãobinário
nerd
oi
oliviatts
pai
palmeirasegrande
palmeiraséminhavida
panfletos
peaceloveandrocknroll
persistênciablogueirasolteira-olindarecifecarnavaaquarianaore
personalizados
pleasechalkboarding
poesia
pontepretano
pride
professor
qualidadedevida
reels
reelsprofissional
rosesanopaisagismo
rumo15k
saude
sertanejo
sertanejouniversitárioinscrevase
simplepocketparty
singer
storys
maedemeninose
mamaede3princesas
maquiagem
mariaclara
mascaradetecido
mesaposta
meumikael
meupacotinhodeamor
meuslivros
moda
mônica
oncoinfluencer
pedagogia
pepitafã
professordiretor
qgstepteam
receitas
redesdecomputadoreso
reeducaçãoalimentar
reels
rihannanavy
riobonito
riodejaneiro
rjabençoada
rotina
sendomilhodepipoca
sigojesuscristo
sonhadora
sp
taekwondo
tecenfermagem
tendojeustenhotudosou
tentandoserfitness
tentantes2020
transiçãocapilar

teamclauderomaoconsultoria:	tudo
teatro	uelinghtonpedagoga
tenhasonhosincriveis	umolharparahistorias
theater	unefparceriassigame
umboloporsemana	usemascara
uolou	venciocancer
vanessameirelesassessoria	venciolinfoma
vaquejadalegal	vidareal
vemcomuolou	videonovo
vestibular	virginiano

Tabela B.1: *Hashtags* mais comuns selecionadas para cada faixa etária.

Apêndice C

Termos selecionados para área profissional

Humanas	**Exatas**	**Biológicas**
17	10	10
23	20	21
aba	22	abaixo
acadêmica	48	acadêmica
acompanhamento	55	adolescente
adm	68	adolescentes
administração	3d	adulto
administrador	abaixo	adultos
administradora	acadêmica	agendamento
administrar	acesse	agendamentos
ajudar	acompanhamento	agende
ajudo	adm	agora
alfabetização	ajudo	ajudando
amo	ambiental	ajudar
amor	ambientes	ajudo
ano	amo	amamentação
anos	amor	amor
apaixonada	anos	anos
aprender	apaixonada	ansiedade
aprendizado	aqui	apaixonada
aprendizagem	área	aprender
aqui	arquiteta	aqui

área
arte
artes
assessoria
atendimento
atividades
auditoria
aula
aulas
autismo
avaliação
bacharel
be
bem
blog
brasil
brincar
casa
casada
casado
ciências
clinica
clínica
compartilhando
compartilhar
compliance
concurseira
conhecimento
constante
construção
consultor
consultoria
contato
conteúdo
conteúdos
arquiteto
arquitetônico
arquitetônicos
arquitetura
arte
artista
assessoria
através
autorais
bem
bim
brasil
casa
casada
chama
ciências
civil
clique
comerciais
comercial
comigo
compromisso
concreto
conforto
conhecimento
conosco
construção
construções
consultor
consultoria
consultorias
contábeis
contábil
contabilidade
contador
arte
atendimento
atendimentos
através
autoconhecimento
autocuidado
autoestima
avaliação
bacharel
bem
brasil
busca
clínica
clínico
clique
coach
cognitivo
cognitivocomportamental
comigo
compartilhando
comportamental
comportamento
conhecimento
consulta
consultoria
contato
conteúdo
conteúdos
corpo
cref
crianças
cristã
crp
cuidar
curso

