UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

ARTUR GALVÃO HALLBERG

Uma Análise das Opiniões de Usuários do Twitter sobre Vacinas Utilizando Técnicas de Aprendizado de Máquina

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Engenharia da Computação

Orientador: Prof. Dr. Dante Augusto Couto Barone
Co-orientador: M.Sc. Eduardo Gabriel Cortes

Porto Alegre
2021

“Stay hungry.
Stay foolish.”
— STEVE JOBS

AGRADECIMENTOS

Agradeço à minha família pelo carinho e suporte dado ao decorrer destes anos de graduação.

Agradeço aos meus colegas e amigos pelo companheirismo ao longo desta jornada.

Agradeço ao meu orientador, Prof. Dante Barone e ao meu co-orientador Eduardo Cortes pela oportunidade de realizar este trabalho e por todos os conselhos e orientações dados ao decorrer de sua realização.

Por fim, agradeço a todos os professores e funcionários da UFRGS que de alguma forma contribuíram com minha formação.

RESUMO

Vacinas são uma técnica antiga, conhecida e utilizada há mais de 200 anos. No entanto, é provável que a chegada da pandemia COVID-19 tenha polarizado o debate público em torno dessa tecnologia a um nível nunca antes visto. Assim, este trabalho tem como objetivo determinar e compreender os fatores que levam os usuários brasileiros do Twitter a serem favoráveis ou não a vacinas, determinando primeiro as opiniões dos usuários em relação ao tema vacinação e, em seguida, utilizando técnicas de Aprendizado de Máquina para inferir informações demográficas e determinar quais são os fatores sociodemográficos que causam maior impacto na opinião dos usuários sobre vacinas. Em primeiro lugar, foram gerados conjuntos de dados compostos por informações demográficas relevantes de usuários que são favoráveis ou contrários a vacinas. Em seguida, a partir dos dados coletados, foram gerados gráficos mostrando as distribuições das informações demográficas obtidas e algoritmos de Aprendizado de Máquina foram aplicados aos conjuntos de dados a fim de gerar modelos relevantes para a pesquisa. Por fim, as informações coletadas nas etapas anteriores foram analisadas a fim de tirar conclusões relevantes sobre como cada fator demográfico considerado influencia a formação de opiniões dos usuários do Twitter sobre vacinas e seu uso. A metodologia proposta produziu resultados informativos e pertinentes, sendo possível verificar que idade e a localização são fatores que causam influência significativa nas opiniões dos usuários. Este trabalho propõe uma estrutura eficiente e ágil que pode ser fácil e prontamente implementada e estendida para entender não apenas as opiniões sobre vacinas, mas também as opiniões sobre qualquer assunto de debate público.

Palavras-chave: Artificial intelligence. machine learning. data mining. data analysis. twitter.

ABSTRACT

Vaccines are an old technique, known and used for over 200 years. However, it is likely that the arrival of the COVID-19 pandemic made the public debate around this technology become polarized at a level never seen before. Thus, this work aims to determine and understand factors that lead Brazilian users on Twitter to be favorable or not to vaccines by first determining users' stances in relation to the vaccination topic and then using Machine Learning methods to infer demographic information and determine which are the socio-demographic factors that cause the greatest impact on users' opinions on vaccines. First, data sets composed of relevant demographic information from users who stand for or against vaccines were generated. Then, from the collected data, charts were generated showing the distributions of the obtained demographic information and Machine Learning algorithms were applied to the data sets in order to generate relevant models for the research. Finally, the information collected in the previous steps was analyzed in order to draw relevant conclusions about how each demographic factor considered influences the formation of Twitter users opinions on vaccines and their use. The methodology proposed produced informative and pertinent results, and it was possible to determine that age and location are factors that cause significant influence on users' opinions. This work proposes an efficient and agile framework that can be easily and readily implemented and extended to understand not only stances on vaccines, but also opinions on any subject of public debate.

Keywords: artificial intelligence, machine learning, data mining, data analysis, Twitter.

LISTA DE ABREVIATURAS E SIGLAS

ML	Machine Learning
CNN	Convolutional Neural Network
NLP	Natural Language Processing
IBGE	Instituto Brasileiro de Geografia e Estatística
BoW	Bag-of-Words
API	Application Programming Interface

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A aplicação de vacinas é uma técnica antiga, conhecida e utilizada há mais de 200 anos. Porém, é provável que nunca tenha-se debatido tanto sobre esta tecnologia como nos últimos 18 meses. Desde quando passou-se a considerar a utilização de vacinas como uma possível solução para a pandemia de COVID-19, o assunto tornou-se um foco de discussão global, e apesar de historicamente sempre haverem existido divergências de opiniões em relação à eficiência e segurança da técnica de vacinação como forma de combate a doenças (WOLFE; SHARP, 2002), possivelmente nunca antes essa polarização foi tão forte e evidente como no último ano.

Diante deste cenário, naturalmente torna-se interessante responder à seguinte pergunta: o que faz com que uma pessoa seja a favor ou contra à utilização de vacinas? E a partir desta pergunta surge a principal motivação e objetivo deste trabalho: utilizar métodos computacionais para compreender e determinar quais são os principais fatores que influenciam as opiniões das pessoas em relação ao emprego de técnicas de vacinação como forma de combate a doenças. A partir da resposta desta questão, torna-se possível também entender como ocorre o surgimento e a propagação de ideias favoráveis e contrárias a vacinas.

Para atingir esses objetivos, propoe-se uma metodologia que consiste em coletar *tweets* sobre vacinação, determinar as opiniões sobre vacinas dos usuários que os escreveram, e então utilizar técnicas de inteligência artificial para inferir informações sociodemográficas sobre esses usuários e por fim determinar quais fatores demográficos influenciam mais significativamente as opiniões dos usuários sobre vacinas. Este trabalho apresenta um *framework* eficiente que pode ser facilmente estendido e aplicado não apenas ao tópico da vacinação, mas também a qualquer tópico de debate público. Em suma, as contribuições deste trabalho são as seguintes:

1. Um conjunto de dados de mais de 80000 instâncias composto pelas informações sociodemográficas de usuários do Twitter que posicionam-se de forma contrária ou favorável à utilização de vacinas
2. Um *framework* eficiente que pode ser estendido e aplicado a qualquer assunto de discussão pública.
3. Uma análise que tira conclusões relevantes sobre como cada fator demográfico influencia as opiniões de usuários do Twitter sobre vacinas.

Este trabalho está organizado de acordo com a seguinte estrutura. O capítulo 2 apresenta os conceitos teóricos nos quais este trabalho baseia-se. O capítulo 3 realiza uma análise de outros trabalhos relacionados a este. No capítulo 4, a metodologia e as etapas realizadas para alcançar os objetivos propostos para este trabalho são apresentados em detalhes. Os resultados obtidos com a aplicação do técnicas explicadas no capítulo 4 são apresentados no capítulo 5. No capítulo 6, uma análise detalhada dos resultados apresentados no capítulo 5 é realizada. Finalmente, conclusões gerais sobre este trabalho, seus resultados e direções futuras são apresentadas no capítulo 7.

2 FUNDAMENTAÇÃO TEÓRICA

Os conceitos nos quais este trabalho é baseado são apresentadas abaixo.

2.1 Web scraping

Web scraping (WEBSCRAPING,) é uma técnica de mineração de dados que consiste em extrair informaçoes de páginas *web* para posteriormente usá-las em aplicações diversas. Existem diversas estratégias diferentes para a realização de *web scraping*. As técnicas utilizadas neste trabalho utilizam uma estratégia bastante simples, a programação HTTP (HTTPSCRAPING,), que consite em programar solicitações HTTP e capturar informações das páginas retornadas, automatizando processos que poderiam ser realizados acessando a *web* manualmente.

2.2 Convolutional neural network (CNN)

Uma CNN (VALUEVA et al., 2020) é uma categoria de rede neural articial amplamente utilizada no processamento de imagens digitais. Tem como principal característica demandar pouco pré-processamento quando comparada à outras estratégias de análise de imagens. CNNs utilizam uma variação de perceptrons multicamada desenvolvidos com o objetivo de diminuir o tempo de pré-processamento, e são capazes de "aprender"a otimizar filtros que em outros algoritmos necessitariam ser implementados manualmente.

2.3 Bag-of-words (BoW)

BoW (BAGOFWORDS,) é um modelo de representação utilizado em NLP onde um texto é representado como um multiconjunto de suas palavras. Desconsidera-se a ordenação das palvras e a estrutura gramatical do texto, porém a multiplicidade das palavras é preservada. Neste tipo de representação, geralmente cada objeto (texto) analisado é representado como um conjunto de atributos, onde cada atributo representa uma palavra diferente e armazena a quantidade de vezes que a palavra representada está presente no texto. Neste trabalho, utiliza-se uma variação de BoW binária, que ao invés de conter atributos que armazenam a quantidade de vezes que dada palavra aparece no texto,

possui atributos que simplesmente informam se o texto contém dada palavra (valor "1", verdadeiro) ou não a contém (valor "0", falso).

2.4 Aprendizado baseado em árvores de decisão

Técnica de ML onde um modelo de árvore de decisão é utilizado para inferir a classe do objeto analisado. Neste trabalho, utilizam-se árvores de classificação, árvores utilizadas quando a classe que deseja-se inferir é discreta. Neste tipo de árvore, folhas representam rótulos de classe e ramos representam conjunções de atributos que levam a esses rótulos. Esta categoria de algoritmos gera estruturas visuais simples e fáceis de enteder, e portanto é um dos tipos de algoritmos de ML mais utilizados (WU et al., 2007).

3 TRABALHOS RELACIONADOS

O trabalho (RECUERO; ZAGO; BASTOS, 2014), de forma semelhante a este, tem como objetivo utilizar técnicas computacionais para compreender quais eram as principais características dos discursos dos protestos ocorridos no Brasil em junho de 2013. Além disso, o trabalho objetiva entender quais são as diferenças de discursos entre as diferentes regiões geográficas brasileiras.

Para atingir tais objetivos, utilizou-se a API pública da rede social Twitter para coletar *tweets* relacionados aos protestos. Estes *tweets* foram então categorizados de acordo com a região do país de onde se originam (Centro-Oeste, Nordeste, Norte, Sudeste e Sul). Foi então contabilizada a ocorrência de certas palavras-chave nas mensagens coletadas, as frequências obtidas foram ordenadas de forma decrescente e a partir destes dados determinou-se quais eram os cem assuntos mais discutidos em cada região e em todo o país.

Apesar de utilizar uma estratégia de implementação relativamente simples, o trabalho apresentou resultados bastante informativos, demonstrando a eficiência da metodologia proposta pelos autores. A ideia geral desta metodologia, coletar e analisar mensagens de redes sociais a fim de compreender as ideias e opiniões expressas por seus usuários, serviu como inspiração para a metodologia proposta neste trabalho.

Além do trabalho mencionado anteriormente, outros trabalhos relacionados merecem ser citados. No trabalho (DARWISH et al., 2020), palavras-chave, *hashtags* e interações entre usuários (*retweets*) são usados em combinação com técnicas de redução de dimensionalidade e *clustering* para detectar as opiniões de usuários famosos do Twitter em relação a tópicos controversos.

No trabalho (BECHINI et al., 2020), a fim de compreender as opiniões de usuários do Twitter em relação ao tópico de vacinação na Itália, um modelo de análise de sentimentos foi usado para classificar os *tweets* em três classes distintas: *Neutral*, *InFavor* e *NotInFavor*. Com base nesses *tweets* categorizados, determinou-se as opiniões dos usuários, e cada região administrativa do país foi analisada em termos de suas porcentagens de usuários a favor e contra vacinas.

O trabalho (GOMIDE et al., 2011) tem como objetivo utilizar informações obtidas através do Twitter para monitorar epidemias de dengue no Brasil. Este trabalho, propõe uma metodologia de vigilância baseada em quatro dimensões: volume, localização, tempo e percepção do público. Análise de sentimentos é realizada a fim de selecionar apenas os

tweets que são relevantes para a vigilância da dengue, e os *tweets* obtidos são analisados sob uma perspectiva espaço-temporal.

4 METODOLOGIA

De forma geral, a estratégia de implementação adotada por este trabalho consiste em utilizar algoritmos de ML para gerar modelos capazes de prever as opiniões de usuários da rede social Twitter sobre vacinas com base em suas informações sociodemográficas, e a partir destes modelos determinar quais são os fatores que mais influenciam as opiniões sobre vacinas dos usuários. Esta estratégia pode de forma geral ser dividida em três etapas. Primeiramente, é necessário gerar um conjunto de dados de qualidade. Depois, deve-se escolher algoritmos relevantes e aplicá-los ao conjunto gerado anteriormente. Por fim, com a etapa anterior concluída, torna-se possível analisar os modelos gerados e extrair conclusões relevantes.

Para a primeira etapa, optou-se por coletar dados da rede social Twitter. Redes sociais, de forma geral, contém um grande volume de mensagens geradas por pessoas reais, o que as torna um terreno fértil para coletar informações a serem utilizadas em trabalhos de análise comportamental. Mais especificamente, escolheu-se neste trabalho utilizar a rede social Twitter a partir do conhecimento prévio de que esta rede possui um altíssimo número de usuários globalmente (TWITTERUSUARIOS,), usuários estes que sabe-se frequentemente utilizam a rede para expressar opiniões pessoais sobre assuntos diversos. Além disso, já existe uma grande variedade de ferramentas voltadas para extração de dados do Twitter, o que faz com que a utilização desta rede social seja também bastante conveniente.

