UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS MURAKAMI ROCHA DA COSTA

UM FRAMEWORK PARA ANÁLISE DE DISCURSO TRANSFÓBICO A PARTIR DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

RIO DE JANEIRO
2022

LUCAS MURAKAMI ROCHA DA COSTA

UM FRAMEWORK PARA ANÁLISE DE DISCURSO TRANSFÓBICO A PARTIR DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Profª. Drª. Jonice Oliveira
Co-orientadora: Profª. Drª. Lívia Ruback

RIO DE JANEIRO
2022

CIP - Catalogação na Publicação

C837f

Costa, Lucas Murakami Rocha da
Um framework para análise de discurso transfóbico a partir de técnicas de aprendizado de máquina / Lucas Murakami Rocha da Costa. -- Rio de Janeiro, 2022.
60 f.

Orientadora: Jonice de Oliveira Sampaio.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Computação, Bacharel em Ciência da Computação, 2022.

1. Transfobia. 2. Aprendizado de máquina. 3. Shapley additive explanations. I. Sampaio, Jonice de Oliveira, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

LUCAS MURAKAMI ROCHA DA COSTA

UM FRAMEWORK PARA ANÁLISE DE DISCURSO TRANSFÓBICO A PARTIR DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 29 de setembro de 2022.

BANCA EXAMINADORA:

Prof. Jonice de Oliveira Sampaio, D.Sc. (UFRJ)

Prof. Lívia Couto Ruback Rodrigues, D.Sc. (UFRRJ) - participação por videoconferência

Prof. Valeria Menezes Bastos, D.Sc.(UFRJ) - participação por videoconferência

Prof. Felipe Fink Grael, M.Sc. (Infnet/Twist) - participação por videoconferência

Luiz Paulo Carvalho, M.Sc.(UFRJ) - participação por videoconferência

Dedico esse trabalho aos meus pais,
sempre presentes e me apoiando.

AGRADECIMENTOS

Agradeço ao Luiz Paulo Carvalho pela recomendação do tema e orientações quanto à convocação de participantes trans para o trabalho.

Também agradeço a todas as pessoas que participaram do processo de geração da base de dados, especialmente as trans, dada a natureza difícil da tarefa, ao reservarem parte de seu tempo para ler e anotar *tweets* potencialmente transfóbicos e agressivos.

Agradeço às minhas orientadoras Jonice Oliveira e Lívia Ruback, por todo o apoio na formação deste trabalho com valiosas orientações e novas referências para que eu pudesse me aprofundar ainda mais no tema que abordamos. Também agradeço por toda a paciência, sobretudo no difícil período de pandemia de COVID-19.

Também agradeço a meu pai Flávio da Costa pela ajuda com a geração de diversas das imagens usadas neste trabalho e minha mãe Satsumi Murakami pela ajuda na formatação do trabalho.

Por fim, segue um agradecimento muito especial ao Felipe Fink Grael, por toda a ajuda com as técnicas de Aprendizado de Máquina e Interpretabilidade de modelos treinados. Sem suas orientações não teria sido possível alcançarmos os resultados que alcançamos.

RESUMO

O crescente uso das redes sociais online impactou grandemente a vida das pessoas e a sociedade. A comunicação entre pessoas, organização de eventos, grupos e negócios, por exemplo, atingiram novos patamares graças a essas novas tecnologias online. Porém, é possível notar que novos e antigos desafios também são impulsionados a partir desta nova dinâmica de comunicação. O discurso de ódio alcança a muito mais alvos e com maior velocidade graças às redes sociais e a automatização. Muitas plataformas procuram seus próprios meios de moderação dos seus espaços online, mas esta é uma tarefa que ainda está longe de chegar ao fim, se é que tal fim é alcançável. O Aprendizado de Máquina é uma área da Inteligência Artificial que busca criar modelos matemáticos para previsão de valores baseados em dados históricos. Diversos trabalhos procuram combater e estudar a difusão de discursos de ódio online se utilizando de tais modelos. Porém, muitos destes modelos são vistos como “caixas pretas”, não possibilitando uma interpretação total de seu funcionamento. A Interpretabilidade de Aprendizado de Máquina é uma área em crescente relevância por estes motivos. Neste trabalho, procuramos criar um classificador de discurso de ódio online em português, com foco em transfobia, a partir de mensagens da plataforma Twitter acerca da repercussão online de uma matéria jornalística da Rede Globo sobre a vida de mulheres presidiárias trans. Em seguida, usamos um método da área de Interpretabilidade de Aprendizado de Máquina para entender o funcionamento do classificador criado e, a partir dessas informações, analisar como os discursos transfóbicos se manifestaram nos *tweets* coletados. Com o classificador criado, os resultados alcançados mostram que a maioria dos termos mais importantes para a detecção de transfobia, no cenário estudado, são ofensivos e no gênero masculino, e muitos deles são usados para se referir a uma mulher trans, o que configura transfobia. Vimos também que a ocorrência destes termos em *tweets* tendem a influenciar o classificador a dar a resposta positiva (*tweet* transfóbico), enquanto termos que tendem a influenciá-lo para a resposta contrária (*tweet* não-transfóbico) enunciam uma variedade maior de sentimentos, tanto agressivos, quanto neutros e não-agressivos.

Palavras-chave: transfobia; discurso de ódio; aprendizado de máquina; interpretabilidade de aprendizado de máquina; SHAP.

ABSTRACT

The growing use of online social networks has greatly impacted people's lives and society. Communication between people, organization of events, groups and businesses, for example, have reached new heights thanks to these new online technologies. However, it is possible to notice that new and old challenges are also driven by these new communication dynamics. Hate speech reaches many more targets and with greater speed thanks to social networks and automation. Many platforms look for their own means of moderating their online spaces, but this is a task that is still far from over, if such an end is even achievable. Machine Learning is an area of Artificial Intelligence that seeks to create mathematical models for predicting values based on historical data. Several works seek to combat and study the spread of hate speech online using such models. However, many of these models are seen as “black boxes”, not allowing a full interpretation of their operation. Machine Learning Interpretability is an area of increasing relevance for these reasons. In this work, we seek to create an online hate speech classifier in Portuguese, focusing on transphobia, based on messages from the Twitter platform about the online repercussion of a journalistic article from Rede Globo about the lives of trans women prisoners. Then, we used a method from the Machine Learning Interpretability area to understand the functioning of the created classifier and, from this information, analyze how transphobic discourses were manifested in the collected tweets. With the created classifier, the obtained results show that the majority of the most important terms for transphobia detection, in the studied scenario, is offensive and in the masculine gender, and many of them are used to refer to a trans woman, which configures transphobia. We have also seen that the occurrence of these terms in tweets tend to influence the classifier to give a positive response (transphobic tweet), while terms that tend to influence it for the opposite response (non-transphobic tweet) enunciate a greater variety of feelings, both aggressive, as well as neutral and non-aggressive.

Keywords: transphobia; hate speech; machine learning; machine learning interpretability; SHAP.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE FÓRMULAS

SUMÁRIO

1 INTRODUÇÃO

O surgimento das redes sociais na internet na última década vem possibilitando formas de comunicação entre as pessoas nunca antes vivenciadas. Amigos, familiares e colegas de trabalho, por exemplo, podem trocar mensagens de texto e multimídia instantaneamente em canais de comunicação privados ou públicos, como o Whatsapp, o Facebook, o Twitter, entre outros.

Neste cenário, o número de usuários conectados à internet vem crescendo a cada ano. Em 2022, o número de usuários de internet de todo o mundo se aproxima da marca de 5 bilhões[1]. Destes, aproximadamente, 4.6 bilhões usam redes sociais[2]. Segundo a última pesquisa TIC Domicílios[3], realizada pelo Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br), 86% dos brasileiros utilizavam a Internet em 2021, isto é, aproximadamente 183,4 milhões de pessoas. Em relação ao uso de redes sociais, frequentemente o Brasil é listado entre os países com mais usuários no mundo[4][5].

São inegáveis as vantagens e os impactos do uso das redes sociais online na vida das pessoas e a nas novas dinâmicas da sociedade. Por exemplo, muitas comunidades vêm se organizando nestes espaços de forma orgânica, agregando pessoas com perfis e interesses em comum. As redes também oferecem possibilidades de organização e divulgação de eventos e negócios, além de permitir e facilitar o contato entre pessoas fisicamente distantes, esse fator sendo ainda mais evidenciado na pandemia de COVID-19.

Porém, embora as redes ofereçam todas essas novas possibilidades, alguns desafios vêm surgindo com o uso dessas novas tecnologias. Zeynep (2015), em seu trabalho, relata um experimento realizado pelo Facebook em 2012, que separou os usuários da rede em dois grupos e, sem o consentimento dos usuários, direcionou conteúdos negativos para um dos grupos e conteúdos positivos para o outro grupo. Ao final do experimento, notou-se que ambos os grupos tiveram humor de seus usuários alterado e os usuários de cada grupo passaram a publicar mais mensagens de acordo com o teor das mensagens que lhes eram direcionadas. Este exemplo ilustra um caso importante de uso inapropriado de dados gerados nas redes e que levanta sérias

---

[1] https://www.statista.com/topics/1145/internet-usage-worldwide/#dossierContents__outerWrapper, acessado em 05/09/2022

[2] https://www.insper.edu.br/noticias/mundo-se-aproxima-da-marca-de-5-bilhoes-de-usuarios-de-internet-63-da-populacao/, acessado em 05/09/2022

[3] https://cetic.br/pt/tics/domicilios/2021/individuos/C1/, acessado em 29/08/2022

[4] https://www.businessofapps.com/data/whatsapp-statistics/, acessado em 29/08/2022

[5] https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/, acessado em 29/08/2022

questões relacionadas à ética no uso dos dados pessoais em plataformas digitais e à manipulação intencional do humor dos usuários das redes.