controladoria
coordenadora
crianças
criativa
cursando
curso
desenvolvimento
deus
dia
dicas
dificuldades
digital
direct
direito
drive
duas
ead
ed
educacao
educação
educacional
educar
email
empreendedor
empreendedora
empresa
empresarial
empresas
ensinar
ensino
escola
escolar
esp
espaço
especial
contadora
contato
contatos
conteúdo
conteúdos
crea
curiosidades
curso
cursos
decoração
departamento
desde
design
designer
deus
dia
dicas
digital
direct
direito
edificações
elaboração
elétrico
email
empreendedor
empreendedora
empresa
empresas
eng
engenharia
engenheira
engenheiro
escritório
esp
especialista
cursos
desenvolvimento
deus
dia
dicas
direct
docente
domiciliar
ed
educação
emagrecimento
emergência
emocional
emoções
encontrar
enf
enfermagem
enfermeira
enfermeiro
enfermeiros
ensino
esp
especialista
esposa
estudante
estudantes
estudo
estudos
experiências
família
familiar
fazer
fé
feridas
fisica

especialista
estagiária
estudante
estudos
eterna
experiências
faço
família
fé
física
forma
formação
formada
freire
fundamental
futura
gestão
graduada
graduanda
humor
ideias
inclusiva
infância
infantil
informações
inspirações
institucional
intervenção
jogos
letramento
link
livros
lúdicas
mãe
mamãe
esposa
estruturais
estrutural
estruturas
estudante
estudos
execução
fale
finanças
financeira
financeiro
fiscal
forma
fundações
futura
futuro
geral
gerenciamento
gerente
gestão
graduanda
hidrossanitário
https
imóveis
incêndio
informações
inspirações
instalações
interiores
jesus
judicial
juntos
laudos
link
mãe
física
físico
fisiologia
forma
formação
funcional
futura
gestão
graduação
graduada
graduanda
graduando
https
idosos
individual
infantil
informações
leve
link
livros
mãe
marketing
melhor
mental
mente
mestre
motivação
mulher
mulheres
mundo
musculação
neuropsicologia
novo
obstetrícia
olhar

marketing
materiais
melhor
mestre
mg
motivação
mundo
municipal
negócio
negócios
neuropsicopedagoga
neuropsicopedagogia
nova
oficinas
olá
pais
palestrante
papelaria
parcerias
particular
paulo
pedagoga
pedagogapsicopedagoga
pedagogia
pedagógica
pedagógicas
pedagógico
pedagógicos
pedagogo
perfil
pesquisador
pessoal
pessoas
pode
pós
mba
mei
melhor
mercado
mg
mundo
negócio
obra
obras
online
orçamento
orçamentos
pai
palestrante
parcerias
patologia
paulo
perícia
pessoa
pessoal
planejamento
pós
pra
prática
prof
professor
profissional
projetar
projeto
projetos
qualidade
realidade
realizar
referência
reforma
on
online
orientação
pai
palestrante
parcerias
parto
perfil
personal
pessoal
pessoas
pode
pós
pósgraduanda
pra
presenciais
presencial
processo
prof
professor
professora
profissionais
profissional
psi
psicanalista
psicologa
psicóloga
psicologia
psicológica
psicólogo
psicoterapeuta
psicoterapia
qualidade
quer
relacionamentos

pósgraduanda
potencial
práticas
prof
profa
professor
professora
professores
profissional
projetos
psicóloga
psicologia
psicomotricidade
psicopedagoga
psicopedagogia
pública
qualidade
rede
reflexões
reforço
resumos
rj
rotina
sempre
ser
sobre
sp
stories
sugestões
toda
tudo
universidade
universitário
vendas
via

reformas
regularização
residenciais
residencial
rio
rs
sc
segurança
sempre
serviços
simples
sobre
sócio
soluções
sonho
sonhos
sp
stories
sucesso
téc
técnica
técnico
tecnologia
ter
toda
todo
todos
tornar
trabalho
transformo
tributária
tudo
urbanismo
urbanista
vamos

residência
residente
resumos
rj
rotina
saude
saúde
sempre
ser
serviços
sessão
sim
sobre
social
sp
stories
tcc
tec
téc
técnica
técnico
tempo
ter
terapeuta
terapia
todo
trabalho
trainer
tratamento
treinamento
treino
tudo
urgência
vamos
via

vida	via	vida
whatsapp	vida	vidas
youtube	whatsapp	whatsapp

Tabela C.1: Termos mais significativos selecionados para cada área profissional.