Para a segunda etapa, priorizou-se naturalmente a escolha de algoritmos que gerassem como saída estruturas altamente informativas e preferencialmente de interpretação direta e simples. Na terceira etapa, o foco da análise realizada é principalmente determinar o nível de influência que cada fator demográfico considerado exerce sobre as opiniões dos usuários a respeito de vacinas e vacinações, visando por consequência determinar os fatores mais influentes.

4.1 Geraçao do conjunto de dados

A primeira etapa do desenvolvimento consiste em gerar conjuntos de dados compostos por informações demográficas relevantes de usuários do Twitter que se posicionam a favor ou contra vacinas (REPOSITORIO,). Para gerar estes conjuntos de dados, primeiramente separaram-se conjuntos de *hashtags* relevantes. Sabendo-se que a utilização

de *hashtags* é uma forte característica desta rede social, e sabendo-se também que frequentemente, ao expressar opiniões, usuários desta rede incluem nos textos dos *tweets* que escrevem diferentes *hashtags* associadas às diferentes opiniões que desejam expressar, é previsível que usuários com opiniões favoráveis à vacinas utilizem certas *hashtags* enquanto usuários com opiniões contrárias utilizem outras. Desta forma, gerando-se dois conjuntos, um composto por *hashtags* utilizadas por usuários que expressam-se a favor de vacinas e outro composto por *hashtags* utilizadas por usuários que expressam-se contra, torna-se simples, através de pesquisas baseadas nos elementos presentes em cada conjunto, encontrar e categorizar usuários de acordo com suas opiniões.

Para encontrar e categorizar essas *hashtags*, utilizaram-se dois métodos. O primeiro consistiu em utilizar *search engines* para encontrar artigos científicos e notícias que citassem *hashtags* e relacionassem sua utilização no Twitter à expressão de opiniões favoráveis ou contrárias a vacinas. O segundo consistiu em utilizar a funcionalidade de busca do Twitter para pesquisar palavras-chave relevantes, analisar os resultados, e através de "tentativa e erro", encontrar *hashtags* apropriadas e categorizá-las. A estratégia utilizada provou-se eficiente, e as *hashtags* obtidas foram separadas em dois conjuntos disjuntos, de acordo com as opiniões que expressam, batizados de *anti-vaccine* e *pro-vaccine* (ver Tabela 4.1 abaixo). Nesta fase do trabalho, escolheu-se focar a análise em usuários brasileiros, assim, selecionou-se apenas *hashtags* em português utilizadas por brasileiros.

Tabela 4.1: Conjuntos de *hashtags* anti e pró-vacina.

Opinião	*Hashtags*
anti-vacina	#VacinaNao, #EuNaoVouTomarVacina, #VacinaMata, #NaoVouTomarVacina, #VacinasNao, #NaoÀsVacinas, #ChegaDeVacina, #VacinasMatam, #NaoQueroVacina, #NaoVouTomar, #ContraVacina, #VacinasCausamAutismo, #NaoAVacina, #NaoTomoVacina, #ForaVacina
pró-vacina	#VacinaParaTodosJa, #VacinaParaTodos,#VacinaSim, #VacinaJa, #VemVacina, #VacinaSalva, #TodosPelasVacinas, #VacinaFunciona, #VacinaÉAmorAoPróximo, #VacinaAgora, #QueroSerVacinado, #QueroSerVacinada, #ExijoVacina, #VivaAVacina, #QueroVacina, #VacinasFuncionam, #ProVacina, #VacinasPelaVida, #VacinasSalvamVidas, #Vacinese, #EuQueroVacina, #EuQueroVacina, #VacinasSalvam, #VacinasJa, #VacinasFuncionam

Os conjuntos construídos foram então utilizados para gerar duas listas de IDs, uma contendo IDs de *tweets* que contém em seu texto *hashtags* contidas no conjunto *anti-vaccine*, e outra contendo IDs de *tweets* que contém em seu texto *hashtags* contidas no conjunto *pro-vaccine*. Para gerar tais listas, foram escritos *scripts* utilizando-se a linguagem de programação Pyhton, versão 3.8.5 (mesma linguagem utilizada em todos os *scripts* escritos para este trabalho), e um módulo *open-source* conhecido como Twint (TWINT. . . ,). O Twitter possui uma API oficial para coletar informações de *tweets*, porém esta ferramenta possui uma limitação considerável no contexto dos objetivos deste trabalho: permite apenas coletar informações de *tweets* de até sete dias atrás, a partir da data e hora da utilização da API. A ferramenta Twint, por sua vez, é capaz de contornar esta limitação e acessar informações de *tweets* publicados em qualquer data. Para tal, ao invés de utilizar a API oficial, utiliza- a funcionalidade de busca do Twitter para buscar *tweets*.

Utilizando-se a ferramenta Twint, escreveu-se dois *scripts*, um para cada um dos dois conjuntos de *hashtags* obtidos anteriormente. Em cada *script*, utilizou-se como ter-

mos de busca as *hashtags* do conjunto associado, separadas pelo operador booleano OR, definiu-se como data inicial de busca o dia primeiro de março de 2019, e como data final de busca o dia primeiro de março de 2021. Assim, ao executar-se ambos os *scripts*, obteve-se duas listas, uma contendo IDs de *tweets* que expressam ideias anti-vacina, outra contendo IDs de *tweets* que expressam ideias pró-vacina, ambas contendo *tweets* escritos entre primeiro de março de 2019 e primeiro de março de 2021. As datas inicial e final de busca foram escolhidas de forma estratégica, de forma a capturar um período de dois anos, onde aproximadamente o primeiro ano corresponde a um período pré-pandemia de Covid-19 e aproximadamente o segundo ano corresponde a um período com pandemia. Desta forma, capturando-se estes dois contextos diferentes, torna-se possível analisar como a ocorrência da pandemia de Covid-19 influenciou as opiniões sobre vacinas.

Naturalmente, a partir apenas de IDs de *tweets* não é possível realizar-se análises informativas a respeito das opiniões de usuários. É necessário também coletar outras informações a respeito destes *tweets*, como suas datas de publicação, e dos usuários que os escreveram, como suas descrições de usuário (popularmente chamadas de biografias, ou *bios)*, suas localizações geográficas e outras informações demográficas relevantes. Para tal, utilizou-se uma ferramenta *open-source* conhecida como Twarc (TWARC,), e mais especificamente, o comando *hydrate* desta ferramenta. O comando *hydrate* recebe como entrada uma lista de IDs de *tweets*, e para cada id da lista, acessa a API oficial do Twitter e obtém todas as informações fornecidas pela API referentes ao *tweet* que este id representa, gerando como saída uma lista contendo todas as informações obtidas para cada *ID*. Assim, ao aplicar este comando a cada uma das duas listas obtidas anteriormente, obtém-se duas novas listas contendo para cada um dos *tweets* capturados pelos procedimentos anteriores uma rica gama de informações referentes a estes *tweets*, informações estas que podem ser utilizadas posteriormente para extrair conclusões relevantes acerca das opiniões dos usuários.

A partir das listas obtidas com a aplicação do comando *hydrate*, torna-se possível selecionar e extrair informações relevantes relacionadas aos *tweets* capturados. O primeiro passo deste processo consiste em obter a localização geográfica do usuário que publicou o *tweet*. O Twitter permite que cada usuário, de forma opcional, preencha nos dados de sua conta um campo contendo sua localização geográfica, porém este campo aceita qualquer tipo de texto, ou seja, um usuário pode por exemplo inserir uma localização que não existe, assim como é possível inserir diferentes textos que se referem à mesma localização. Por exemplo, pode escrever-se “porto alegre, rio grande do sul” para

referir-se à cidade de Porto Alegre localizada no Rio Grande do Sul, assim como pode escrever-se apenas “poa, rs” para referir-se a esta cidade. Nestas circunstâncias, de forma a facilitar a análise computacional dos dados, naturalmente torna-se necessário empregar algum método para “normalizar” as localizações, ou seja, detectar as diferentes formas de referir-se a uma mesma localização e definir a que localização de fato elas se referem. Ou seja, dado o exemplo acima, é necessário ser capaz de detectar que tanto “porto alegre, rio grande do sul” como “poa, rs” referem-se à mesma cidade, a cidade de Porto Alegre, localizada no estado do Rio Grande do Sul.

A ferramenta Twarc retorna um campo contendo uma *string* que contém a localização inserida pelo usuário, ou uma *string* vazia, caso nenhuma localização tenha sido inserida. Esta informação, porém, não foi normalizada. Portanto, a fim de normalizar esta localização, utilizou-se um módulo *open-source* em Python chamado de GeoPy (GEOPY,). Este módulo reúne as APIs oficiais de diversos serviços de mapeamento online, como Google Maps, Apple Maps, entre outros, e simplifica suas utilizações. Sabe-se que a funcionalidade de busca deste tipo de serviço frequentemente têm a capacidade de normalizar localizações geográficas, o que significa que, utilizando o exemplo anterior, ao buscar-se pelos termos “porto alegre, rio grande do sul” ou “poa, rs”, as ferramentas retornarão os mesmos resultados. Assim, a utilização das APIs destes serviços torna-se uma forma extremamente eficaz de resolver este problema.

Mais especificamente, dentre as APIs oferecidas pela ferramenta GeoPy, optou-se pela utilização da API do serviço de mapeamento conhecido como OpenStreetMap OpenStreetMap contributors. O OpenStreetMap é um serviço de mapeamento online de dados abertos, amplamente utilizado em aplicações comerciais desenvolvidas por grandes empresas, como Facebook, Amazon e Uber, e que possui uma API que gera resultados precisos e informativos. Assim, para normalizar as localizações, escreveu-se um *script* que percorre as listas obtidas através do comando *hydrate* e, para cada *tweet*, utiliza o campo correspondente de localização do usuário como entrada da API do OpenStreetMap, o que seria equivalente de certa forma a inserir o texto do campo na funcionalidade de busca deste serviço. Caso a ferramenta retorne algum resultado (a localização geográfica correspondente normalizada), este resultado é armazenado juntamente com as outras informações do *tweet*. Caso a ferramenta não seja capaz de retornar um resultado, pode-se deduzir que a localização não existe ou não foi informada pelo usuário (a *string* de entrada é vazia). Assim, através da utilização deste método, torna-se possível normalizar localizações reais informadas por usuários, informações esta que são muito úteis para

determinar quais são os fatores demográficos que influenciam mais significativamente as opiniões dos usuários.

Naturalmente, também é interessante obter algumas outras informações sóciodemográficas relativas aos usuários que escrevem os *tweets*, como idade, gênero e grupo étnico, ou raça. A aplicação da ferramenta Twarc não retorna estas três informações, o que faz com que seja necessário utilizar outro método para obtê-las. Mais especificamente, optou-se pela utilização de um módulo em Python conhecido como DeepFace (SERENGIL; OZPINAR, 2020). Esta ferramenta implementa uma CNN capaz de receber como entrada uma imagem contendo uma face humana e gerar como saída predições de idade, gênero e raça relativas à pessoa presente na foto. O comando *hydrate* retorna a URL da foto de perfil dos usuários que escreveram os *tweets*, portanto, escreveu-se um *script* que percorre as duas listas geradas anteriormente, salva as fotos de perfil localmente a partir de suas URLs, e aplica a ferramenta DeepFace sobre cada uma destas fotos, armazenando os resultados de idade, gênero e raça obtidos para cada foto juntamente com as outras informações já obtidas para seu respectivo usuário. Naturalmente, há casos em que a foto de perfil não contém nenhum rosto. Nestes casos a ferramenta utilizada, ao ser incapaz de detectar um rosto, dispara uma exceção, que é simplesmente ignorada e nenhum resultado é armazenado.

Por fim, com todos os dados necessários devidamente coletados e processados, torna-se possível gerar os conjuntos de dados finais. Para tal, escreveu-se dois *scripts*, um para cada lista, que percorre a lista e para cada um de seus *tweets* armazena informações relevantes obtidas anteriormente, que são a data de publicação do *tweet*, a descrição do usuário (biografia) que o escreveu, a localização do usuário que o escreveu, a idade, gênero e raça do usuário que o escreveu, e se o *tweet* em questão expressa uma opinião anti-vacina ou pró-vacina. Assim, ao final da execução dos *scripts*, geram-se dois *data sets*, ambos formados por instâncias que representam um usuário do Twitter e compostas por seis atributos (data de publicação do *tweet*, descrição do usuário, localização, idade, raça e gênero) e um atributo-alvo (classe) que representa a opinião do usuário, expressa pelo *tweet* que escreveu, acerca de vacinas, ou seja, informa se o usuário é anti-vacina ou pró-vacina. Naturalmente, um mesmo usuário pode escrever diversos *tweets* diferentes expressando suas opiniões, porém como no caso deste trabalho tem-se interesse em coletar dados de usuários únicos, escreveu-se nos *scripts* um filtro que não inclui uma instância no *dataset* final caso o usuário que a instância representa já tenha sido incluído no *dataset* anteriormente, evitando assim a geração de *data sets* com repetições de usuários.

4.2 Técnicas de análise

Os *datasets* obtidos foram analisados através da geração de dois tipos de estruturas visuais: gráficos e árvores de decisão.

4.2.1 Gráficos de distribuição absoluta

Após concluir a geração dos *datasets*, a primeira etapa de análise dos dados consistiu em calcular as distribuições absolutas dos atributos dos *datasets*, ou seja, para cada valor possível de cada atributo, calculou-se qual é a porcentagem de instâncias do conjunto de dados que possui este valor. O objetivo deste procedimento é observar o comportamento das ferramentas de inferência utilizadas e avaliar a representatividade estatística da amostra coletada, ou seja, julgar o quão bem a amostra representa a população brasileira. As porcentagens foram calculadas de forma direta através de um *script* em Python que percorre os datasets, e os gráficos foram gerados utilizando-se o módulo MatPlotLib (HUNTER, 2007), bastante conhecido e amplamente utilizado para realizar análises estatísticas.