A falta de ética no uso de dados de redes sociais também vem influenciando processos democráticos nos últimos anos, como processos eleitorais, por manipulação da visibilidade de conteúdos online, favorecendo lados específicos. Temos acompanhado casos, por exemplo, de coleta de dados que violam a privacidade de dados pessoais de usuários, como o escândalo da Cambridge Analytica e sua influência na eleição presidencial estadunidense em 2016 (ZEYNEP, 2015).

A utilização das redes sociais, contudo, ainda encontra muitas barreiras em sua regulamentação pelo mundo. Este cenário acaba favorecendo um ambiente descentralizado e pouco controlado, que permite a utilização antiética de dados de usuários, a viralização de fake news, a propagação de discursos de ódio, entre outros. Porém, recentemente muitos governos e organizações estão se preocupando com a questão do uso e tratamento dos dados pessoais por empresas. No Brasil, temos o caso da Lei Geral de Proteção dos Dados (LGPD)[6], que entrou em vigor em agosto de 2018.

A propagação de discursos de ódio, embora seja um problema que já existia antes do surgimento das redes sociais, toma outra proporção neste ambiente virtual, onde o compartilhamento de mensagens pode alcançar um comportamento viral, de alta taxa de disseminação (MELO et al., 2019). Um exemplo de como notícias falsas e discursos de ódio podem ser impulsionados através das redes sociais foram as mensagens difamando a vereadora do Rio de Janeiro, Marielle Franco. Logo após o seu assassinato, em março de 2018, uma série de fake news circularam nas redes sociais replicando o boato de que Marielle foi casada com um traficante e tinha associação com o crime[7] (RUBACK e OLIVEIRA, 2018). A publicação original acabou sendo removida, mas a informação falsa já havia circulado de forma massiva nas redes sociais.

Os discursos de ódio representam atos de violência simbólica que almejam inferiorizar determinados grupos e podem, além dos danos emocionais causados, instigar a violência física contra seus alvos. Um dos grupos alvos destes tipos de discurso de ódio é a população LGBTQ+ (Lésbicas, Gays, Bissexuais, Travestis, Transexuais, Transgêneros, Queers e demais)[8].

A população trans, no Brasil, é um dos grupos que mais sofre violências de diversas formas (BENEVIDES, 2022). No Brasil, segundo um estudo de 2021 da Escola de Medicina de Botucatu da Universidade do Estado de São Paulo (UNESP), 2% da

[6] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm, acessado em 06/09/2022

[7] https://www.aosfatos.org/noticias/nao-marielle-nao-foi-casada-com-marcinho-vp-nao-engravidou-ao-16-e-nao-foi-eleita-pelo-comando-vermelho/

[8] http://www.nohomophobes.com/#!/today/, acessado em 03/08/2022. Site mostra quantidades diárias de termos homofóbicos encontrados no Twitter.

população é composta por pessoas trans ou não-binárias[9]. O dossiê de 2021 da Agência Nacional de Travestis e Transexuais (ANTRA) relata a alta incidência de violência física, incluindo assassinatos, de pessoas trans ao longo dos anos (BENEVIDES, 2022). A população trans também enfrenta outras formas de violência, como a falta de reconhecimento, acolhimento e até mesmo tentativas de normalização de identidades trans em escolas, acarretando em fracassos ou abandono/evasão escolares de alunas e alunos trans (SALVADOR et al., 2021). Nosso trabalho busca contribuir, sob o ponto de vista da computação, no combate ao discurso de ódio online como um todo, mas especificamente trazemos um estudo de caso sobre a transfobia disseminada no Twitter.

1.1 CENÁRIO

Nosso trabalho investiga a detecção e análise de discurso de ódio disseminado online. Como estudo de caso, analisamos um caso específico de transfobia, disseminada na plataforma Twitter, acerca da repercussão da entrevista do médico Dráuzio Varella com a detenta Suzy Oliveira (uma mulher trans), em uma matéria sobre a realidade trans nas penitenciárias, onde várias detentas trans foram entrevistadas[10]. A matéria foi ao ar em 1° de Março de 2020 no programa Fantástico, da TV Globo.

Escolhemos o caso da Suzy, em particular, pois este repercutiu bastante nas redes na semana seguinte ao programa. Pelo seu relato ao entrevistador, onde ela conta que não recebia visitas há oito anos e compartilha a realidade de mulheres trans no cárcere, muitas pessoas mostraram solidariedade à ela através de cartas e outros presentes[11]. Porém, após difundirem nas redes o crime cometido por Suzy, se iniciaram, em grande quantidade, ofensas e ataques direcionados à ela e ao entrevistador. Muitas dessas mensagens continham discursos transfóbicos e outros tipos de discursos de ódio, às vezes de forma sutil e outras de forma explícita.

Para este estudo, escolhemos a plataforma do Twitter como fonte de dados por propiciar maior facilidade na coleta das mensagens de perfis públicos, além de se tratar de uma rede social com foco em textos. Com o grande pico de repercussão do caso nas redes, juntou-se um volume considerável de mensagens comentando o assunto, o que

---

[9] https://agenciabrasil.ebc.com.br/en/saude/noticia/2021-11/trans-non-binary-make-2-brazilians-study-shows, acessado em 29/08/2022

[10] https://g1.globo.com/fantastico/noticia/2020/03/01/mulheres-trans-presas-enfrentam-preconceito-abandono-e-violencia.ghtml, acessado em 14/09/2022

[11] https://g1.globo.com/sp/sao-paulo/noticia/2020/03/07/detenta-trans-suzy-ja-recebeu-234-cartas-apos-reportagem-do-fantastico-diz-secretaria-de-sp.ghtml, acessado em 14/09/2022

caracterizou uma boa oportunidade de análise desses dados textuais e os discursos transfóbicos contidos neles.

1.2 OBJETIVOS GERAIS

Neste trabalho, criamos um classificador de transfobia, dentre outros discursos de ódio e categorias pertinentes, e usamos um método de interpretabilidade de modelos de aprendizado de máquina para investigar quais palavras ou termos são as que mais impactam a decisão deste classificador.

O modelo classificador foi treinado com dados referentes ao cenário específico da Suzy e, portanto, não é recomendado o seu uso em diferentes contextos. Porém, o levantamento de palavras mais importantes para este classificador nos traz mais informações para uma análise do discurso transfóbico do cenário em questão, além de contribuir com um maior entendimento do modelo de aprendizado de máquina treinado.

Podemos definir como objetivo geral do nosso trabalho: investigar métodos computacionais de detecção e análise de discurso de ódio online, em particular a detecção de transfobia no Twitter.

1.3 OBJETIVOS ESPECÍFICOS

1. Criar uma base de dados anotada manualmente por voluntários, que pode ser utilizada como base de treinamento para classificadores automáticos de transfobia online
2. Gerar modelos de classificação de transfobia utilizando Multinomial Naive Bayes, Support Vector Machine e NBSVM (uma combinação dos dois primeiros)
3. Avaliar o desempenho dos classificadores utilizando métricas de avaliação, como o *F1-score*.
4. Explorar a interpretabilidade de modelos de aprendizado de máquina para analisar os termos mais importantes para a classificação, através do método SHAP (SHapley Additive exPlanations) (LUNDBERG e LEE, 2017).

1.4 ESTRUTURA DO TEXTO

Na Seção 2, apresentamos o referencial teórico relacionado ao nosso trabalho, com conceitos necessários ao entendimento da nossa proposta: discurso de ódio, transfobia, racismo, aprendizado de máquina e interpretabilidade na Inteligência Artificial.

Na Seção 3, mostramos trabalhos relacionados às áreas de aprendizado de máquina para detectar discurso de ódio online e também mais aplicações do método SHAP em trabalhos acadêmicos da mesma área.

Na Seção 4 apresentamos o método utilizado neste trabalho para a criação do detector de discurso de ódio e a análise dos termos que mais o influenciam, desde a coleta dos dados até a criação dos classificadores e cálculo dos valores SHAP.

Na seção 5 detalhamos os experimentos com cada classificador e mostramos os resultados obtidos, bem como os resultados e interpretações que a aplicação de SHAP nos permitiu.

Por fim, na Seção 6 fazemos nossas considerações finais. Na Seção 7 elucidamos trabalhos futuros e, por fim, na Seção 8, listamos nossas referências.

2 REFERENCIAL TEÓRICO

Nesta seção, começamos definindo alguns conceitos relacionados às ciências humanas e sociais que são essenciais à compreensão da nossa proposta: Discurso de ódio, Transfobia e Racismo. Em seguida, apresentamos alguns conceitos computacionais relacionados ao aprendizado de máquina: o processo de aprendizagem, os modelos de classificação, as métricas para avaliação e conceitos relacionados à interpretabilidade em Inteligência Artificial.