Apêndice D

Termos selecionados para faixa etária

Adolescentes	Adultos
11	18
15	22
16	23
16y	24
17	25
17y	26
18	78
18y	abençoada
19	acadêmica
19y	acima
2003	adm
21	administração
23	al
28	alegria
55	alma
abaixo	amante
acadêmica	amo
aceitamos	amor
acesse	and
adm	aninhos
administração	anos
ajudo	apaixonada

al	apostólica
alma	aquariana
além	aqui
amante	ariana
amo	assista
amor	assistam
and	ba
aninhos	bacharel
anos	bacharela
apaixonada	baiana
apenas	beleza
aquariana	bem
aqui	bom
arte	brasil
assista	cachos
assistam	caminho
atendimento	canal
atriz	canceriana
ba	carioca
be	casa
beleza	casada
bem	casal
bolos	católica
brasil	celular
cabelo	ceo
cada	civil
canal	ciências
carioca	clínica
cartões	comigo
casada	compartilhando
católica	consultora
ce	contato
cima	contábeis
clique	cristo
coisa	cristã

coisas	cursando
consultora	design
contato	designer
coração	deus
cristo	dia
cristã	dicas
desde	digital
designer	dinheiro
deus	direct
dia	direito
dicas	dona
digital	educação
direct	empreendedora
direito	enfermagem
doces	enfermeira
dona	engenheira
empreendedora	ensaio
encomendas	escorpiana
enfermagem	especialista
es	esposa
espalhe	estudante
estudante	estética
eventos	família
facebook	faz
família	fazer
faz	felicidade
fazer	feliz
feliz	filha
filha	fisioterapeuta
força	fisioterapia
futura	forte
futuro	futura
fé	fé
gestão	física
gratidão	geminiana
in	gestão

influencer
informações
is
jesus
joão
leonina
libriana
lifestyle
link
loja
luz
lá
make
makes
makeup
mamãe
maquiadora
maquiagem
marketing
medicina
medo
melhor
menina
mg
mim
moda
modelo
momentos
mulher
mundo
my
mãe
nada
novo
nunca
graduada
graduanda
grande
gratidão
heitor
in
infantil
influencer
jesus
joão
laura
leonina
libriana
life
link
loja
love
luz
lá
mae
maior
make
mamãe
maquiadora
maquiagem
maria
marketing
maternidade
medicina
melhor
mg
miguel
mim
moda
modelo

olhos
orçamento
orçamentos
pai
parceria
parcerias
paulo
pe
pedidos
perfil
personalizados
pessoa
pessoal
pessoas
pode
posso
pouco
pr
pra
primaveras
professora
profissional
proprietária
psicologia
quer
rio
rj
rn
sagitariana
segue
sempre
senhor
ser
shine
signo
sob
momentos
mulher
mundo
my
mãe
noiva
nordestina
nova
novo
nunca
nutrição
onde
pai
parcerias
paulo
pb
pe
pedagoga
pedagogia
perfil
pernambucana
pessoa
posso
pra
primaveras
princesa
professora
profissional
proprietária
príncipe
pós
quanto
quer
real
receitas
rei

sobre	rio
sol	rj
sonhos	salmos
sp	saudável
stories	saúde
storys	sempre
tec	senhor
tempo	ser
ter	sobrancelhas
the	sobre
tik	social
tiktok	sonhos
to	sp
toda	stories
todo	tempo
todos	the
tok	todo
trabalho	todos
tudo	trabalho
técnica	tudo
vai	técnica
vem	vai
veterinária	vem
vezes	versão
via	via
viagens	vida
vida	vindos
vivo	virginiana
voce	virtual
vídeo	vivendo
we	viver
whats	vivo
whatsapp	whatsapp
years	years
youtube	you
youtuber	youtube

Tabela D.1: Termos mais significativos selecionados para cada faixa etária.