4.2.2 Gráficos de distribuição por classes

A segunda etapa de análise dos dados consistiu em calcular as distribuições por classe dos atributos, ou seja, para cada valor possível de cada atributo, calculou-se qual é a porcentagem de instâncias que contém este valor que que pertence à classe anti-vacina e qual é a porcentagem que pertence à classe pró-vacina, e então gerou-se gráficos apresentando estes resultados. O objetivo deste procedimento é tornar possível, antes mesmo de executar algoritmos mais complexos e através apenas da análise destes gráficos, chegar-se a conclusões relevantes acerca de como os valores de cada atributo influenciam as opiniões dos usuários. Novamente, as porcentagens foram calculadas de forma direta através de um *script* em Python e os gráficos foram gerados utilizando-se o módulo MatPlotLib.

4.2.3 Análise de palavras-chave

A ferramenta Twarc também retorna um atributo contendo a descrição do usuário, popularmente conhecida como biografia, ou simplesmente *bio*. Neste atributo, incluído nos *data sets*, cada usuário pode, opcionalmente, escrever um pequeno texto de até 280 caracteres descrevendo a si mesmo. Sabendo-se que usuários do Twitter frequentemente usam suas descrições de usuário para divulgar informações pessoais, como por exemplo suas ideologias políticas e profissões, naturalmente torna-se interessante analisar estas biografias visando extrair informações que complementem os dados sociodemográficos já obtidos anteriormente.

Para tal, primeiramente utiliza-se a classe CountVectorizer de um módulo em Python conhecido como Scikit-learn (PEDREGOSA et al., 2011) para converter as descrições para representações do tipo BoW. Neste tipo de representação, cada objeto (texto) analisado é representado como um conjunto de atributos, onde cada atributo representa uma palavra diferente e armazena a quantidade de vezes que a palavra representada está presente no texto. A classe CountVectorizer extrai todas as palavras presentes no conjunto de descrições do *dataset* e converte cada descrição para uma representação BoW que utiliza as palvras extraídas, retornando uma matriz onde cada linha representa uma descrição de usuário e cada coluna uma palavra que pode ou não estar presente em dada descrição (linha). Como uma pequena variação do algoritmo, optou-se por utilizar uma representação BoW binária, que ao invés de conter atributos que armazenam a quantidade de vezes que dada palavra aparece na descrição, possui atributos que simplesmente informam se a descrição contém dada palavra (valor "1", verdadeiro) ou não a contém (valor "0", falso).

Decidiu-se, então, a partir das representações BoW gerada anteriormente, obter as 20 palavras que aparecem com maior frequência nas descrições dos usuário anti-vacina, e também as 20 palavras que aparecem com maior frequência nas descrições dos usuário pró-vacina. Para tal, simplesmente selecionou-se todas as linhas da matriz gerada anteriormente que representam descrições anti-vacina e somou-se as colunas, obtendo-se um vetor de atributos onde cada atributo representa uma palavra e informa em quantas descrições de usuários anti-vacina esta palavra apareceu. Este vetor é então ordenado de forma de forma decrescente e as 20 primeiras palavras são selecionadas. Um processo semelhante foi utilizado para obter as 20 palavras que aparecem com maior frequência nas descrições dos usuário pró-vacina, selecionando linhas que representam usuários pró-

vacina ao invés de linhas que representam usuário anti-vacina. Os resultados obtidos são representados visualmente através de gráficos.

Após obter-se as 20 palavras mais frequentes para cada classe, constatou-se que que palavras que representam ideologias políticas e profissões aparecem consistemente entre as mais frequentes. Assim, decidiu-se que seria interessante analisar palavras-chave relacionadas a ideologias políticas e profissões a fim de determinar as distribuições percentuais das classes anti-vacina e pró-vacina em relação a ideologias políticas e profissões.

De forma a analisar as distribuições percentuais das ideologias políticas dos usuários incluídos no *dataset*, escreveu-se uma lista de palavras-chave que representam ideologias políticas. Localizou-se as palavras que compõem a lista na matriz gerada anteriormente. Dividiu-se a matriz em duas novas matrizes, onde uma é formada por linhas que representam descrições de usuários anti-vacina e outra por linhas que representam descrições de usuários pró-vacina. Somou-se as colunas destas matrizes que representam as palavras da lista, ou seja, obteve-se a contagem do total de vezes em que cada palavra da lista aparece nas descrições de usuários anti-vacina e o total de vezes em que cada palavra da lista aparece nas descrições de usuários pró-vacina. Então, Para cada umas das palavras selecionadas, determinou-se em que proporção a palavra é utilizada nas descrições de usuários anti-vacina e em que proporção é utilizada nas descrições de usuários pró-vacina. Para analisar as distribuições percentuais das classes em relação a profissões, o mesmo procedimento foi realizado, apenas substituindo-se a lista de ideologias políticas por uma lista de profissões. Os resultados para ambas as listas são representados visualmente através de gráficos.

4.2.4 Árvores de decisão

Por fim, após finalizar a geração de gráficos, optou-se por aplicar sobre os dados algoritmos para gerar árvores de decisão. A geração de árvores de decisão é uma técnica tradicional de ML, e no contexto deste trabalho, sua principal utilidade é o fato de que gera como saída estruturas simples, intuitivas e de fácil visualização e interpretação, o que facilita uma análise do comportamento dos dados. Geraram-se três árvores distintas: uma contendo apenas os atributos sociodemográficos inseridos no dataset, uma contendo apenas palavras-chave extraídas das descrições dos usuários e uma contendo atributos sociodemográficos e palavras-chave.

Para gerar as árvores, utilizou-se novamente o módulo Scikit-learn, e a classe

DecisionTreeClassifier. Como parâmetros de geração, utilizou-se como critério de seleção de atributos para divisão de nodos o critério "entropia"(ganho de informação) e limitou-se a profundidade máxima da árvore para o valor três, a fim de gerar uma árvore de mais fácil interpretação e evitar *overfitting* aos dados de treinamento. Para visualizar as árvores, utilizou-se o módulo Graphviz (GRAPHVIZ,), que inclui ao grafo de saída elementos que facilitam a interpretação de sua estrutura, como por exemplo, colore em azul nodos que contém uma maior proporção de instâncias pertencentes à classe pró-vacina e em laranja nodos que contém uma maior proporção de instâncias pertencentes à classe anti-vacina. Em ambos os casos, quanto mais escuras as cores maiores são as diferenças entre as proporções do nodo. Nodos brancos possuem proporções de classes iguais.

Também é importante observar que para gerar árvores de decisão a ferramenta Scikit-learn aceita apenas atributos numéricos, assim, os três atributos categóricos presentes nos *datasets* originais (localização, gênero e raça) tiveram de ser codificados para atributos numéricos, e para tal optou-se pela utilização da função OneHotEncoder do Scikit-learn, que, para cada valor de cada atributo presente nos *datasets* originais, cria um novo atributo que recebe "1"caso a instância contenha aquele valor e "0"caso contrário. A fim de facilitar a visualização das árvores, considerou-se apenas as regiões geográficas correspondentes às localizações dos atributos.

5 RESULTADOS

A união dos datasets gerados pela etapa anterior possui um total de 88062 instâncias, onde 73482 delas, ou seja, 83,44% das instâncias representam usuários pró-vacina e 14580 delas, ou seja, 16,56% das instâncias representam usuários anti-vacina. A partir da filtragem realizada na etapa anterior, constatou-se que, em média, cada usuário anti-vacina escreveu 1,89 *tweets* expressando opiniões desfavoráveis a vacinas no período considerado enquanto cada usuário pró-vacina escreveu em média 2,96 *tweets* expressando opiniões favoráveis a vacinas no mesmo período.

5.1 Gráficos de distribuição absoluta

Figura 5.1: Distribuição absoluta do atributo idade.

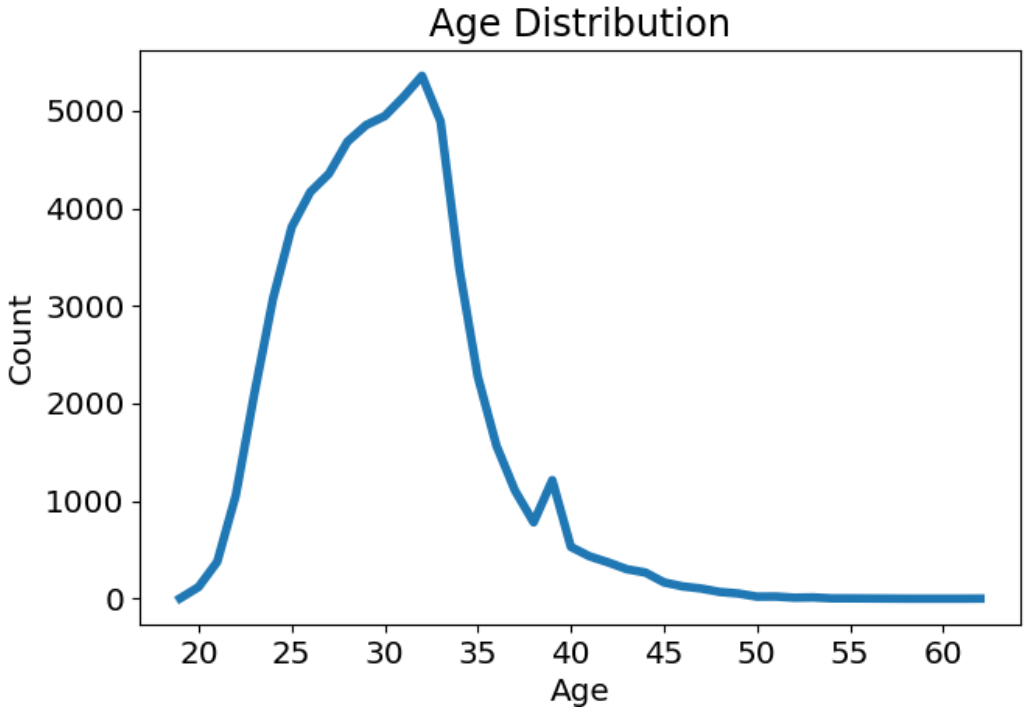


Figura 5.2: Distribuição absoluta do atributo gênero.

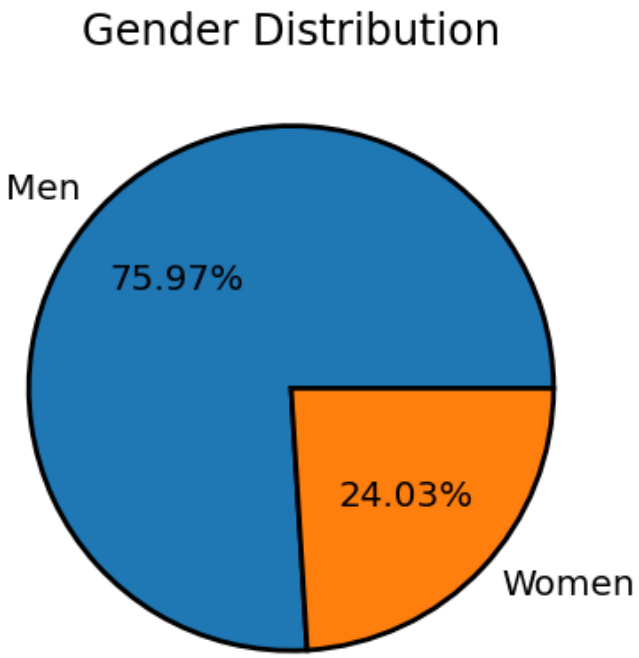


Figura 5.3: Distribuição absoluta do atributo raça.

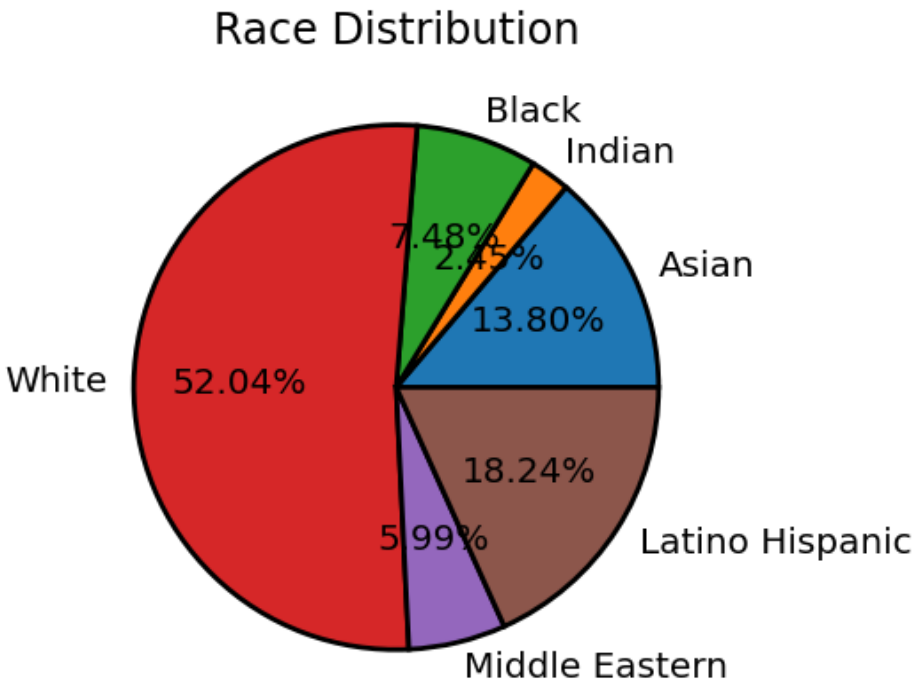


Figura 5.4: Distribuição absoluta de cada região geográfica brasileira, com base no atributo localização.