2.1 DISCURSO DE ÓDIO E TRANSFOBIA

Neste trabalho, nos apoiaremos em definições, tanto de transfobia quanto de racismo, de teóricos brasileiros considerados referências nas áreas. Para a definição de transfobia, consideramos os trabalhos da teórica e mulher trans Jaqueline Gomes de Jesus, que define transfobia como: “Preconceito e/ou discriminação em função da identidade de gênero de pessoas transexuais ou travestis.” (JESUS, 2012). No caso de textos online, são exemplos quaisquer ataques, ameaças ou desrespeito ao corpo trans,

como por exemplo: negar a identidade de gênero de uma pessoa (chamar, tratar ou se referir a uma mulher trans com pronomes direcionados a homens, e o contrário para homens trans).

Já para o racismo, utilizamos a definição do filósofo e intelectual brasileiro, especialista sobre a questão racial no Brasil, Silvio de Almeida. Silvio define racismo como “uma forma sistemática de discriminação que tem a raça como fundamento, e que se manifesta por meio de práticas conscientes ou inconscientes que culminam em desvantagens ou privilégios para indivíduos, a depender do grupo racial ao qual pertençam.” (ALMEIDA, 2018).

Para discursos de ódio, de forma geral, podemos citar a definição de Paula Fortuna e Sérgio Nunes em sua revisão sistemática da literatura acerca do tema de detecção automática de discurso de ódio:

> Discurso de ódio é linguagem que ataca ou diminui, que incita violência ou ódio contra grupos, baseado em características específicas como aparência física, religião, descendência, origem nacional ou étnica, orientação sexual, identidade de gênero ou outras, e pode ocorrer em diferentes estilos linguísticos, até mesmo de formas sutis ou quando humor é utilizado.(FORTUNA e NUNES, 2018. Tradução nossa)

Portanto, discursos de ódio podem se manifestar de outras formas, além da transfobia e do racismo como, por exemplo, sexismo, preconceitos de classe, intolerância religiosa, xenofobia, entre outros. Estes, porém, não são o foco do nosso estudo de caso, mas também são formas de discursos de ódio que devem ser combatidas.

Muitas vezes, podemos encontrar no discurso uma linguagem ofensiva, mas que não se utiliza de discriminação. Neste trabalho, nos referimos a estes casos como “Linguagem puramente ofensiva”.

É importante mencionar uma particularidade importante neste contexto: Muitas pessoas usam termos que podem ser considerados ofensivos de forma geral, mas que se são manifestados por alguém da própria comunidade que seria atacada, não seriam ofensivos. Como por exemplo, alguns perfis de drag queens que utilizam linguagem comumente vista como mais agressiva e até podendo conter xingamentos de origem LGBTQfóbica, mas que são utilizados pela comunidade, muitas vezes num tom de protesto ou de ressignificação dos termos[12][13].

---

[12] https://tab.uol.com.br/noticias/redacao/2020/03/06/bom-dia-gay-guerra-ao-discurso-de-odio-nas-redes-derruba-posts-lgbtq.htm, acessado em 07/08/2022

[13] https://revistaforum.com.br/politica/2022/7/28/pre-candidata-do-pt-tem-conta-do-twitter-bloqueada-apos-foto-com-l-de-lula-termo-sapato-120857.html, acessado em 29/08/2022

2.2 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina é um subcampo da Inteligência Artificial que busca criar modelos matemáticos de previsão de valores ou tomadas de decisão, baseados em dados sobre um fenômeno de interesse. Os dados coletados sobre o fenômeno são alimentados ao modelo como exemplos para a extração de padrões que o ajudarão a retornar respostas com o menor erro possível (RUBACK et al. 2022). Existem diversos algoritmos de aprendizado para criar tais modelos, cada um deles com seus próprios hiperparâmetros, parâmetros externos aos dados do fenômeno em estudo, que controlam o comportamento do algoritmo na geração dos modelos.

Dentro da área de Aprendizado de Máquina, existem diversas vertentes. Em nosso trabalho, utilizamos a chamada de Aprendizado Supervisionado. Nesta vertente, os dados de exemplo para o treino da máquina são acompanhados da classe à qual o exemplo pertence, muitas vezes chamada de “anotação” (em inglês “annotation”). A anotação funciona como um “gabarito” que a máquina usa de referência tanto no processo de treinamento (aprendizado) quanto no de teste (avaliação). Ou seja, é esperado do treinamento da máquina um modelo que seja capaz de responder valores correspondentes às anotações fornecidas para cada instância do dataset (BURKOV, 2019). Tarefas comuns de aprendizado supervisionado são a previsão de dados futuros (como a previsão do tempo), categorização de textos (a qual nosso trabalho se enquadra), entre outras.

O Aprendizado Supervisionado pode ser usado para valores categóricos (ou discretos), onde tentamos prever valores categóricos (ex. decisão entre gatos ou cachorros presentes em imagens) e dizemos que temos uma tarefa de classificação. Ele também pode ser usado para valores reais (ou contínuos), onde prevemos valores contínuos (ex. previsão de flutuações na bolsa de valores), onde denomina-se uma tarefa de regressão. Neste trabalho, propomos um classificador de categorias de discurso de ódio, como “Transfobia” e “Racismo”, portanto, estamos lidando com uma tarefa de classificação.

A figura 1 mostra uma visão geral das etapas do aprendizado de máquina supervisionado. A primeira etapa é a de coleta dos dados (ou exemplos) que o modelo utiliza para “aprender”. A próxima etapa é a anotação, na qual é feita a classificação manual dos exemplos e serve para dizermos ao modelo qual o resultado esperado para cada instância do dataset, ou, em outras palavras, estamos fornecendo o gabarito dos exemplos para que o modelo possa aprender e extrair os padrões que ele utilizará para classificar os novos exemplos.

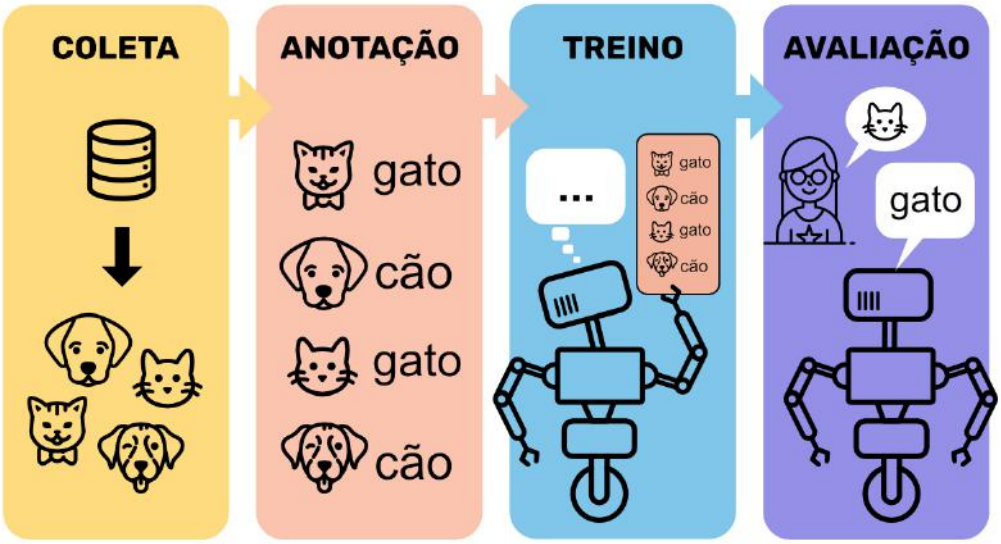

Figura 1: Etapas do processo de aprendizado de máquina.

Após a anotação, chegamos na etapa de treino, onde o modelo irá aprender com os dados coletados e, de acordo com o algoritmo de aprendizado escolhido, será capaz de prever a classe de novos exemplos. É nesta etapa que o modelo procura extrair os padrões dos dados de forma a auxiliar na tomada de decisão para as suas respostas. Durante o treino, quanto mais dados disponíveis, mais exemplos a máquina tem para detectar padrões e, portanto, melhor tendem a ser os resultados.

Por fim, temos a etapa de avaliação/teste, onde o modelo criado irá classificar dados não utilizados no treinamento e os resultados retornados são avaliados ao serem comparados com as categorias dadas pela etapa de anotação, que serve como uma espécie de gabarito para checar se as respostas da máquina estão corretas.

A etapa de avaliação é importante para entendermos o desempenho do modelo treinado para novos dados e garantir que ele tenha boa generalização. Caso contrário, é possível que um modelo aprenda a classificar as instâncias de treino muito bem, mas falhe em muitos dos dados de teste. Quando isto ocorre, o modelo apresenta baixa capacidade de generalização e dizemos que ela sofreu overfitting. É como se o modelo estivesse “decorando” as respostas certas para os dados de treino que ela tem acesso, se tornando inflexível para dados não vistos no treino e mostrando-se incapaz de detectar os padrões dos dados quando considerada uma gama/escopo maior de instâncias.

É imprescindível a separação de parte dos dados de exemplo para treino e outra para teste/avaliação. No geral, o modelo deve aprender com o primeiro conjunto e ser testado com um outro conjunto. E, quanto maior for o conjunto de treino, a tendência é que melhor seja o desempenho do modelo. Muitas vezes essa necessidade representa um desafio em tarefas de classificação, onde não há muitos dados coletados para o treino

e parte deles ainda será utilizada somente para testes. Uma das técnicas que podem ser usadas para mitigar tal situação, quando não temos muitos dados para teste e treino, é o uso de k-fold cross-validation (validação cruzada, em português) (BISHOP e NASRABADI, 2006).