Apêndice E

Termos selecionados nos textos alternativos das imagem no conjunto de dados rotulado por área profissional

Termo	Humanas	Biológicas	Exatas
jewelry	Sim	Sim	Sim
xray	Não	Sim	Não
bicycle	Sim	Sim	Não
burger	Sim	Sim	Não
car	Sim	Sim	Sim
sofa	Sim	Sim	Sim
furniture	Sim	Sim	Sim
msc	Não	Sim	Sim
money	Sim	Não	Não
more	Sim	Não	Não
activewear	Sim	Sim	Não
shoes	Sim	Sim	Não
cat	Sim	Sim	Sim
nature	Sim	Sim	Sim
ocean	Sim	Sim	Sim
suit	Sim	Não	Não
cup	Sim	Sim	Sim
room	Sim	Não	Não

eyeglasses	Sim	Sim	Sim
fragrance	Não	Sim	Não
food	Sim	Sim	Sim
book	Sim	Sim	Sim
french	Sim	Sim	Não
me	Não	Não	Sim
tree	Sim	Sim	Sim
lake	Sim	Sim	Sim
trees	Sim	Sim	Sim
dog	Sim	Sim	Sim
may	Sim	Não	Não
water	Sim	Sim	Sim
fries	Sim	Sim	Não
wall	Sim	Não	Não
fire	Sim	Sim	Não
footwear	Sim	Sim	Sim
civil	Não	Não	Sim
swimming	Sim	Não	Não
dr	Sim	Sim	Não
sunglasses	Sim	Sim	Sim
and	Sim	Sim	Sim
vehicle	Sim	Não	Sim
dra	Não	Sim	Não
coffee	Sim	Sim	Sim
parrot	Sim	Não	Não
november	Sim	Não	Não
person	Sim	Não	Não
cosmetics	Sim	Sim	Sim
drink	Sim	Não	Não
superman	Não	Sim	Não
plan	Sim	Não	Sim
instrument	Sim	Não	Não
hand	Sim	Não	Não
watermelon	Sim	Não	Não
laptop	Sim	Sim	Sim

on	Sim	Não	Não
meme	Sim	Sim	Sim
black-and-white	Sim	Sim	Sim
one	Sim	Não	Não
kissing	Sim	Não	Não
people	Sim	Sim	Sim
skyscraper	Sim	Sim	Sim
of	Sim	Sim	Sim
apple	Não	Sim	Não
heart	Não	Sim	Não
december	Sim	Não	Não
soccer	Não	Sim	Não
sitting	Sim	Não	Não
biceps	Não	Sim	Não
sp	Não	Sim	Não
grass	Sim	Sim	Sim
floor	Sim	Não	Sim
image	Sim	Sim	Sim
bizarria	Sim	Não	Não
bird	Sim	Sim	Sim
closeup	Sim	Não	Não
outdoors	Sim	Sim	Sim
solanea	Sim	Não	Não
cartoon	Sim	Não	Não
pb	Sim	Não	Não
big	Sim	Sim	Sim
construction	Sim	Não	Não
pet	Não	Sim	Não
road	Sim	Sim	Sim
bedroom	Sim	Não	Sim
braids	Sim	Não	Não
mota	Não	Sim	Não
long	Sim	Não	Não
sports	Sim	Não	Não
snail	Sim	Não	Não

musical	Sim	Não	Não
tool	Sim	Não	Não
shorts	Sim	Não	Não
twilight	Sim	Sim	Sim
wallet	Sim	Sim	Não
railroad	Não	Sim	Sim
saddle-stitched	Sim	Sim	Sim
standing	Sim	Não	Não
necklace	Sim	Não	Não
bottle	Sim	Sim	Sim
tattoo	Sim	Sim	Não
monument	Sim	Não	Não
watch	Sim	Sim	Sim
purse	Sim	Sim	Não
snake	Não	Não	Sim
rose	Sim	Sim	Sim
beach	Sim	Sim	Sim
map	Sim	Sim	Sim
rt	Sim	Sim	Sim
flower	Sim	Sim	Sim
coast	Não	Sim	Não
leather	Sim	Sim	Sim
enf	Não	Sim	Não
sky	Sim	Sim	Sim
sandals	Sim	Sim	Sim
april	Sim	Não	Não
chess	Sim	Sim	Não
office	Sim	Não	Sim
cloud	Sim	Sim	Sim
helicopter	Sim	Não	Não
phone	Sim	Não	Não
cake	Sim	Sim	Sim
iss	Sim	Não	Não
cupcake	Não	Sim	Não
pool	Sim	Sim	Sim