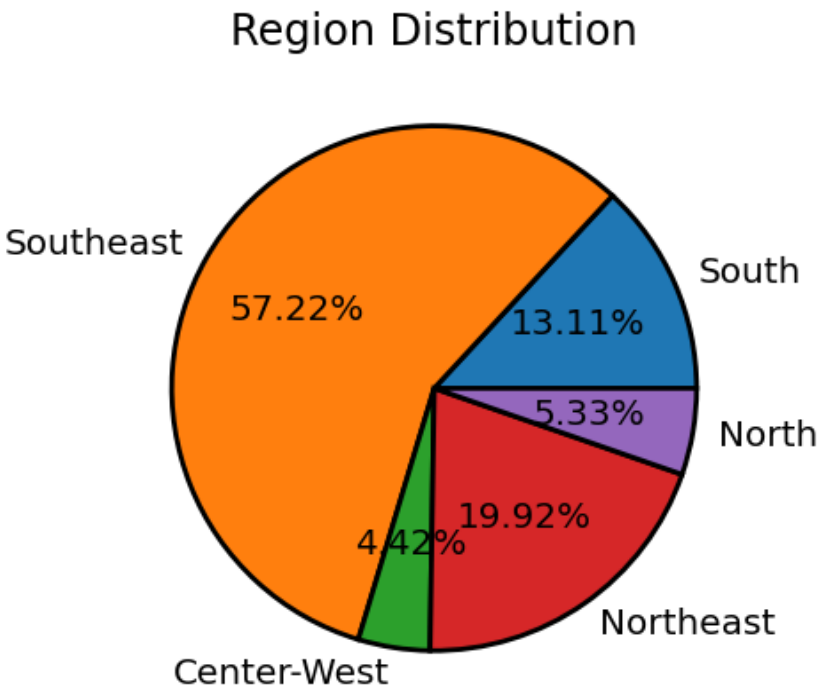
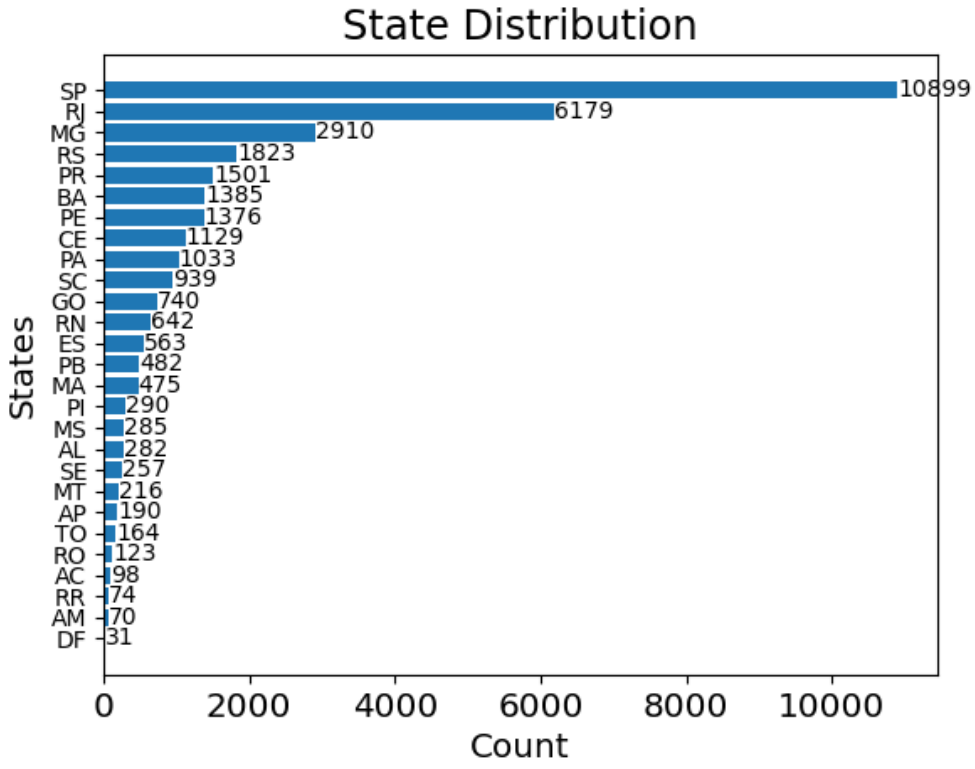


Figura 5.5: Distribuição absoluta de cada estado brasileiro, com base no atributo localização.



5.2 Gráficos de distribuição por classe

Figura 5.6: Ditribuição por classe do atributo idade.

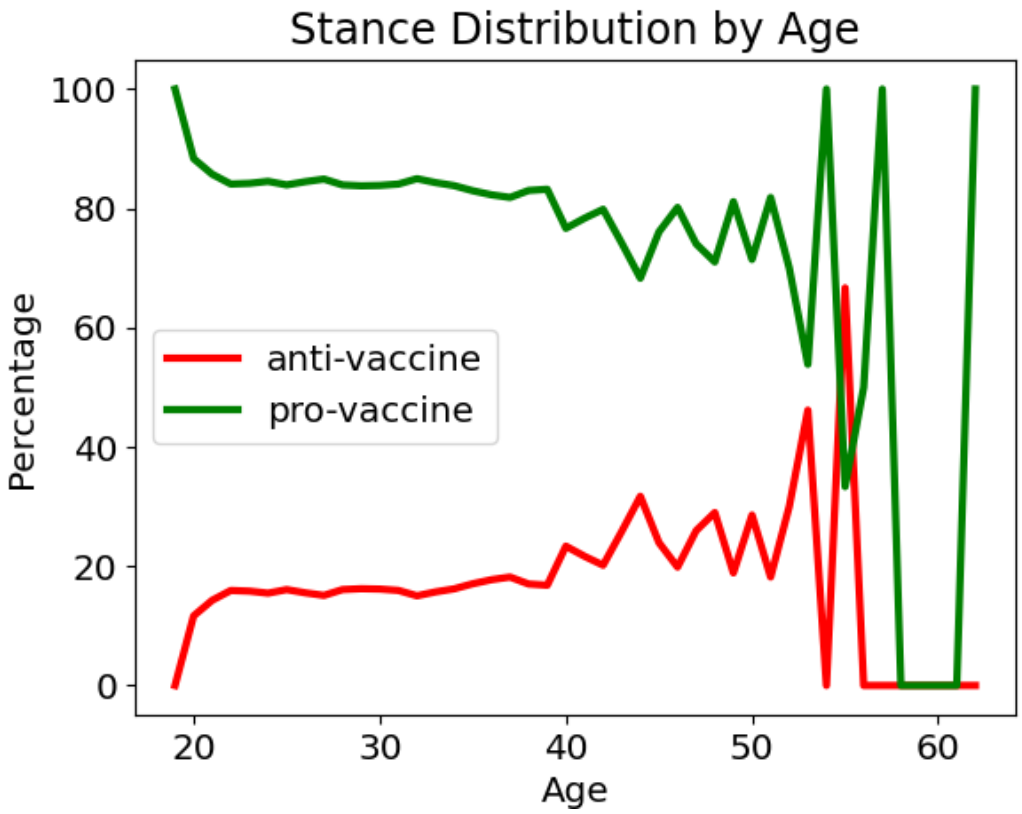


Figura 5.7: Número de *tweets* de cada classe publicados em cada dia do período analisado, em escala logarítmica.

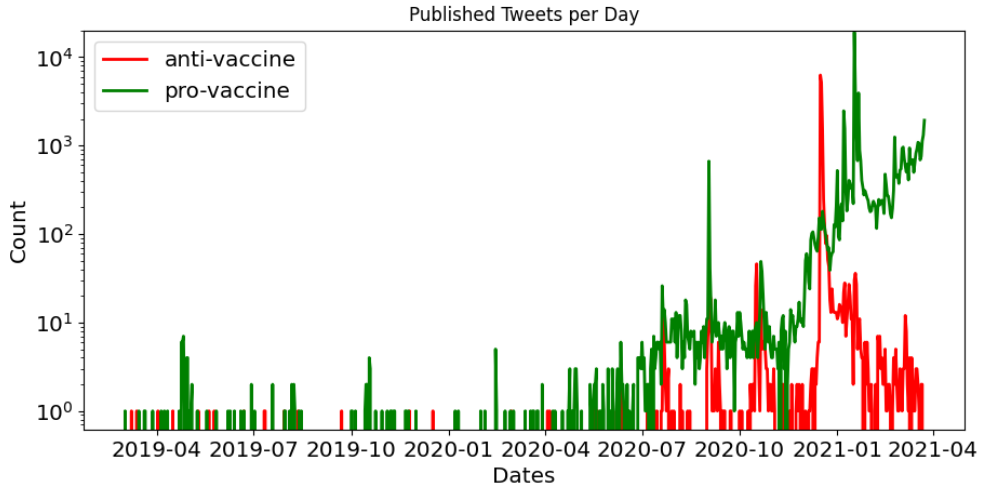


Figura 5.8: Ditribuição por classe do atributo idade.

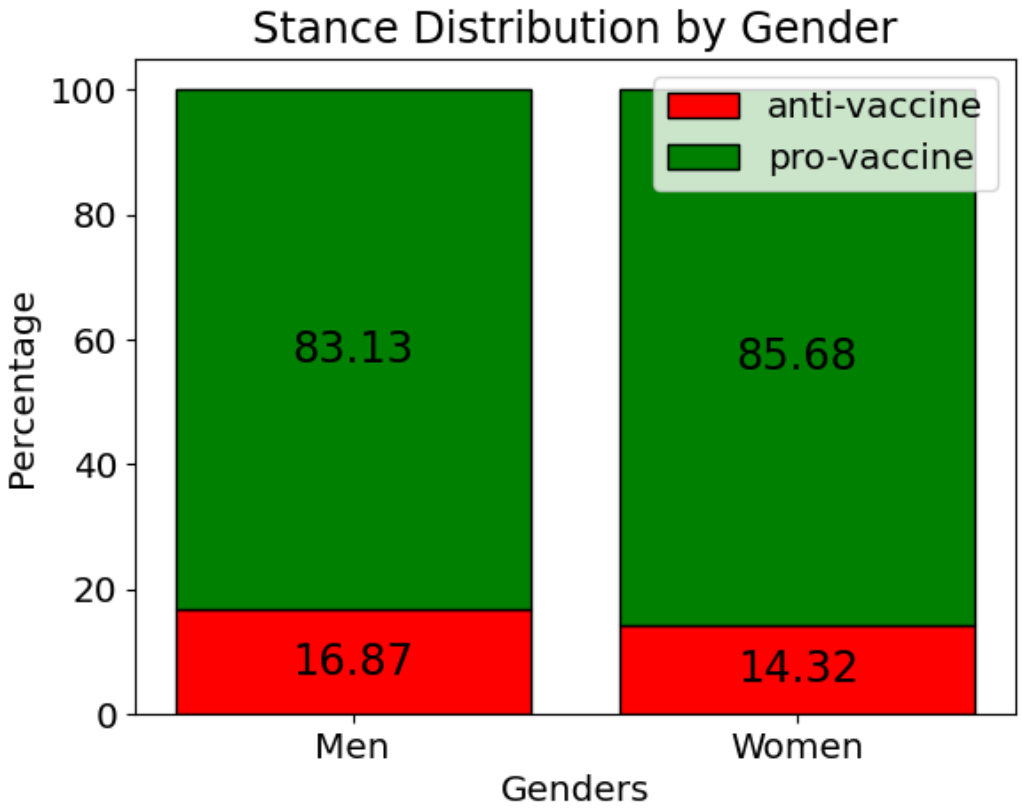


Figura 5.9: Ditribuição por classe do atributo raça.

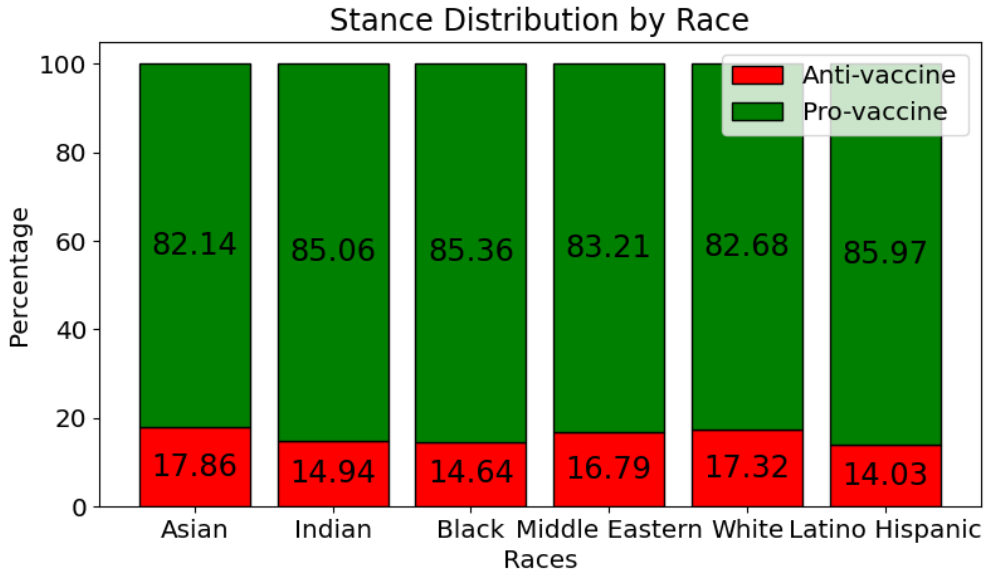


Figura 5.10: Ditribuição por classe de cada região geográfica brasileira, com base no atributo localização.