A técnica K-fold cross-validation consiste em dividir o conjunto de dados em K partes de igual tamanho, chamadas de folds, separar uma delas para o teste do modelo e o resto para o treinamento. Este procedimento é repetido, em diversas rodadas, para todas as combinações possíveis de subconjuntos (folds), de forma que todos os K folds sejam utilizados como grupo de teste uma vez. As métricas de desempenho são calculadas para todas as rodadas e tem-se a média delas como a medida final de desempenho do modelo. Dessa forma, com todos os folds sendo usados para teste do modelo em alguma rodada, garante-se o uso do conjunto inteiro de dados disponíveis para teste.

A seguir mostramos o funcionamento dos algoritmos de aprendizado de máquina e as métricas de avaliação utilizadas neste trabalho.

2.2.1 Classificador Multinomial Naive Bayes (MNB)

De acordo com Manning et al. (2008), o cerne do funcionamento do Multinomial Naive Bayes (MNB) é o Teorema de Bayes. O algoritmo acha a maior probabilidade condicional (ou likelihood, para casos discretos) P(c|x) de uma classe c ∈ C ser a classe certa para uma instância que tem as features x ∈ X, sendo C o conjunto de todas as classes e X o conjunto de todas as features de cada instância em um dataset. Esta probabilidade é calculada através do Teorema de Bayes, conforme descreve a equação 1.

$$P(c|\vec{x}) = \frac{P(\vec{x}|c)P(c)}{P(\vec{x})}$$

Equação 1: Teorema de Bayes

P(**c**) sendo a probabilidade a priori da instância em questão ser da classe **c** e P(**x**) um vetor com probabilidades a priori de cada *feature* de **x** estar presente em uma instância. Porém, como P(**x**) é constante para todas as classes, podemos desconsiderar este denominador, ficando com uma fórmula de proporcionalidade, como descrita na fórmula 1:

$$P(c|\vec{x}) \propto P(\vec{x}|c)P(c)$$

Fórmula 1: Proporcionalidade a partir do Teorema de Bayes ao assumir P(**x**) constante

Se assumirmos independência entre as *features* (o que nem sempre é possível, nos evidenciando uma das motivações do nome do algoritmo ser *Naive*), podemos substituir a probabilidade P(**x**|c) pelo produto das probabilidades condicionais de suas *features* x_i, com 0 ≤ i < k, sendo k o número de *features*, como descrito na fórmula 2.

$$P(c|\vec{x}) \propto \prod_{i=1}^{k} P(x_i|c)P(c)$$

Fórmula 2: Proporcionalidade a partir do Teorema de Bayes ao assumir P(**x**) constante e independência das *features* de **x**

A classe de resposta do modelo será a classe de maior probabilidade c_{mp}, como na equação 2 a seguir.

$$c_{mp} = \arg\max_{c \in C} \prod_{i=1}^{k} P(x_i|c)P(c)$$

Equação 2: Classe de resposta do modelo MNB

Porém, com esta abordagem, um problema surge quando pelo menos uma das probabilidades de *features* dada uma classe é zero. Isto zeraria todo o produtório e a multiplicação por P(**c**), independente de quão altos sejam os outros valores. Para resolver este problema, introduz-se um fator α (alfa) que soma uma constante nas contagens de probabilidade de cada *feature*, evitando valores zerados.

Um último ajuste a ser feito existe por motivos computacionais. Como o cálculo das probabilidades de cada classe envolve o produto de diversas probabilidades condicionais, existe a possibilidade de underflow de ponto flutuante. Para resolver este problema, passa-se então a calcular o log destas probabilidades de classe, trocando o problema de multiplicação para um de soma, como descrito na equação 3. A classe de maior probabilidade em log ainda será a classe mais provável.

$$c_{mp} = \arg\max_{c \in C} [logP(c) + \sum_{i=1}^{k} logP(x_i|c)]$$

Equação 3: Classe de resposta do modelo MNB adaptada para log

2.2.2 Classificador Support Vector Machine (SVM)

O algoritmo SVM (Support Vector Machine) procura achar o hiperplano de N dimensões, sendo N o número de features, que melhor divide os pontos da base de dados em regiões de suas respectivas classes. Novos pontos a serem classificados serão atribuídos a classe pertencente a região em que se encontrarem. Existem infinitos hiperplanos que separam dois ou mais conjuntos de pontos, mas apenas um deles é o melhor divisor que maximiza a distância entre tais conjuntos (BISHOP e NASRABADI, 2006).

Para achar o melhor hiperplano, o algoritmo se baseia na posição dos pontos nas extremidades de cada subconjunto, chamados de vetores de suporte, que vão, a cada iteração, influenciando a posição e orientação deste hiperplano. A figura 2 ilustra os diversos hiperplanos possíveis separando duas classes e a escolha do melhor.

Na figura 2, no gráfico da esquerda, podemos ver que podemos posicionar um hiperplano de separação das classes de cães e gatos de infinitas formas diferentes. Porém só uma das posições será a melhor. O gráfico da direita na figura 2 mostra o hiperplano que melhor separa cães de gatos, permitindo um hiperplano correto e que permite maior margem entre si e os primeiros exemplos de cada classe.

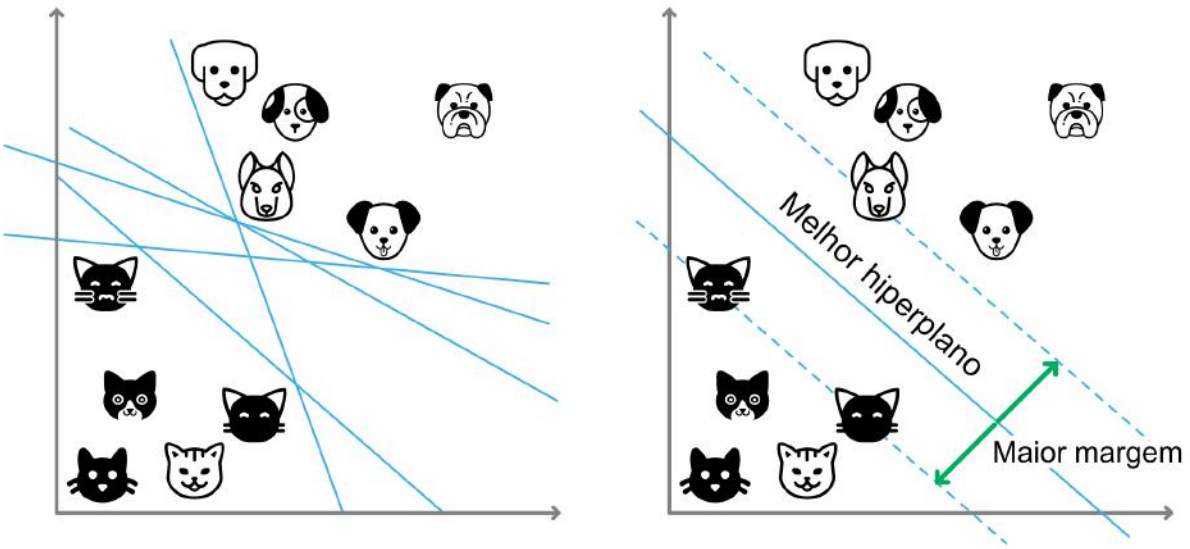

Figura 2: Na esquerda, possíveis hiperplanos separam os dados das classes cão e gato. Na direita temos o hiperplano que melhor divide tais dados, com a maior margem possível entre os vetores de suporte (pontos nas linhas de margem).

O treino de um modelo SVM é um método iterativo que começa com um dos possíveis hiperplanos de separação e, a cada iteração, seus parâmetros são ajustados

para reposicionar o hiperplano de separação das classes, baseado em seus erros de classificação.

O algoritmo SVM conta com um hiperparâmetro C, que pode ser interpretado como uma intolerância a erros durante o treino. Quanto maior o valor de C, mais o algoritmo é penalizado por erros e, portanto, é forçado a estreitar mais a margem entre os subconjuntos. A permissividade a erros pode ser desejável para evitar um possível overfitting do modelo, pois um modelo completamente intolerante a erros pode ter um posicionamento do hiperplano de separação das classes com margens muito restritas, de forma que novos pontos a serem classificados que aparecerem perto do hiperplano, têm mais chance de serem classificados erroneamente.

A figura 3 ilustra um exemplo desta situação, onde um outlier de uma das classes faz com que a margem entre vetores de suporte seja muito pequena quando não há tolerância a erros durante o treino da máquina. Por consequência disto, novos dados sendo classificados têm maior chance de atribuição de classes erradas. Uma maior tolerância a erros de treinamento pode acarretar em uma melhor generalização do modelo e, portanto, uma maior taxa de acertos no total.