guitar	Sim	Não	Não
planet	Não	Não	Sim
screenshot	Sim	Sim	Sim
march	Sim	Não	Não
living	Sim	Não	Não
playing	Sim	Não	Não
strawberry	Sim	Não	Sim
ring	Sim	Não	Não
screen	Sim	Sim	Sim
maltese	Sim	Sim	Não
esp	Não	Sim	Sim
child	Sim	Sim	Sim
hair	Sim	Não	Não
fireworks	Não	Sim	Não
baby	Sim	Sim	Sim
beard	Sim	Sim	Sim
night	Não	Sim	Não
elephant	Não	Sim	Não
tower	Sim	Não	Não
toy	Sim	Sim	Não
street	Sim	Não	Sim
persian	Não	Sim	Não
sculpture	Não	Sim	Não
illustration	Sim	Não	Não
palm	Sim	Sim	Sim
animal	Sim	Sim	Não
dessert	Sim	Não	Sim
in	Sim	Não	Não
park	Sim	Não	Não
body	Sim	Sim	Sim
babys-breath	Sim	Não	Não
indoor	Sim	Não	Não
capital	Sim	Não	Não
instruments	Sim	Não	Não
high-heeled	Sim	Sim	Não

boots	Não	Não	Sim
anime-style	Sim	Sim	Não
january	Sim	Não	Não
or	Sim	Não	Não
june	Sim	Não	Não
camera	Não	Sim	Sim
table	Sim	Não	Sim
woodwork	Sim	Não	Não
brick	Sim	Não	Não
corn	Sim	Não	Não
with	Sim	Sim	Sim
balloon	Sim	Não	Não
television	Sim	Não	Não
wrist	Sim	Sim	Sim
bridge	Sim	Sim	Sim
mountain	Sim	Não	Não
motorcycle	Sim	Sim	Sim
plant	Não	Não	Sim

Tabela E.1: Termos extraídos dos textos alternativos das imagens para cada área profissional.

Apêndice F

Termos selecionados nos textos alternativos das imagem no conjunto de dados rotulado por faixa etária

Termo	Adolescentes	Adultos
railroad	Sim	Sim
on	Não	Sim
eyeglasses	Não	Sim
water	Sim	Sim
illustration	Não	Sim
cat	Sim	Sim
hair	Sim	Sim
heart	Sim	Não
buggy	Não	Sim
shorts	Não	Sim
trees	Sim	Sim
body	Sim	Sim
purse	Sim	Sim
image	Sim	Sim
balloon	Não	Sim
cupcake	Sim	Sim
drink	Não	Sim
guitar	Não	Sim

fishing	Não	Sim
nature	Sim	Sim
meme	Sim	Sim
lake	Sim	Sim
mountain	Não	Sim
twitter	Não	Sim
cake	Sim	Sim
of	Sim	Sim
otto	Não	Sim
closeup	Não	Sim
child	Sim	Sim
cup	Sim	Sim
fire	Não	Sim
may	Não	Sim
strawberry	Sim	Sim
bicycle	Não	Sim
motorcycle	Não	Sim
sunglasses	Sim	Sim
biceps	Não	Sim
shoes	Sim	Sim
headscarf	Não	Sim
night	Não	Sim
boots	Não	Sim
fries	Sim	Não
ring	Não	Sim
rose	Sim	Sim
park	Não	Sim
november	Não	Sim
phone	Não	Sim
tattoo	Não	Sim
jewelry	Sim	Sim
rod	Não	Sim
candy	Sim	Sim
anime-style	Sim	Sim
cartoon	Não	Sim