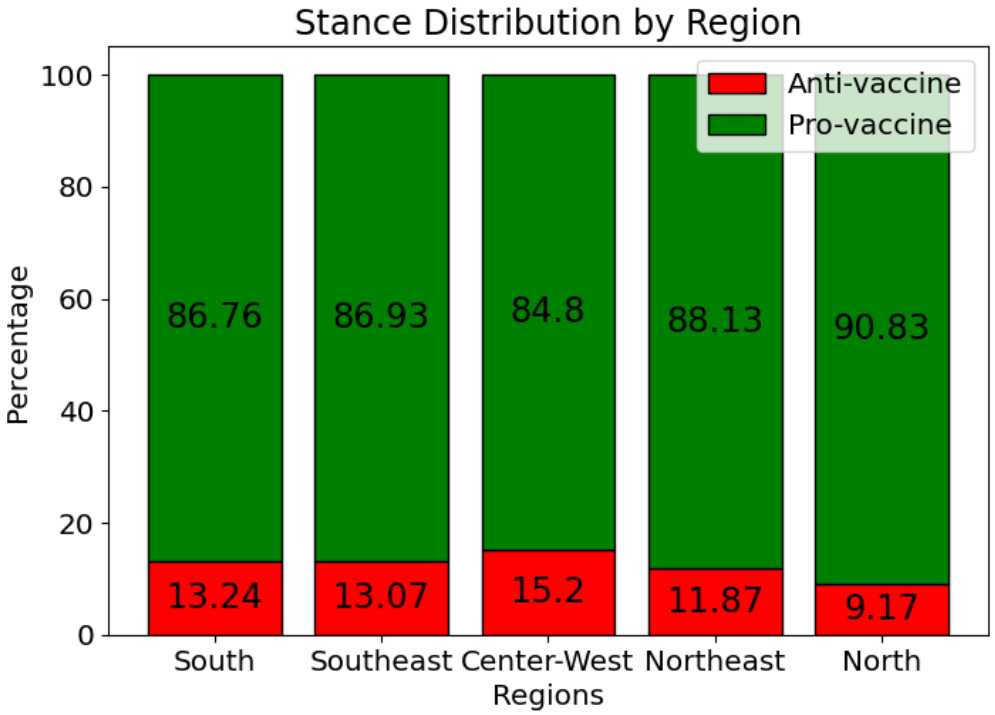
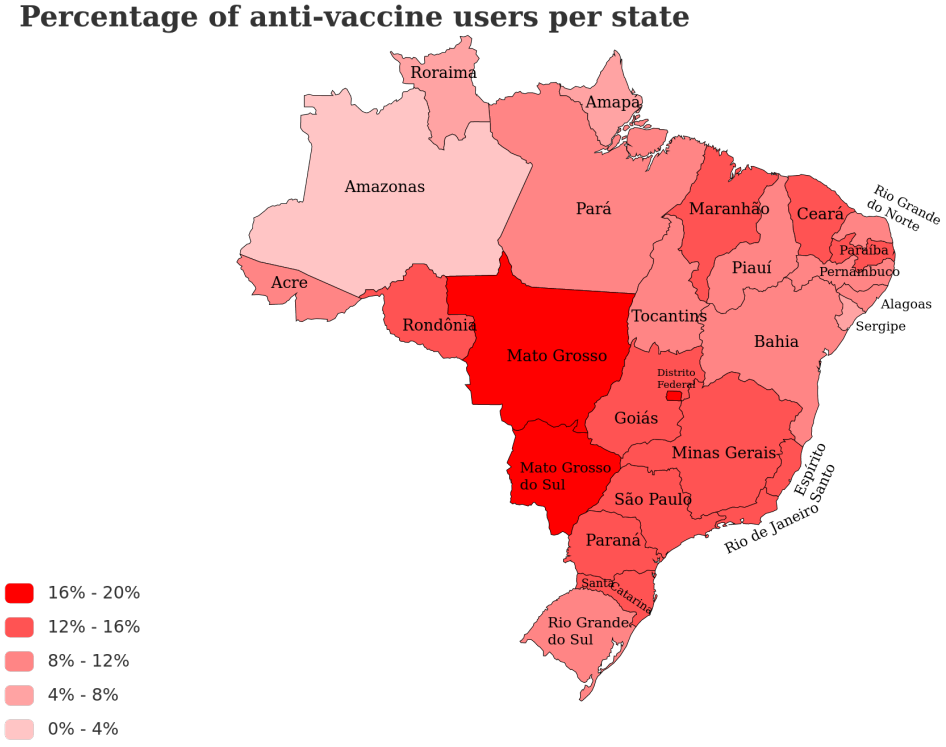


Figura 5.11: Porcentagem de usuários anti-vacina em cada estado brasileiro, com base no atributo localização. Tons mais escuros indicam porcentagens maiores.



5.3 Análise de palavras-chave

Figura 5.12: Vinte palavras-chave mais usadas na descrição por usuários anti-vacina.

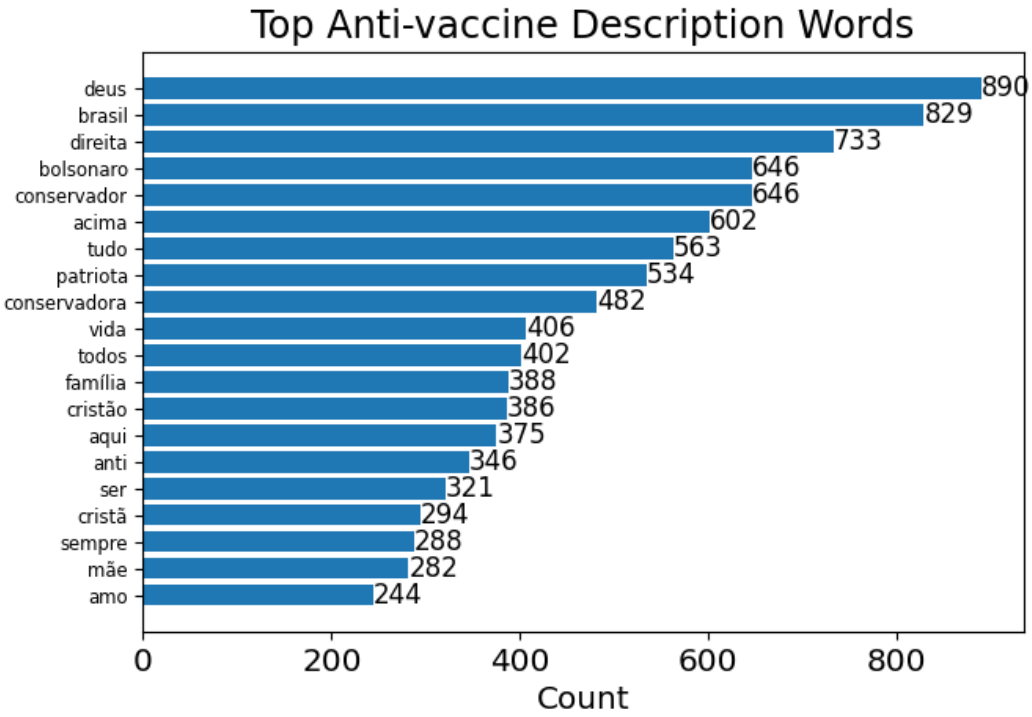


Figura 5.13: Vinte palavras-chave mais usadas na descrição por usuários pró-vacina.

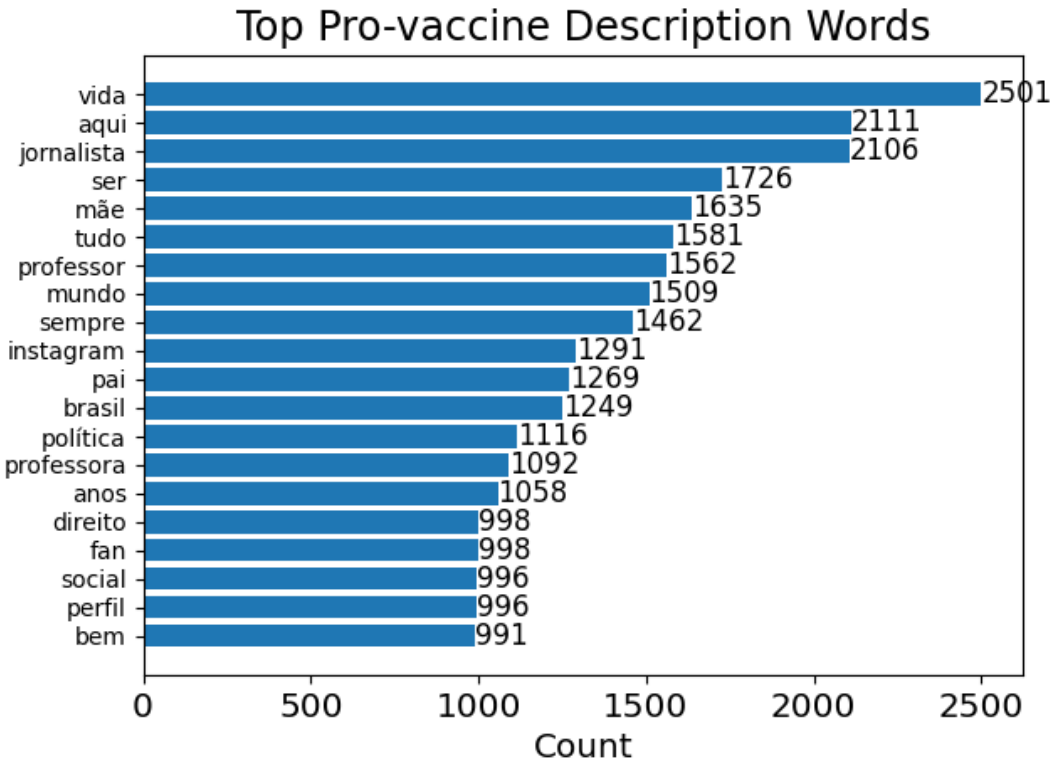


Figura 5.14: Ditribuição por classe de ideologias políticas.

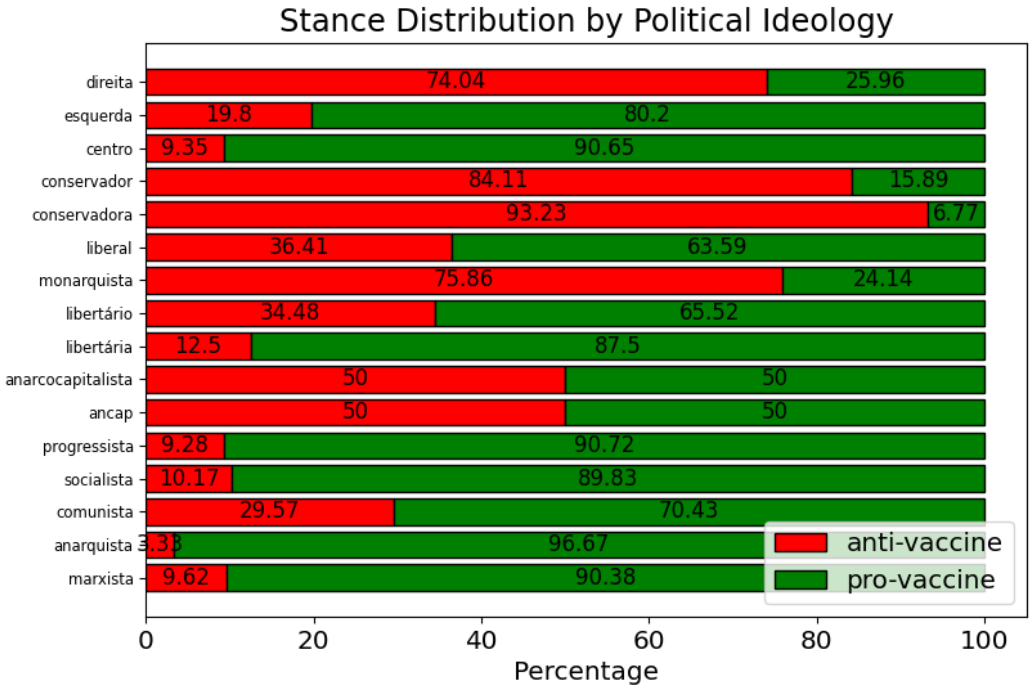


Figura 5.15: Ditribuição por classe de profissões (parte 1).