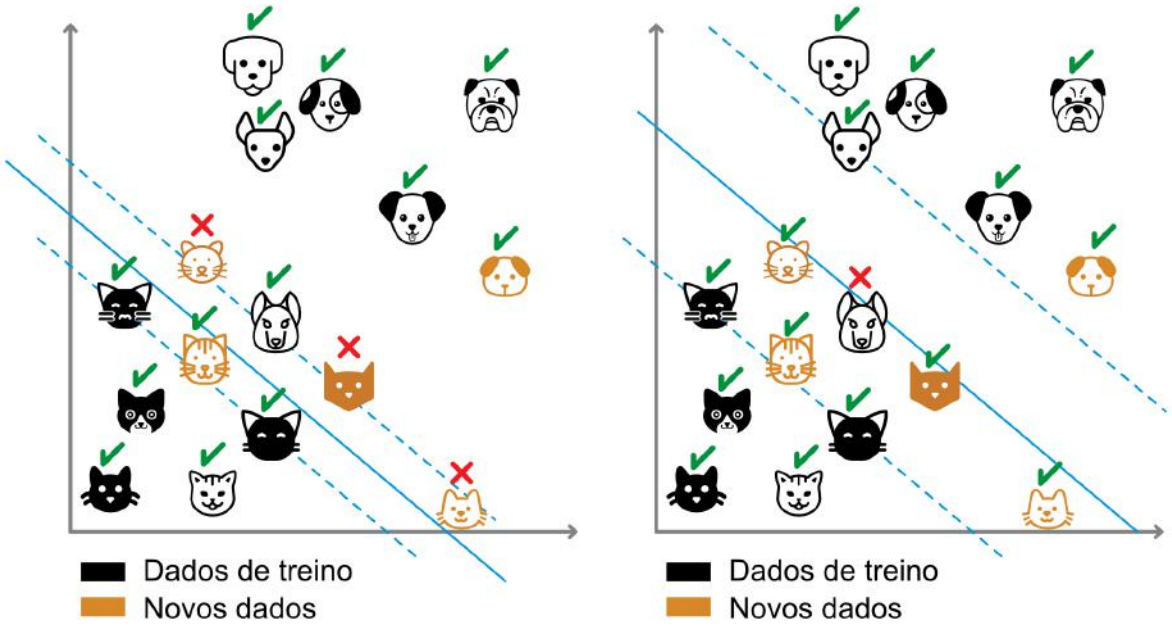

Figura 3: Exemplos de como uma maior permissividade a erros durante o treino pode gerar melhor desempenho no geral, quando novos dados não vistos durante o treino entram para classificação.

O algoritmo do SVM acha um separador linear dos dados. Porém, para casos onde os dados não são linearmente separáveis, o algoritmo ainda pode contar com uma manipulação conhecida como Kernel Trick. Neste método, o algoritmo trabalha com os dados transformados para maiores dimensões, com a esperança de existência de uma separação linear dos dados em maiores dimensões.

2.2.3 Classificador NBSVM (combinação de MNB com SVM)

Wang e Manning (2012), em seu trabalho, destacam que modelos MNB têm boa performance quando lidam com textos curtos, já os modelos SVM mostram melhor desempenho em textos maiores. O modelo proposto por eles é uma combinação dos dois: Um modelo SVM que usa como features, os logs das probabilidades de cada termo nos textos, como calculado no MNB. Este modelo recebeu o nome de NBSVM.

Os autores, ao testar seu modelo proposto, notaram que, novamente, a performance para textos longos era muito boa, mas para aprimorar os resultados para textos mais curtos, fizeram mais um ajuste em seu modelo. Eles incorporaram ao modelo uma interpolação com a variável β, de 0 a 1, entre os resultados fornecidos pelo NBSVM definido no parágrafo anterior e um MNB tradicional. Esta nova abordagem mostrou conseguir bons resultados para diversos tamanhos de texto e os autores sugeriram o uso desta versão do modelo como uma nova baseline para projetos de aprendizado de máquina supervisionado.

2.3 MÉTRICAS DE AVALIAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA

Existem diversas métricas para avaliar o desempenho de classificadores. As que utilizamos neste trabalho são: *accuracy*, *precision*, *recall* e F_1*-score* (BURKOV, 2019).

Accuracy, muitas vezes em português chamada de “acurácia” ou “assertividade”, para não gerar confusão com a tradução da outra métrica *precision*, é uma métrica que nos diz a taxa de acertos do classificador. É definida pela taxa de verdadeiros positivos (VP) + verdadeiros negativos (VN) sobre a quantidade total de instâncias avaliadas, que é a soma VP + VN + falsos positivos (FP) + falsos negativos (FN), como na equação 4.

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

Equação 4: Cálculo da métrica *accuracy*

Precision, ou precisão, de um classificador é a taxa de verdadeiros positivos dentre todos as instâncias classificadas como positivas. Ou seja, de todas as instâncias que o

classificador disse serem positivas (pertencentes à categoria em questão), quais realmente são. Sua definição é dada pela equação 5:

$$Precision = \frac{VP}{VP + FP}$$

Equação 5: Cálculo da métrica *precision*

Já *recall*, ou revocação, é a taxa de verdadeiros positivos dentre todas as instâncias de fato positivas. Essa métrica nos diz, de todos os casos em que o classificador deveria dizer que são positivos, quais ele realmente disse que são. E é dada pela equação 6:

$$Recall = \frac{VP}{VP + FN}$$

Equação 6: Cálculo da métrica *recall*

A métrica F_β*-score* é uma forma de unir os valores de *precision* e *recall* em um único valor, o parâmetro β, um valor positivo real, serve para ponderar as influências das duas métricas, dando mais importância para *recall* do que *precision*. Essa métrica é definida pela equação 7:

$$F_\beta = \frac{1 + \beta^2}{\frac{\beta^2}{recall} + \frac{1}{precision}}$$

Equação 7: Cálculo da métrica F_β*-score*

Quando se dá igual importância para *recall* e *precision*, com $\beta = 1$, temos a métrica F_1*-score*, que é uma média harmônica das duas grandezas. Essa é a forma utilizada neste trabalho e é definida na equação 8:

$$F_1 = 2\frac{precision \cdot recall}{precision + recal}$$

Equação 8: Cálculo da métrica F_1*-score*

2.4 INTERPRETABILIDADE DE MODELOS DE APRENDIZADO DE MÁQUINA

Modelos de aprendizado de máquina muitas vezes não são fáceis de ser interpretados, principalmente modelos mais complexos como os baseados em redes neurais profundas (LUNDBERG e LEE, 2017). Muitas vezes é impossível se interpretar as suas predições de alguns modelos, chamados de caixas-pretas. E até mesmo modelos que proporcionam entendimento mais fácil de seu funcionamento, como os baseados em árvores de decisão, podem precisar de mais explicações para se obter maior compreensão do porque certos modelos tomam suas decisões (MENG et al., 2020).

A capacidade de interpretar como os modelos estão realizando escolhas é importante, já que estes podem ser utilizados nas mais diversas aplicações com impactos de diferentes graus em nossas vidas. Mesmo que modelos mais complexos usados como “caixa preta” possam oferecer melhor desempenho, para muitas tarefas cruciais (como as da área de saúde, por exemplo), em alguns casos prioriza-se a utilização de métodos de menor desempenho, mas que propiciam uma melhor interpretação do modelo e suas decisões.

Por esses motivos, a área de Interpretabilidade de Modelos de Aprendizado de Máquina vem ganhando relevância nos últimos anos. Essa área busca alcançar meios de explicar o funcionamento por trás de modelos de aprendizado de máquina de todas as origens, o que proporciona maior confiança no uso dos mesmos. Um dos métodos dessa área é o que exploramos neste trabalho, o método *SHapley Additive exPlanations* (SHAP) (LUNDBERG e LEE, 2017).

Os valores SHAP foram derivados do cálculo de valores Shapley (primeira parte da sigla), da área de teoria de jogos cooperativos, onde se almejava medir a contribuição de cada jogador pertencente a um time no resultado de um jogo ou partida. O cálculo desse valor é feito para cada jogador e também leva em consideração que a contribuição deles pode variar de acordo com os membros participantes do time, pois, por exemplo, dois jogadores que têm maior sinergia em conjunto podem ter melhor êxito em uma partida que jogaram juntos do que em times separados, com outras pessoas de menos afinidade.

O cálculo de valores Shapley para um jogador j**,** leva este tipo de relação em conta ao medir a contribuição de j em todos os times possíveis que podem ser formados a partir de um conjunto de todos os demais jogadores (todos os subgrupos possíveis de jogadores). É calculada a diferença entre os resultados da partida no cenário em que j faz parte do time e no cenário complementar, onde o time não conta com a participação de j e, por fim, tira-se a média de todas essas contribuições. Em outras palavras, é calculada a contribuição marginal do jogador j.

A figura 4 ilustra, com um exemplo, o cálculo do valor Shapley para um jogador em um jogo como apresentado no parágrafo acima. A coluna da esquerda representa todas as possibilidades de formações de grupos sem o jogador em questão (jogador azul) e os

resultados de um jogo com tal grupo de participantes. A coluna do meio, representa o mesmo, porém com grupos que contêm o jogador azul. A coluna da direita calcula a contribuição do jogador azul para cada par de cenários, que será a diferença dos resultados dos jogos quando jogados pelos grupos com e sem o jogador em questão. O valor Shapley do jogador azul será a média de suas contribuições em todos os cenários.