footwear	Sim	Sim
dog	Sim	Sim
food	Sim	Sim
coast	Não	Sim
toy	Sim	Sim
person	Não	Sim
screen	Não	Sim
bike	Não	Sim
leather	Sim	Sim
bottle	Sim	Não
superman	Não	Sim
playing	Não	Sim
watch	Sim	Sim
long	Não	Sim
one	Não	Sim
april	Não	Sim
smiling	Não	Sim
musical	Não	Sim
pool	Sim	Sim
wrist	Sim	Sim
saddle-stitched	Sim	Sim
coffee	Sim	Sim
football	Não	Sim
train	Não	Sim
standing	Não	Sim
beard	Sim	Sim
with	Sim	Sim
book	Sim	Sim
people	Sim	Sim
flower	Sim	Sim
rt	Não	Sim
horse	Não	Sim
palm	Sim	Sim
plant	Não	Sim
dessert	Sim	Sim

fireworks	Não	Sim
or	Não	Sim
dirt	Não	Sim
bedroom	Não	Sim
stone-fruit	Não	Sim
parrot	Sim	Não
tree	Sim	Sim
screenshot	Não	Sim
indoor	Não	Sim
animal	Não	Sim
brick	Não	Sim
high-heeled	Sim	Sim
hospital	Não	Sim
french	Sim	Não
black-and-white	Sim	Sim
sky	Sim	Sim
cosmetics	Sim	Sim
beach	Sim	Sim
ocean	Sim	Sim
outdoors	Sim	Sim
braids	Sim	Sim
baby	Sim	Sim
kissing	Não	Sim
dune	Não	Sim
more	Não	Sim
march	Não	Sim
car	Sim	Não
grass	Sim	Sim
toucan	Não	Sim
instruments	Não	Sim
twilight	Sim	Sim
wall	Não	Sim
cloud	Sim	Sim
road	Não	Sim
sitting	Não	Sim

Tabela F.1: Termos extraídos dos textos alternativos das imagens para faixa etária.

Referências

Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villaseñor-Pineda, L., and Meza, I. (2016). Evaluating topic-based representations for author profiling in social media. In Montes y Gómez, M., Escalante, H. J., Segura, A., and Murillo, J. d. D., editors, *Advances in Artificial Intelligence - IBERAMIA 2016*, pages 151–162, Cham. Springer International Publishing.

Anderson, M., Jiang, J., et al. (2018). Teens, social media & technology 2018. *Pew Research Center*, 31(2018):1673–1689.

Aragão, F. B. P., Farias, F. G., de Oliveira Mota, M., and de Freitas, A. A. F. (2016). Curtiu, comentou, comprou. a mídia social digital instagram e o consumo. *Revista Ciências Administrativas*, 22(1):130–161.

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Campos, Gabriela Enes e Costa, H. (2016). Caracterização dos perfis comerciais na rede social instagram. In *Anais do V Brazilian Workshop on Social Network Analysis and Mining*, pages 37–48. SBC.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR.

Cohen, W. W. (1995). Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Estival, D., Gaustad, T., Pham, S. B., Radford, W., and Hutchinson, B. (2007). Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Filho, J. A. B. L., Pasti, R., and de Castro, L. N. (2016). Gender classification of twitter data based on textual meta-attributes extraction. In Rocha, Á., Correia, A. M., Adeli, H., Reis, L. P., and Mendonça Teixeira, M., editors, *New Advances in Information Systems and Technologies*, pages 1025–1034, Cham. Springer International Publishing.