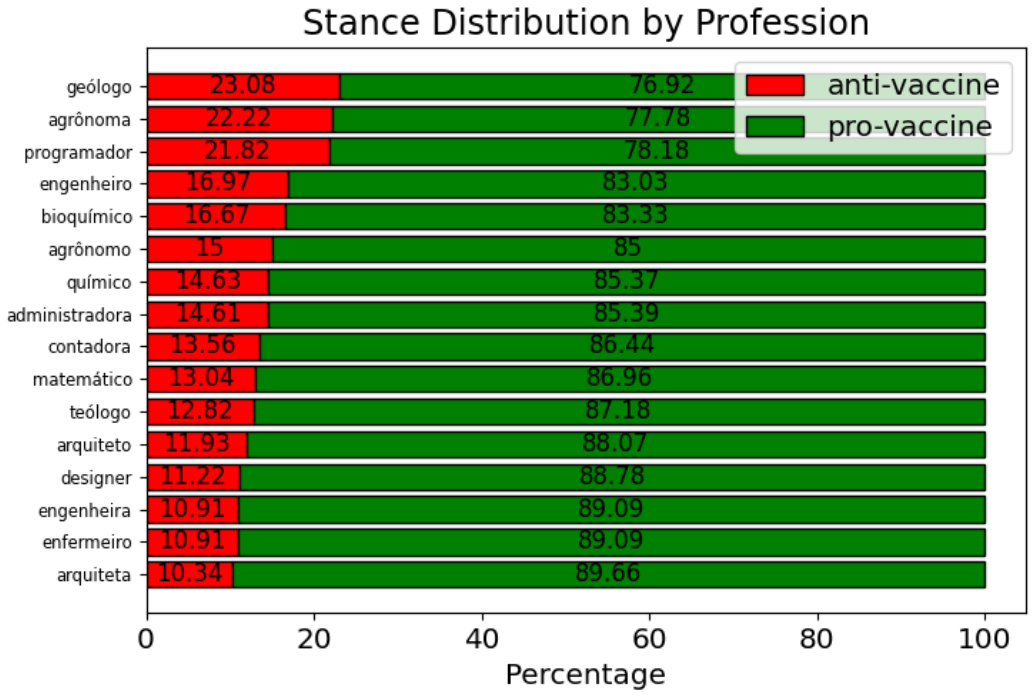


Figura 5.16: Ditribuição por classe de profissões (parte 2).

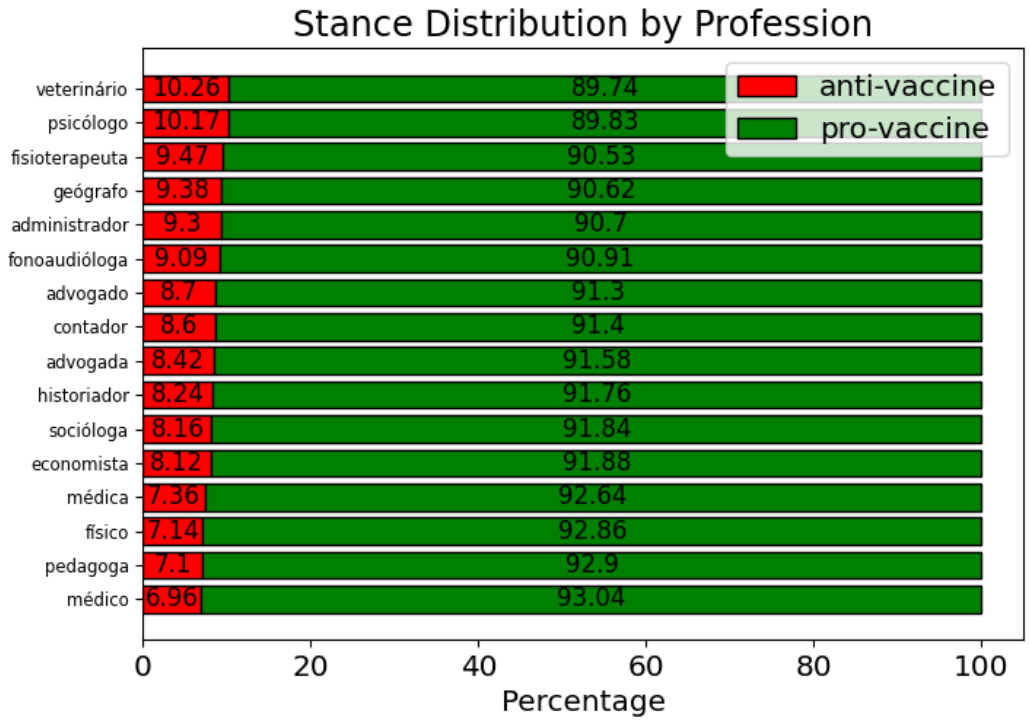
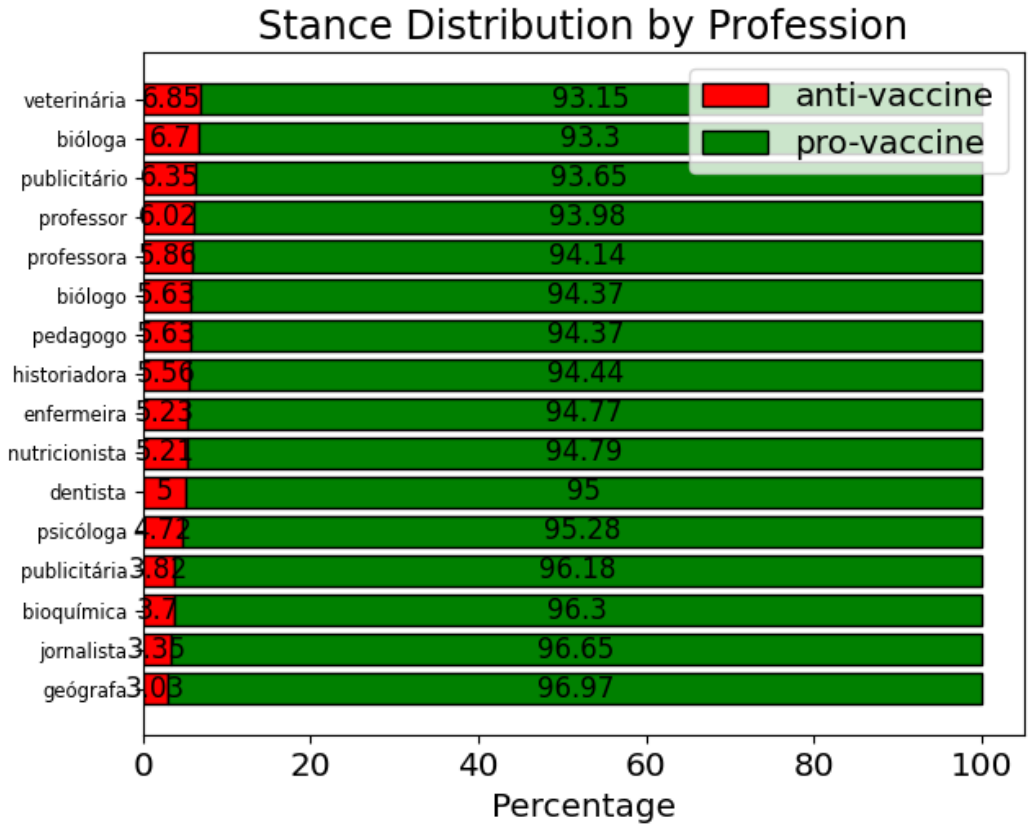


Figura 5.17: Ditribuição por classe de profissões (parte 3).



5.4 Árvores de decisão

Figura 5.18: Árvore de decisão contendo apenas os atributos sociodemográficos inseridos no dataset. Nodos azuis possuem uma proporção maior de instâncias pertencentes à classe pró-vacina, e nodos laranjas possuem uma proporção maior de instâncias pertencentes à classe anti-vacina. Em ambos os casos, tons mais escuros indicam diferenças maiores entre as proporções das classes do nodo. Nodos brancos possuem proporções de classe idênticas.

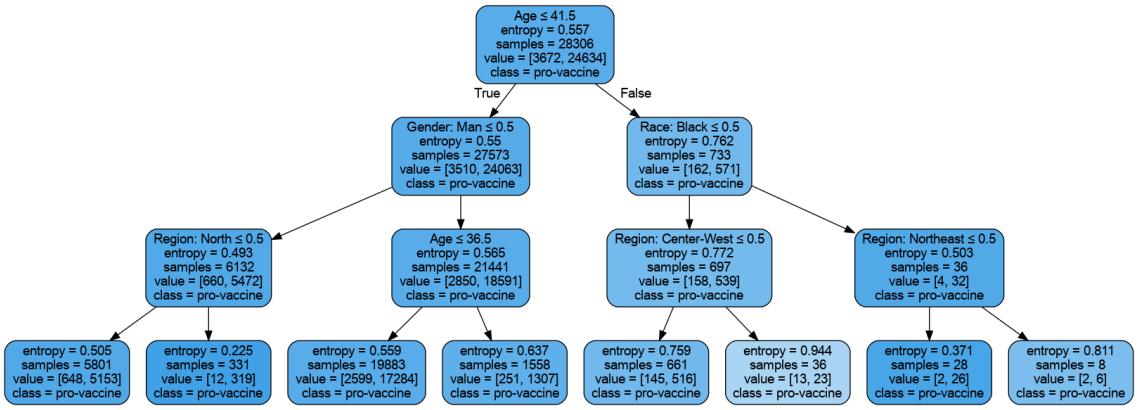


Figura 5.19: Árvore de decisão contendo apenas palavras-chave extraídas das descrições dos usuários. Nodos azuis possuem uma proporção maior de instâncias pertencentes à classe pró-vacina, e nodos laranjas possuem uma proporção maior de instâncias pertencentes à classe anti-vacina. Em ambos os casos, tons mais escuros indicam diferenças maiores entre as proporções das classes do nodo. Nodos brancos possuem proporções de classe idênticas.

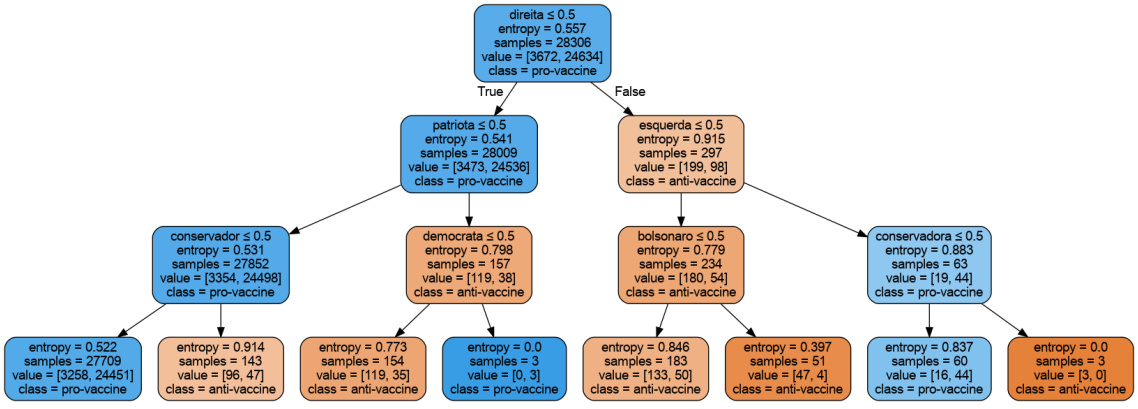
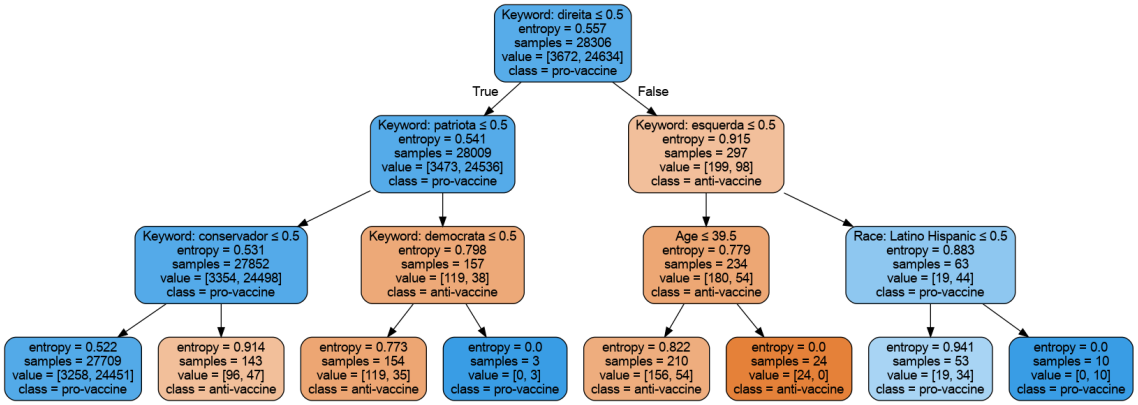


Figura 5.20: Árvore de decisão contendo atributos sociodemográficos e palavras-chave. Nodos azuis possuem uma proporção maior de instâncias pertencentes à classe pró-vacina, e nodos laranjas possuem uma proporção maior de instâncias pertencentes à classe anti-vacina. Em ambos os casos, tons mais escuros indicam diferenças maiores entre as proporções das classes do nodo. Nodos brancos possuem proporções de classe idênticas.