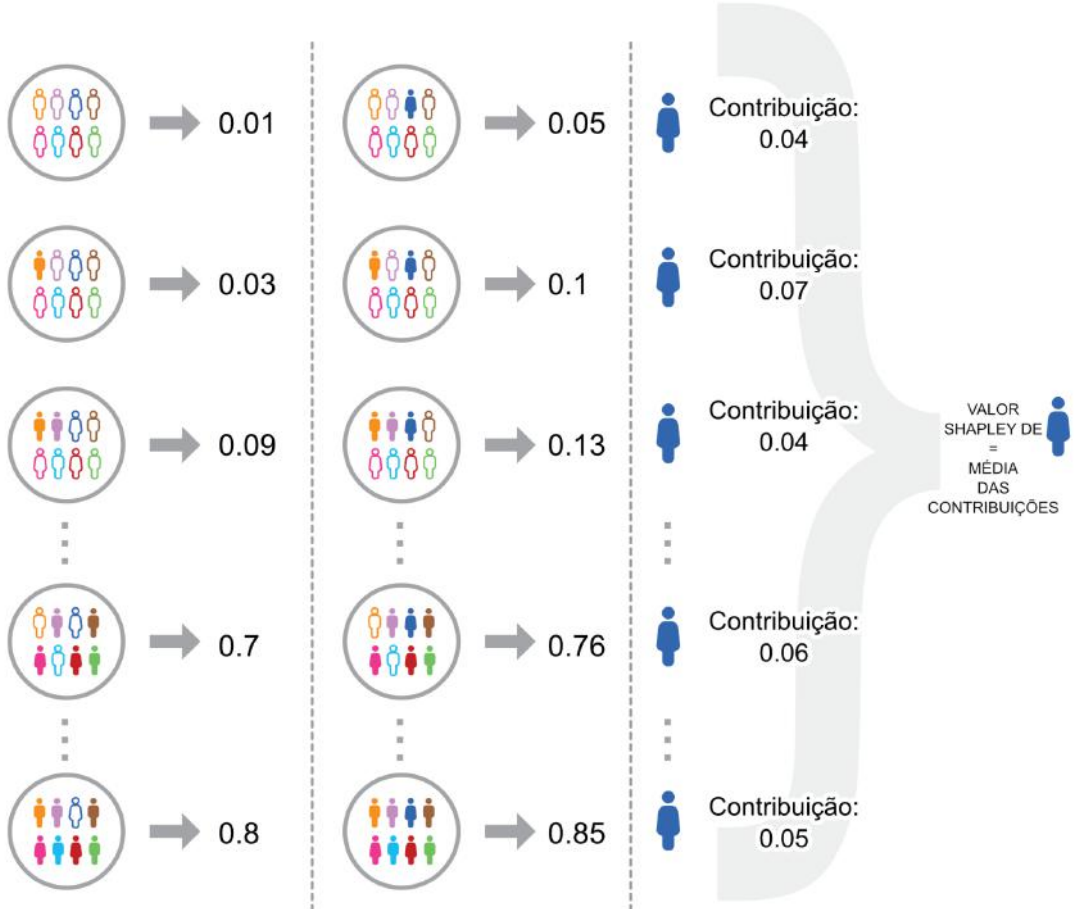

Figura 4: Exemplo de cálculo de valor Shapley para um jogador.

Um fator interessante dos valores Shapley é que eles têm uma propriedade aditiva. Os valores Shapley dos jogadores, quando somados, são iguais ao resultado da partida.

Lundberg e Lee (2017), então, propõem os valores SHAP como uma métrica unificada de importância de *features* na área de aprendizado de máquina. Os valores SHAP são os valores Shapley de uma função de esperança condicional de um modelo de aprendizado de máquina f com *input* de um vetor de *features* binarizado **z**, onde, sem as informações de nenhuma *feature* tem-se simplesmente o valor E[f(**z**)], e conforme as *features* vão sendo adicionadas, suas contribuições (valores Shapley) alteram essa esperança ao colocarem informações a priori, até chegar em E[f(**z**) | **z** = **x**] = f(**x**), o *output* do modelo em questão para uma instância **x**, que é aproximado pela função de esperança do método SHAP quando se tem a influência de todas as *features* disponíveis.

Apenas mais um ajuste precisa ser feito para a adaptação de valores Shapley para o aprendizado de máquina. O cálculo desses valores, como visto anteriormente, precisa

calcular a contribuição de uma *feature* em cenários quando há ausência de outras *features*, porém, não são todos os modelos de aprendizado de máquina que aceitam vetores de *input* de tamanhos diferentes. Para esses casos, uma das soluções elaboradas pelos autores do método foi a substituição dos valores das *features* ausentes pela média dos valores dessa *feature* em todo o *dataset*.

O método SHAP para aprendizado de máquina proporciona uma avaliação local, ou seja, feita para cada instância de dados de forma individual, mas é possível alcançar um entendimento global do modelo ao analisar todas as contribuições das *features* em todos os dados de forma agregada, como faremos neste trabalho.

3 TRABALHOS RELACIONADOS

As publicações de trabalhos sobre discurso de ódio na área de Ciência da Computação e Engenharia eram escassas antes de 2014, mas desde então vêm aumentando em número consideravelmente (FORTUNA e NUNES, 2017). Um exemplo inclui o trabalho de Gambäck e Sikdar (2017) que, para criar um modelo detector de racismo ou sexismo, utilizaram diferentes métodos com modelos de Rede Neural Convolucional, um algoritmo de Rede Neural Artificial (RNA) que é muito utilizado para detecção de padrões em imagens, mas que apresenta robustos resultados também em outras tarefas. Eles obtiveram bons resultados em comparação com trabalhos anteriores no mesmo conjunto de dados, mas ressaltaram que os sistemas construídos não tiveram bom desempenho nas classes de racismo e “ambas” (racismo e sexismo juntos), pois houve falta de exemplos suficientes para o treinamento da máquina nestas classes.

Exemplos de trabalhos de detecção e discurso de ódio em português são os de Fortuna e Nunes (2017) e Camelo (2018). Fortuna realizou e publicou uma revisão sistemática da literatura em detecção automática de discurso de ódio, além disso montou uma base de dados anotada de tweets em português de Portugal e do Brasil e investigou modelos que consideram uma ampla gama de discursos de ódio. Já Camelo, utilizou diversos modelos de aprendizado de máquina fora da família das RNAs em uma tarefa de detecção que focou especificamente em discursos de ódio LGBTfóbicos na sessão de comentários de matérias sobre a população LGBT do G1[14]. Além disso, os participantes do processo de anotação também foram dessa população.

Um trabalho que analisou o resultado de um modelo criado a partir de dados anotados por diversas pessoas, sem restrições de participantes, é o de Davidson et al.

[14] https://g1.globo.com/, acessado em 14/09/2022

(2017), que levanta cuidados necessários neste processo de anotação humana dos dados de treino da máquina. O trabalho procurou lidar com a dificuldade de modelos automáticos de diferenciar textos com discurso de ódio de textos ofensivos, mas sem ódio. Eles trabalharam com *tweets* e utilizaram três classes no trabalho: discurso de ódio, linguagem ofensiva (sem configurar ódio) e texto neutro. Analisando os casos em que a máquina fez classificações erradas, descobriram indícios de que os anotadores tendiam a ter dificuldades em identificar tweets com discurso de ódio na ausência de palavras fortes, palavrões ou em textos com ódio porém de forma sutil. Também mostram que seus participantes tendiam a considerar textos com termos racistas e homofóbicos como contendo discurso de ódio, porém não os com termos sexistas, que em geral eram classificados como apenas ofensivos, o que demonstra um viés humano que pode afetar o treinamento da máquina.

Quanto aos trabalhos que exploram a transparência e interpretabilidade de modelos de aprendizado de máquina com valores SHAP, existem pesquisas como a de Meng et al. (2020), onde tais valores são utilizados para gerar um entendimento maior de quais fatores podem fazer com que avaliações de produtos *online* tenham maior qualidade e, portanto, chances de serem relevantes a futuros clientes. O trabalho aplica o cálculo dos valores SHAP e, com eles, cria diversas visualizações em cima de algoritmos classificadores baseados em árvores de decisão.

Marcílio e Eler (2020), em seu trabalho, também exploraram o uso de SHAP para a tarefa de seleção de *features*. O método proposto pelos autores calcula a contribuição média de cada *feature* para um *dataset* em um classificador e seleciona as que possuem os maiores valores. Eles compararam a performance do método proposto contra três outros métodos comuns de seleção de *features*. Para as comparações, foi utilizado um modelo XGBoost, para classificação e regressão, em oito *datasets* públicos distintos e diferentes cenários, com variações na porcentagem de *features*, foram avaliados.

Os resultados mostram que a utilização de SHAP para a tarefa de seleção de *features*, quando comparado com as outras três alternativas, obtém os melhores resultados em todas as tarefas testadas, com exceção de regressão em um dos *datasets*. Porém, mesmo neste caso, a diferença de *Area Under Curve* (AUC) com o melhor modelo foi de apenas 0.02.

Nosso trabalho procura, com influência destas experiências, trabalhar na detecção de discursos de ódio (em um recorte específico, do tema da Suzy) com enfoque na transfobia em *tweets* em português. Tema que, no melhor do nosso conhecimento, ainda é pouco abordado na literatura. E após criação do classificador, avaliar suas tendências para tomadas de decisão. Como nossos dados são textuais e trabalharemos com *tokens*, como descrito na Seção 4, o método SHAP pode nos elucidar termos mais importantes

para as classificações, potencialmente enriquecendo possibilidades de análise de discursos em textos.

4 CLASSIFICADOR DE TRANSFOBIA

Neste trabalho propomos um método para criar um classificador de transfobia, que nos permite avaliar os termos mais importantes utilizados para a classificação. A Figura 5 mostra as etapas que utilizamos no método.