Filho, R. M., Carvalho, A., and Pappa, G. (2014). Inferência de sexo e idade de usuários no twitter. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, pages 200–211, Porto Alegre, RS, Brasil. SBC.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression : A statistical view of boosting. *Annals of statistics*, 28(2):337–407.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Han, K., Lee, S., Jang, J. Y., Jung, Y., and Lee, D. (2016). Teens are from mars, adults are from venus: analyzing and predicting age groups with behavioral characteristics in instagram. In *Proceedings of the 8th ACM Conference on Web Science*, pages 35–44.

Hoorn, J. F., Frank, S. L., Kowalczyk, W., and van Der Ham, F. (1999). Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, 14(3):311–338.

Jang, J. Y., Han, K., Shih, P. C., and Lee, D. (2015). Generation like: comparative characteristics in instagram. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4039–4042. ACM.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Kapovciute-Dzikiene, J., Venckauskas, A., and Damasvicius, R. (2017). A comparison of authorship attribution approaches applied on the lithuanian language. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 347–351. IEEE.

Khosla, A., Das Sarma, A., and Hamid, R. (2014). What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Lorena, A. C. and de Carvalho, A. C. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67.

Lui, M. and Baldwin, T. (2011). Cross-domain feature selection for language identification. In *Proceedings of 5th international joint conference on natural language processing*, pages 553–561.

Marcus, A., Wu, E., Karger, D. R., Madden, S., and Miller, R. C. (2011). Crowdsourced databases: Query processing with people. Cidr.

Mechti, S., Jaoua, M. B., and Belguith, L. H. (2013). A framework for plagiarism detection based on author profiling. *Notebook for PAN at CLEF*.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 9(214):237–249.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Oliveira, Y. R. d. (2014). O instagram como uma nova ferramenta para estratégias publicitárias.

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

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pfeil, U., Arjan, R., and Zaphiris, P. (2009). Age differences in online social networking–a study of user profiles and the social capital divide among teenagers and older users in myspace. *Computers in Human Behavior*, 25(3):643–654.

Rennie, J. D., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*.

Rodrigues, R. G., Pereira, W. W., Bezerra, E., and Guedes, G. P. (2017). Inferência de idade utilizando o liwc: identificando potenciais predadores sexuais. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil. SBC.

Roh, Y., Heo, G., and Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*.

Sasaki, Y. et al. (2007). The truth of the f-measure. 2007.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference*, pages 37–52. Springer.

Schwartz, R., Tsur, O., Rappoport, A., and Koppel, M. (2013). Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Skiena, S. S. (2017). *The data science design manual*. Springer.

Smith, A. and Anderson, M. (2018). Social media use in 2018. *Pew Research Center*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, page 254–263, USA. Association for Computational Linguistics.

Song, J., Han, K., Lee, D., and Kim, S.-W. (2018). "is a picture really worth a thousand words?": A case study on classifying user attributes on instagram. *PLOS ONE*, 13(10):1–22.

Souza, F., de Las Casas, D., Flores, V., Youn, S., Cha, M., Quercia, D., and Almeida, V. (2015). Dawn of the selfie era: The whos, wheres, and hows of selfies on instagram. In *Proceedings of the 2015 ACM on conference on online social networks*, pages 221–231.

Tambi, R., Kale, A., and King, T. H. (2020). Search query language identification using weak labeling. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3520–3527.

Terrizzano, I. G., Schwarz, P. M., Roth, M., and Colino, J. E. (2015). Data wrangling: The challenging yourney from the wild to the lake. In *CIDR*.

Valiati, V. A. D., Faleiro, L. G., and Quadro, K. R. (2020). Seja um pato: características da produção de conteúdo do instagram tudo orna. *Cambiassu: Estudos em Comunicação*, 15(25):223–242.

Yakout, M., Ganjam, K., Chakrabarti, K., and Chaudhuri, S. (2012). Infogather: entity augmentation and attribute discovery by holistic matching with web tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 97–108.

Zhang, Y., Baghirov, F., Hashim, H., and Murphy, J. (2016). Gender and instagram hashtags: A study of# malaysianfood. In *Conference on Information and Communication Technologies in Tourism*.