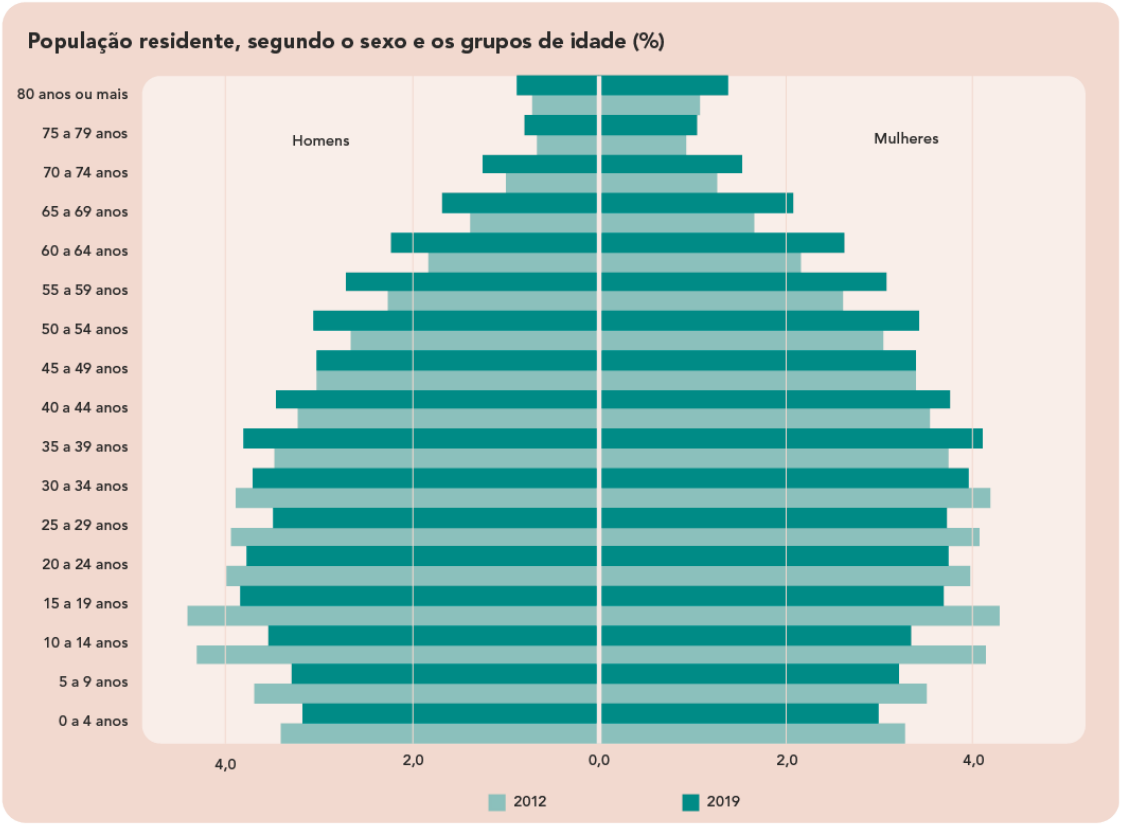
6 ANÁLISE DOS RESULTADOS

Uma análise detalhada dos resultados obtidos através da implementação da metodologia proposta neste trabalho é apresentada abaixo.

6.1 Gráficos de distribuição absoluta

A partir dos algoritmos executados, é possível chegar-se a conclusões interessantes. Observando o gráfico de distribuição absoluta de idades apresentado na Figura 5.1, observa-se que a distribuição de idades da amostra coletada é significativamente diferente da distribuição de idades da população brasileira (ver Figura 6.1 abaixo). Observa-se que que a menor idade detectada pelo algoritmo de predição de idades utilizado foi 19, enquanto a maior foi 62, porém a grande maioria dos usuários amostrados apresentam idades entre 25 e 35 anos, e há pouquíssimas amostras de usuários acima de 50 anos. Este comportamento é esperado, a partir do conhecimento prévio de que a rede social Twitter possui uma porcentagem de usuários entre 25 e 35 anos consideravelmente maior que sua porcentagem de usuários acima de 50 anos (TWITTERIDADES,).

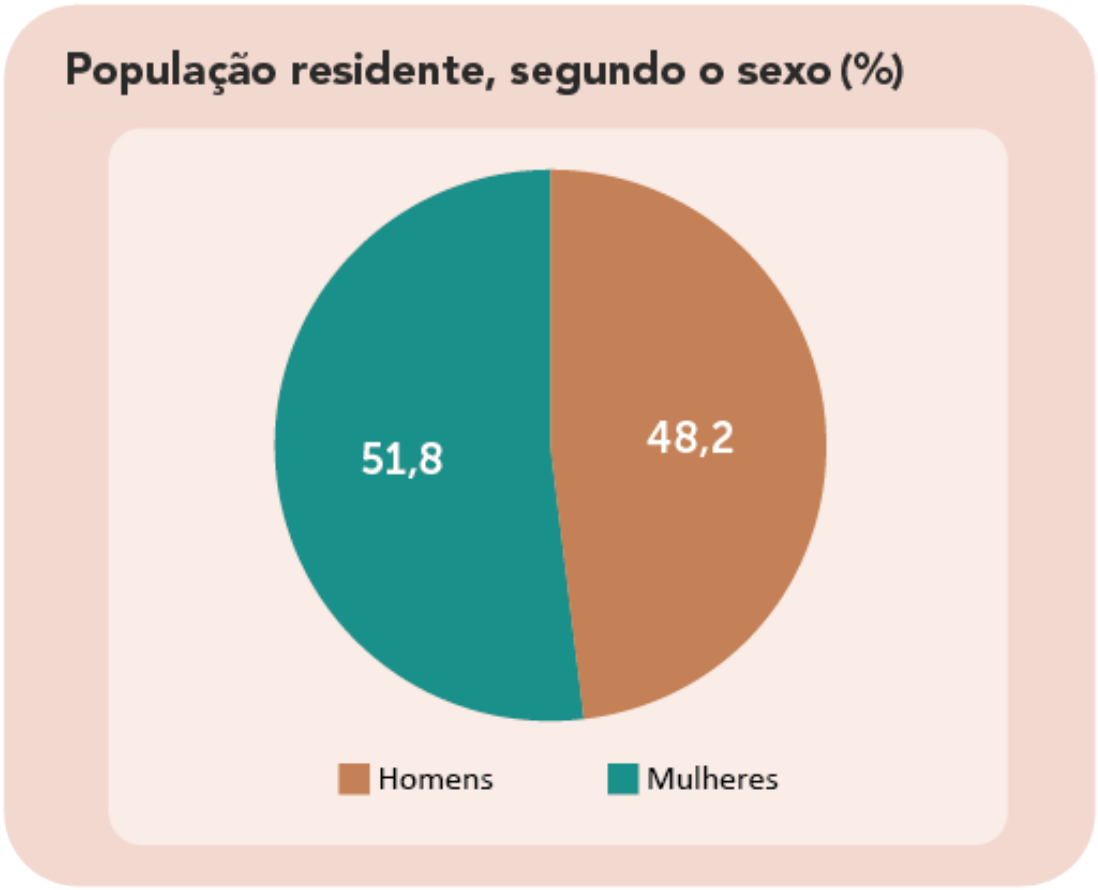
Figura 6.1: Pirâmida etária da população brasileira, segundo dados do IBGE. Fonte: https://educa.ibge.gov.br/jovens/conheca–brasil/populacao/18320-quantidade-de-homens-e-mulheres.html



Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional por Amostra de Domicílios Contínua 2012/2019.

Observando-se o gráfico de de distribuição absoluta de gêneros apresentado na Figura 5.2, e comparando-o à distribuição de gêneros da população brasileira (ver Figura 6.2 abaixo) observa-se que a distribuição de gêneros da amostra coletada é significativamente diferente da distribuição de gêneros da população brasileira. Aproximadamente 76% das instâncias coletadas correspondem a usuários homens, contra 24% que correspondem a usuárias mulheres. Este comportamento não é surpreendente, visto que sabe-se que de fato o Twitter possui uma porcentagem de usuários homens consideravelmente maior (TWITTERGENEROS,).

Figura 6.2: Distribuição de gêneros da população brasileira, segundo dados do IBGE. Fonte: https://educa.ibge.gov.br/jovens/conheca–brasil/populacao/18320-quantidade-de-homens-e-mulheres.html



Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional por Amostra de Domicílios Contínua 2012-2019.

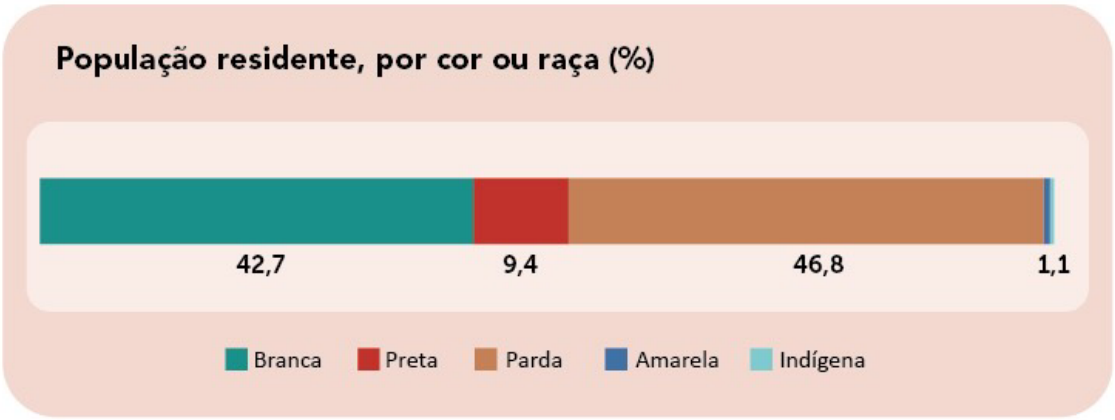
É necessário fazer algumas observações sobre o gráfico de raças apresentado na Figura 5.3, antes de analisar seus resultados. Observa-se que frequentemente diferentes países possuem diferentes definições de raças. Por exemplo, a definição de raça "parda", frequentemente utilizada no Brasil, não existe e não é utilizada em diversos outros países. O IBGE em seus censos considera oficialmente que a socio-demografia brasileira é composta por cinco raças: branco, pardo, preto, amarelo e indígena (IBGE-RACAS,). Naturalmente, torna-se desejável utilizar um algoritmo de predição de raças que seja capaz de capturar adequadamente a demografia brasileira, e de preferência, as cinco raças oficiais consideradas pelo IBGE. Infelizmente, não encontrou-se nenhuma ferramenta que capturasse exatamente estas raças, e portanto utilizou-se uma ferramenta que é capaz de capturá-las parcialmente. O pacote utilizado, DeepFace, é capaz de predizer seis raças, que são: asiático, indiano, negro, branco, do oriente médio e latino hispânico. Destas seis

raças, três delas fazem pouco sentido na demografia brasileira: indiano, do oriente médio e latino hispânico. Por outro lado, asiático, negro e branco, são raças que fazem sentido na demografia brasileira, e portanto ainda pode-se extrair algumas conclusões acerca dos resultados que o gráfico apresenta.

Além disso, é importante observar que o modelo implementado pelo DeepFace foi treinado utilizando-se a base de fotos do IMDB (Internet Movie Database), e estas fotos são geralmente de alta qualidade, possuem boa luminosidade e apresentam faces claras. Portanto, é de se esperar que o sistema não seja robusto à fotos de baixa qualidade, que frequentemente é o caso de fotos de perfil de usuários de redes sociais. Através de observações empíricas, constatou-se que quando as fotos possuem algumas características, como baixa luminosidade e utilização de óculos escuros, de forma geral as predições tendem a ser consideravelmente menos confiáveis.

Observando o gráfico de distribuição absoluta de raças apresentado na Figura 5.3, e comparando-o à distribuição de raças, ou cores, da população brasileira (ver Figura 6.3 abaixo), observa-se que a amostra coletada possui proporções de usuários brancos e negros (pretos) relativamente semelhantes às proporções presentes na população brasileira, porém a proporção de asiáticos (amarelos) da amostra é consideravelmente maior que a proporção da população brasileira, e as raças parda e indígena não são representadas no gráfico da Figura 5.3. É interessante observar que através de observações empíricas constatou-se que usuários brancos com imagens de perfil de baixa qualidade (luminosidade precária, utilização de óculos escuros) frequentemente são classificados como asiáticos, que usuários que usuários que foram classificados pelo DeepFace como pertencentes à raça latino hispânica frequentemente de fato pertencem à raça parda, e que a raça latino hispânica representa porcentagem considerável na distribuição absoluta de amostras.

Figura 6.3: Distribuição de cores, ou raças da população brasileira, segundo dados do IBGE. Fonte: https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18319-cor-ou-raca.html



Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional por Amostra de Domicílios Contínua 2012-2019.

Analisando-se o gráfico de distribuição absoluta de populações de regiões brasileiras apresentado na Figura 5.4, observa-se que obteve-se muito mais amostras de usuários das regiões Sul, Sudeste e Nordeste quando comparado ao número de amostras obtidas de usuários das regiões Centro-Oeste e Norte, o que é um resultado esperado, pois segundo dados do IBGE, as populações das regiões Centro-Oeste e Norte são de fato consideravelmente menores que as populações das regiões Sul, Sudeste e Nordeste (POPULACAOIBGE,). Naturalmente, como selecionou-se *tweets* escritos em português, existe a possibilidade de também obter-se *tweets* escritos por usuários naturais de outros países lusófonos, como Portugal, por exemplo, porém como sabe-se que a população brasileira é muito maior que a população destes países isto ocorre raramente, e nestes casos a instância e sua localização são simplesmente descartadas da análise.

Observando o gráfico de distribuição de populações de estados brasileiros apresentado apresentado na Figura 5.5, nota-se que os seis estados dos quais coletou-se os maiores números de amostras (SP, RJ, MG, RS, PR e BA) são também os estados mais populosos do Brasil, de acordo com dados do IBGE (POPULACAOIBGE,). Analisando-se informações do IBGE (POPULACAOIBGE,), nota-se que, de forma geral, as quantidades de amostras coletadas são compatíveis com a distribuição populacional dos estados brasileiros, conforme esperado.