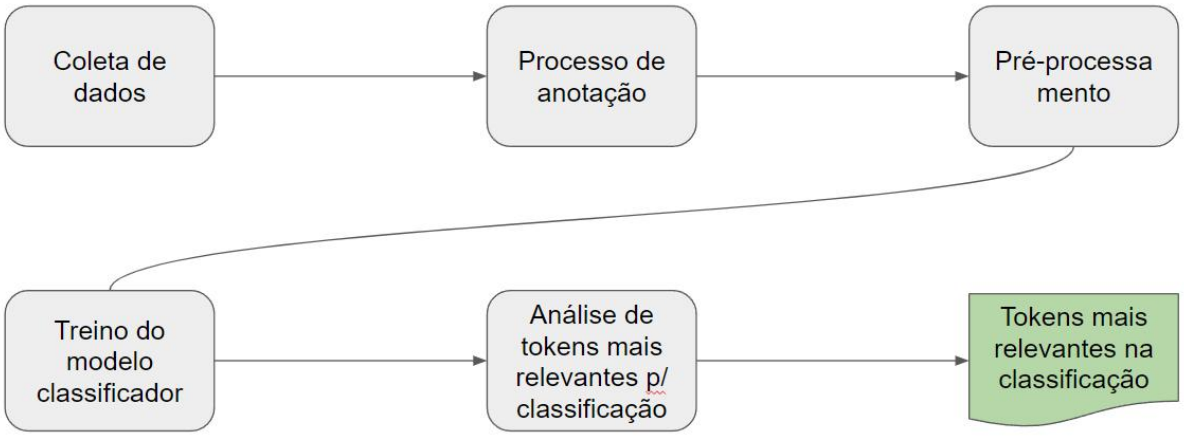

Figura 5: Etapas do método para detecção e análise de transfobia em *tweets*.

Começaremos explicando o processo de coleta de dados para a construção das bases de treino e de teste e, posteriormente, abordaremos o pré-processamento da base para adequação ao uso dos classificadores. Fechamos esta seção explicando o processo do treino e teste dos classificadores e a aplicação do método de valores SHAP para compreender o impacto dos termos na classificação.

4.1 A BASE DE DADOS

Nesta subseção demonstramos o processo de coleta dos *tweets* relacionados ao caso da Suzy Oliveira e sua entrevista com o Dr. Drauzio Varella no programa Fantástico, da Rede Globo. Também realizamos uma análise preliminar dos dados de acordo com a linha do tempo dos *tweets* coletados.

4.1.1 Coleta dos Dados

Para a coleta dos dados, utilizamos o módulo Python chamado Tweepy (versão 3.8)[15], que permite o acesso à API do Twitter. Vale lembrar que existem planos pagos e um plano gratuito para acesso desta API. O último permite apenas a busca por *tweets* publicados no máximo uma semana atrás, contando a partir do momento da requisição. Além disso, a coleta gratuita não retorna necessariamente todos os *tweets* na base de dados do Twitter, pois, nos termos da empresa, ela prioriza relevância no lugar de completude[16]. A documentação da API do Twitter detalha todos os métodos que podem ser acessados para a coleta de dados, assim como os parâmetros de cada método[17].

Em nosso trabalho, a busca por *tweets* na API do Twitter foi feita no período de 2 a 20 de Março de 2020. Foi criada uma *string* de busca que procura por mensagens públicas que continham ao menos uma das seguintes *hashtags*: #Susy, #SuzyEstuprador, #SuzyLivre, #drauziovarella e #DrauzioVarellaLixo. *Hashtags* essas, selecionadas ao analisar os *tweets* sobre o tema em seu período de alta na própria plataforma. Após a coleta, retirando duplicatas exatas (textos 100% idênticos) e *retweets* (republicações feitas por terceiros contendo a mesma mensagem da original), obtivemos 8828 *tweets* únicos.

4.1.2 Análise Preliminar dos Dados

Na figura 6 podemos observar uma maior concentração de *hashtags* nos *tweets* entre os dias 8 e 12 de Março. Nos dias 8 e 9 de março, diversas páginas e sites começavam a discutir sobre a nota de esclarecimento que Dráuzio Varella havia escrito, em resposta às primeiras críticas direcionadas ao médico, devido ao revelamento ao público do crime cometido por Suzy Oliveira[18]. Em sua nota, afirmou que “era médico e não juiz” e que seguia uma conduta de não perguntar o passado de seus pacientes para não deixar que seu julgamento pessoal interfira no seu juramento de médico, e que mantinha este tratamento para seu trabalho na televisão[19]. Porém, o crescimento da hashtag “#drauziovarellalixo” no gráfico da Figura 6, indica que muitas pessoas

[15] https://www.tweepy.org/, acessado em 03/08/2022
[16] https://developer.twitter.com/en/docs/tweets/search/overview/standard, acessado em 14/09/2022
[17] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets, acessado em 14/09/2022
[18] https://emais.estadao.com.br/noticias/comportamento,detenta-trans-drauzio-verella-globo-se-manifestam, 70003226135, acessado em 05/08/2022
[19] https://twitter.com/drauziovarella/status/1236778361130758145, acessado em 05/08/2022

continuaram revoltadas com a situação. As mensagens com críticas e ataques ao médico continuaram, muitas se utilizando de discursos transfóbicos à Suzy no ataque ao Dr. Drauzio.

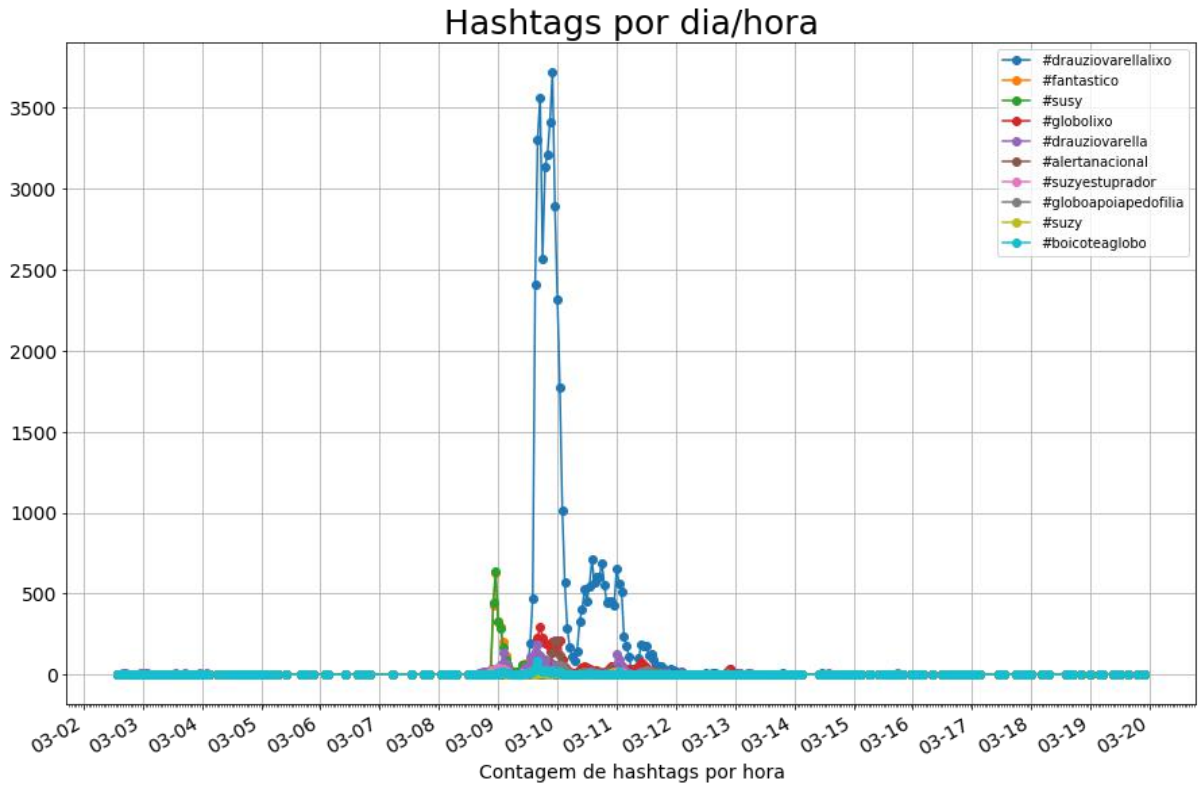

Figura 6: Linha do tempo das *hashtags* nos *tweets* coletados, por dia/hora

4.2 PROCESSO DE ANOTAÇÃO DA BASE

Uma vez coletados os dados, passamos à etapa de anotação dos mesmos, que se refere à classificação manual dos *tweets* para gerar os exemplos de treino que o modelo vai utilizar para identificar de forma automática discursos de ódio (ver seção 2.2). Nesta etapa, os *tweets* foram classificados por seres humanos, de forma manual, com uma ou mais das categorias pré-definidas para este trabalho: “Transfobia”, “Racismo”, “Outro discurso de ódio”, “Linguagem puramente ofensiva” e “Nenhuma das outras”.

Além de transfobia, consideramos também a categoria de racismo, pelo fato de Suzy ser uma mulher trans preta, além de que atualmente a lei que criminaliza o racismo[20] é a mesma que cobre os crimes de transfobia[21].

[20] http://www.planalto.gov.br/ccivil_03/leis/l7716.htm, acessado em 05/09/2022
[21] https://www.conjur.com.br/2019-jun-13/stf-reconhece-criminalizacao-homofobia-lei-racismo, acessado em 06/09/2022

A categoria “Outro discurso de ódio” foi incluída por acreditarmos que todo tipo de discurso de ódio deve ser combatido, por mais que nosso foco neste trabalho seja a transfobia. Já a categoria “Linguagem puramente ofensiva” foi incluída na anotação porque quando palavras ou frases agressivas são usadas, a distinção entre em quais casos se configura ou não discurso de ódio, fica mais complexa (DAVIDSON et al. 2017).