A partir da análise dos gráficos de distribuição absoluta, conforme esperado a partir do conhecimento prévio das características dos usuários da rede social Twitter, nota-se que a população da amostra coletada não possui uma distribuição sociodemográfica perfeitamente compatível com a sociodemografia da população brasileira, e portanto não

deve-se generalizar as análises realizadas neste trabalho para toda a população brasileira. Porém, os dados obtidos neste trabalho representam adequadamente certos setores da população, portanto conclusões relevantes ainda podem ser extraídas. Também observa-se que o comportamento das ferramentas utilizadas para extrair as localizações geográficas e inferir as idades, gêneros e raças dos usuários é compatível com o esperado, considerando o conhecimento prévio das características dos usuários da rede social Twitter, o que indica que as ferramentas utilizadas possuem um bom nível de precisão.

6.2 Gráficos de distribuição por classe

Observando-se o gráfico de idades presente na Figura 5.6, observa-se como a partir de aproximadamente 32 anos em diante a porcentagem de usuários anti-vacina passa a aumentar de forma significativa, enquanto a porcentagem de usuários pró-vacina diminui, indicando que pessoas mais velhas possuem uma tendência maior a serem anti-vacina que pessoas mais jovens. Também observa-se como o gráfico apresenta um comportamento ruidoso para altas idades, a partir de aproximadamente 50 anos. Isto ocorre devido ao fato de existirem poucas instâncias com estes valores, o que naturalmente tende a causar ruídos.

A partir da análise do gráfico de datas da Figura 5.7, pode-se chegar a algumas conclusões interessantes em relação às quantidades de *tweets* anti-vacina e pró-vacina publicados e os eventos que ocorreram ao longo do período amostrado. Em primeiro lugar, é interessante observar como o número de *tweets* publicados durante o período relativo à pandemia de Covid-19 é muito maior que o número publicado antes da pandemia, o que é um comportamento esperado. Observa-se também que a partir de agosto de 2020, que sabe-se foi quando começou a haver notícias mais concretas em relação a utilização de vacinas como forma de combate à pandemia (CREDITOVACINA,) (TESTEVACINA,), tanto o número de *tweets* anti-vacina como de *tweets* pró-vacina publicados começa a crescer consideravelmente. Observa-se também como em meados de dezembro, que foi quando o STF julgou se a aplicação compulsória de vacinas era constitucional ou não (STF,), ocorre um grande pico de *tweets* anti-vacina. Por fim, observa-se como o maior pico de *tweets* pró-vacina ocorre no início de 2021, período que coincide com quando começou-se a aplicar vacinas para Covid-19 no Brasil (COMECOVACINACAO,), e há de forma geral um aumento considerável de publicações de *tweets* pró-vacina neste período.

Observando-se o gráfico de gêneros da Figura 5.8, nota-se que há uma porcentagem maior de homens que expressam ideias anti-vacina, quando comparada à porcentagem de mulheres que expressam estas mesmas ideias.

Observa-se que na Figura 5.9 as raças “asiático”, “branco” e “do oriente médio” possuem porcentagens de usuários anti-vacina semelhantes e maiores que as das raças “indiano”, “negro” e “latino hispânico”, que também são semelhantes entre si. É interessante observar, porém, que através de observações empíricas constatou-se que existe um número considerável casos em que a ferramenta prediz a raça “asiático”, porém de fato a raça do usuário é “branco”. Isto geralmente ocorre quando a qualidade da foto é baixa. Também é interessante observar que constatou-se através de observações empíricas que frequentemente quando a ferramenta prediz a raça “latino hispânico” de fato o que ocorre é que o usuário se enquadra na raça “pardo”. É difícil chegar a conclusões sobre a relação das raças “indiano” e “do oriente médio” com a demografia brasileira.

Analisando-se o gráfico de regiões da Figura 5.10, observa-se que a região Centro-Oeste possui a maior porcentagem de usuários anti-vacina, e que a região Norte possui a menor porcentagem. As regiões Sul, Sudeste e Nordeste possuem porcentagens de usuários anti-vacina relativamente semelhantes, porém a porcentagem da região Nordeste é um pouco menor quando comparada às porcentagens das outras duas.

Observando-se o mapa de estados apresentado pela Figura 5.11, nota-se que os estados que apresentam as maiores porcentagens de usuários anti-vacina são Mato Grosso do Sul, Mato Grosso e Distrito Federal, e o que apresenta a menor porcentagem é o Amazonas. A análise deste gráfico é bastante interessante, pois dentre todos os obtidos é o gráfico onde se observa as maiores diferenças de distribuições de classes entre os valores considerados. Por exemplo, observa-se uma diferença bastante significativa entre as porcentagens de usuários anti-vacina apresentadas pelos estados com as maiores porcentagens anti-vacina (Mato do Grosso do Sul, Mato Grosso e Distrito Federal) e a apresentada pelo estado com a menor porcentagem anti-vacina (Amazonas).

6.3 Análise de palavras-chave

Analisando-se o gráfico da Figura 5.12, que apresenta as palavras mais utilizadas nas descrições de usuário por usuários anti-vacina, observa-se que o gráfico apresenta uma quantidade considerável de palavras relacionadas a ideologias políticas.

Através da análise do gráfico da Figura 5.13, que apresenta as palavras mais utili-

zadas nas descrições de usuário por usuários pró-vacina, nota-se que o gráfico apresenta uma quantidade considerável de palavras relacionadas a profissões.

Observando-se o gráfico de distribuição por classe de ideologias políticas apresentado na Figura 5.14, nota-se que a palavra-chave utilizada pela maior porcentagem de usuários anti-vacina é a palavra "conservadora", enquanto que a palavra-chave utilizada pela maior porcentagem de usuários pró-vacina é a palavra "anarquista".

Nos gráficos de distribuição por classe de profissões apresentados nas Figuras 5.15, 5.16 e 5.17, observa-se que a profissão citada pela maior porcentagem de usuários anti-vacina é a profissão "geólogo", enquanto que a citada pela maior porcentagem de usuários pró-vacina é a profissão "geógrafa". Este gráfico apresenta grandes diferenças de distribuições de classes entre os valores considerados.

6.4 Árvores de decisão

Analisando-se a árvore de decisão apresentada na Figura 5.18, observa-se que o primeiro ponto de corte selecionado pelo algoritmo, ou seja, um dos mais significativos para deduzir as classes das instâncias, é a idade de 41 anos, o que conforme observado anteriormente, indica que há uma diferença significativa entre as opiniões dos usuários a partir de uma certa idade. Também observa-se que, de forma geral, os nodos à direita raíz, que possuem usuários com idade acima de 41 anos, apresentam tons de azul mais claros quando comparados aos nodos presentes à esquerda raíz, o que indica que os agrupamentos deste lado da árvore possuem porcentagens maiores de usuários anti-vacina. Por fim, observa-se que um dos nodos do lado direito da árvore dividem as instâncias com base em se os usuários que elas representam são ou não originários da região Centro-Oeste, e observa-se a partir do resultado desta divisão que a partição formada por usuários do Centro-Oeste possui uma porcentagem maior de usuários anti-vacina quando comparada à formada por usuários que não originários desta região.

Através da análise da árvore de decisão apresentada na Figura 5.19, observa-se que parte considerável das palavras selecionadas pelo algoritmo possuem cunho político. Diferente do que ocorre na árvore da Figura 5.18, a árvore da Figura 5.19 apresenta uma grande variação nas cores de seus nodos, indicando que é mais eficiente para inferir classes que a árvore da Figura 5.18.

Analisando-se a árvore de decisão apresentada na Figura 5.20, observa-se que ao combinar os dados originais do *data set* com palavras-chave as palavras-chave passam

a "dominar"o algoritmo, sendo selecionadas na maioria dos nodos, o que indica que são mais informativas que todos os outros atributos. Novamente, observa-se que as palavras em muitos casos possuem cunho político, e que a árvore apresenta uma grande variação nas cores de seus nodos.

7 CONCLUSÕES

Neste trabalho, utilizou-se algoritmos de ML para determinar quais são os fatores que causam maior influência nas opiniões de usuários do Twitter sobre vacinas. Primeiramente, gerou-se *datasets* onde cada instância representa um usuário e é formada por seis atributos (data de publicação do *tweet* opinativo, descrição do usuário, localização do usuário, idade, gênero e raça do usuário) e um atributo-alvo que representa a opinião do usuário em relação a vacinas (anti-vacina ou pró-vacina). Realizou-se então uma análise estatística sobre estes dados, e gerou-se gráficos relevantes. Através da aplicação de um algoritmo de ML sobre o dataset, gerou-se árvores de decisão. Os gráficos e a árvore foram então analisadas e extraiu-se conclusões relevantes acerca das opiniões dos usuários.

A partir das técnicas executadas nesta etapa do trabalho chegou-se a resultados e conclusões informativas e pertinentes, indicando a eficácia da metodologia proposta. Foi possível determinar que a utilização nas descrições de usuários de palavras-chave relacionadas a política é o fator que causa a influência mais significativa nas opiniões dos usuários. Além destas palavras, idade e localização são também fatores bastante influentes. Este trabalho propõe um framework ágil e eficiente que pode ser facilmente e prontamente implementado e estendido para compreender não apenas opiniões sobre vacinas, mas também opiniões sobre qualquer assunto de debate público.

Como trabalho futuro, seria interessante expandir o processo de extração informações dos conjuntos de dados gerados aplicando algoritmos de NLP nos dados obtidos, como análise de sentimentos, e novamente analisar e interpretar os resultados, a fim de gerar conclusões informativas e complementares às já obtidas. O trabalho foi limitado à análise das opiniões de usuários brasileiros, mas também seria interessante aplicar a mesma metodologia de geração e análise de dados proposta para usuários do Twitter de outros países, a fim de expandir o escopo de análise do trabalho e comparar os resultados obtidos para diferentes países.

REFERÊNCIAS

BAGOFWORDS. <https://scikit-learn.org/stable/modules/feature_extraction.html#the-bag-of-words-representation>. Accessed: 2020-04-30.

BECHINI, A. et al. Stance analysis of twitter users: the case of the vaccination topic in italy. **IEEE Intelligent Systems**, p. 1–1, 2020.

COMECOVACINACAO. <https://agenciabrasil.ebc.com.br/saude/noticia/2021-01/vacinacao-contra-covid-19-come%C3%A7a-em-todo-o-pais#>. Accessed: 2020-04-30.

CREDITOVACINA. <https://www.gov.br/pt-br/noticias/saude-e-vigilancia-sanitaria/2020/08/governo-abre-credito-de-r-1-9-bilhao-para-producao-e-aquisicao-de-vacina-contra-o-coronavirus>. Accessed: 2020-04-30.

DARWISH, K. et al. Unsupervised user stance detection on twitter. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 14, n. 1, p. 141–152, May 2020. Available from Internet: <https://ojs.aaai.org/index.php/ICWSM/article/view/7286>.

GEOPY. <https://github.com/geopy/geopy>. Accessed: 2020-04-30.

GOMIDE, J. et al. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In: **ACM Web Science Conference (WebSci)**. [S.l.: s.n.], 2011.

GRAPHVIZ. <https://gitlab.com/graphviz/graphviz>. Accessed: 2020-04-30.

HTTPSCRAPING. <https://docs.python-guide.org/scenarios/scrape/>. Accessed: 2020-04-30.

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

IBGE-RACAS. <https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18319-cor-ou-raca.html>. Accessed: 2020-04-30.

OpenStreetMap contributors. **Planet dump retrieved from https://planet.osm.org** . 2017. <https://www.openstreetmap.org>. Accessed: 2020-04-30.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

POPULACAOIBGE. <https://cidades.ibge.gov.br/>. Accessed: 2020-04-30.

RECUERO, R. da C.; ZAGO, G.; BASTOS, M. O discurso dos protestosbr: análise de conteúdo do twitter. **Galáxia. Revista do Programa de Pós-Graduação em Comunicação e Semiótica. ISSN 1982-2553**, v. 14, n. 28, 2014. ISSN 1982-2553. Available from Internet: <https://revistas.pucsp.br/index.php/galaxia/article/view/17911>.

REPOSITORIO. <https://github.com/Arturgh0/Vaccines>. Accessed: 2020-04-30.

SERENGIL, S. I.; OZPINAR, A. Lightface: A hybrid deep face recognition framework. In: IEEE. **2020 Innovations in Intelligent Systems and Applications Conference (ASYU)**. [S.l.], 2020. p. 23–27.

STF. <https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=457462&ori=1>. Accessed: 2020-04-30.

TESTEVACINA. <https://g1.globo.com/bemestar/vacina/noticia/2020/08/12/vacina-para-covid-19-da-biontech-e-pfizer-induziu-resposta-imune-robusta-mostram-resultados-preliminaghtml>. Accessed: 2020-04-30.

TWARC. <https://github.com/DocNow/twarc>. Accessed: 2020-04-30.

TWINT - Twitter Intelligence Tool. <https://github.com/twintproject/twint>. Accessed: 2020-04-30.

TWITTERGENEROS. <https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/>. Accessed: 2020-04-30.

TWITTERIDADES. <https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/>. Accessed: 2020-04-30.

TWITTERUSUARIOS. <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>. Accessed: 2020-04-30.

VALUEVA, M. et al. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. **Mathematics and Computers in Simulation**, v. 177, p. 232–243, 2020. ISSN 0378-4754. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0378475420301580>.

WEBSCRAPING. <https://www.techopedia.com/definition/5212/web-scraping>. Accessed: 2020-04-30.

WOLFE, R.; SHARP, L. Anti-vaccinationists past and present. **BMJ (Clinical research ed.)**, v. 325, p. 430–2, 09 2002.

WU, X. et al. Top 10 algorithms in data mining. **Knowl. Inf. Syst.**, Springer-Verlag, Berlin, Heidelberg, v. 14, n. 1, p. 1–37, dec. 2007. ISSN 0219-1377. Available from Internet: <https://doi.org/10.1007/s10115-007-0114-2>.