Por fim, a categoria “Nenhuma das outras” é destinada para os tweets que não se enquadram em qualquer das opções de categorias acima.

Para realizar esta classificação manual, procuramos priorizar participantes trans nesta etapa, para que a classificação de transfobia seja feita de forma representativa.

Para selecionar possíveis participantes para a etapa de anotação da base de *tweets*, recorremos a contatos pessoais e de grupos em redes sociais, solicitando que pessoas mais próximas a nós que fossem trans, ou tivessem contato com comunidades trans, servissem de ponte e contactassem mais pessoas.

Das pessoas que alcançamos, muitas não puderam ou se recusaram a participar. Os motivos foram diversos, desde a falta de tempo por outros compromissos no dia-a-dia, até casos mais delicados e individuais de cada um, como algumas que não se sentiam bem emocionalmente para ler textos com conteúdo de ódio. Algo totalmente compreensível, ainda mais nos tempos de pandemia onde as novas realidades, notícias e rotinas nos tomaram e tomam tanta energia. Por fim, também houve pessoas que decidiram não participar pelo fato desta equipe de pesquisa não conter pessoas trans. É importante ressaltar neste ponto que não queremos nos posicionar como protagonistas na luta trans, mas sim como aliados, especialmente considerando a luta contra o discurso de ódio *online* como um todo. Contudo, é uma crítica que ouvimos com respeito. Agradecemos a consideração da pesquisa por todas as pessoas, independente se aceitaram participar ou não.

Embora a participação de pessoas trans nessa etapa do trabalho seja importante, enfrentamos um grande desafio: garantir um número de participantes de forma a anotar um número significativo de *tweets* para o treinamento da máquina. Caso tivéssemos restringido as anotações apenas à comunidade trans, como havíamos planejado, menos pessoas estariam aptas a participar do processo de anotação, reduzindo substancialmente a quantidade de tweets anotados, impactando diretamente o desempenho do modelo. Dessa forma, foi necessária a admissão também de anotadores cis, para que tivéssemos um número razoável de tweets anotados.

Porém, a garantia de representatividade trans neste projeto é importante não só pelo motivo social, sem o qual perde-se o ponto deste trabalho e seu estudo de caso, mas também para a qualidade dos dados que serão alimentados à máquina durante o processo de aprendizado. Dessa forma, limitamos a participação de anotadores cis

apenas a *tweets* já anotados por pessoas trans. Dessa maneira, embora não seja o cenário ideal, conseguimos aumentar o número de *tweets* anotados enquanto ao menos garantimos que um dos anotadores de cada mensagem é trans. Não havendo tweets para o treino da máquina sem passar pela participação de uma pessoa trans.

O processo de anotação dos *tweets* foi realizado no *software* Doccano (NAKAYAMA et al. 2018), uma ferramenta gratuita e aberta que provê, numa página *web*, uma interface gráfica e intuitiva que utilizamos para a visualização de *tweets* e a marcação dos mesmos com uma ou mais das categorias abordadas no projeto. A ferramenta foi instalada e seu acesso disponibilizado aos participantes em uma máquina na nuvem. Utilizamos a modalidade gratuita do serviço do Heroku[22], para esta disponibilização da plataforma Doccano na nuvem. A figura 7 mostra uma captura de tela da interface de anotação do Doccano, com a visualização de um *tweet* e botões para a seleção de anotações.

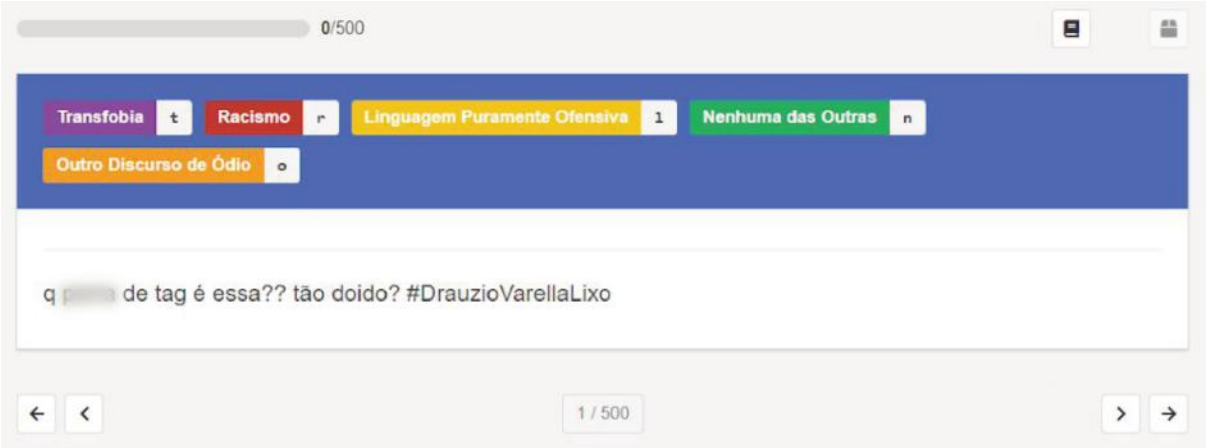

Figura 7: Captura de tela da ferramenta de anotação Doccano.

Criamos, com ajuda de pessoas trans e outras ligadas à causa, um guia de detecção de discurso de ódio com sugestões de definições (também usadas neste texto na seção 2.1) e explicações de utilização da plataforma escolhida para esta etapa, que citaremos mais adiante. Cada participante recebeu acesso a este guia, que está descrito no Anexo A.

Cada participante, de acordo com a sua disponibilidade, recebeu pelo menos um conjunto de 150 *tweets* para a classificação manual. E cada um destes conjuntos de *tweets* foi anotado por três pessoas, de forma a garantir votação ímpar para cada categoria nas mensagens, no caso de empate.

[22] https://www.heroku.com/, acessado em 14/08/2022

Conseguimos a participação de um total de 11 pessoas, algumas se dispondo a anotar mais *tweets* do que o mínimo sugerido a cada participante, já outras, não conseguiram completar o lote mínimo sugerido. Das pessoas participantes, 5 são trans, as outras 6 são cis. Após todos os participantes concluírem o processo de anotação, conseguimos 1050 *tweets* anotados. A figura 8 mostra o número de *tweets* classificados com cada categoria, lembrando que há a possibilidade de interseção de categorias, ou seja, um mesmo *tweet* pode ser classificado tanto com contendo Transfobia quanto como contendo Racismo, por exemplo.

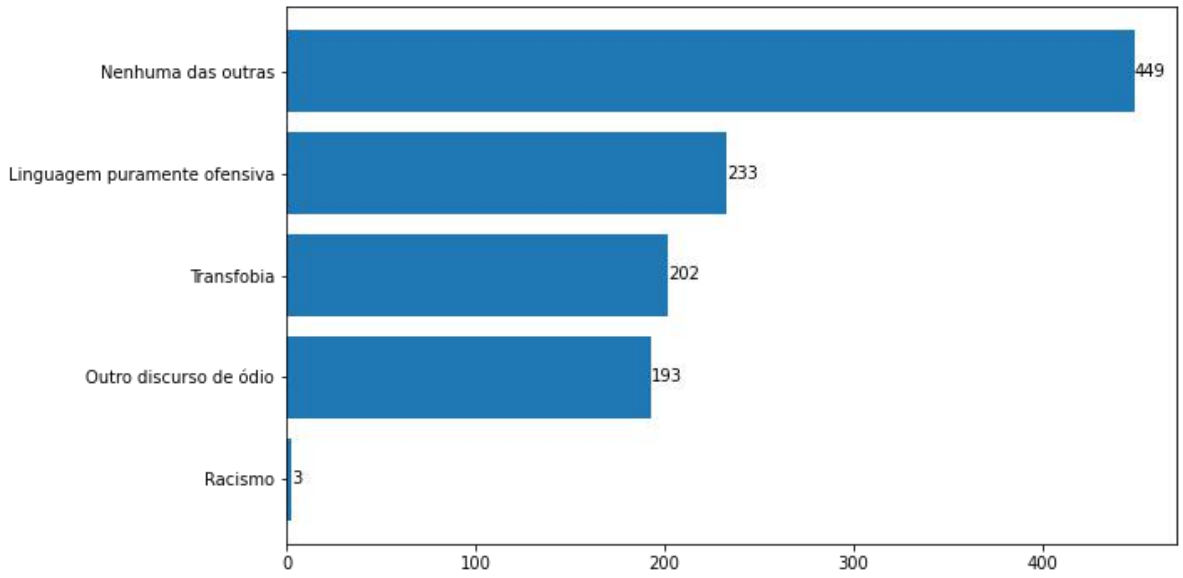

Figura 8: Número de *tweets* classificados com cada categoria (existem casos de interseção).

Conforme podemos ver na figura 8, após votação, apenas 3 *tweets* obtiveram a anotação de “Racismo” e, destes, somente 1 *tweet* foi marcado unicamente com ela, os outros dois tendo também interseção com outras categorias além de “Racismo”. Considerando esta quantidade irrelevante de tweets anotados com racismo, decidimos retirar a categoria “Racismo” da base, já que, de acordo com os anotadores, esta classe não está presente de forma significativa no subconjunto de *tweets* que estamos lidando. O *tweet* marcado somente com “Racismo” foi desconsiderado na base, e os que pertenciam também a outras categorias, seguiram apenas com suas outras classes.

Ainda, em 110 dos 1050 *tweets* passados aos anotadores não houve acordo para as categorias, ou seja, nenhuma delas obteve mais de um voto no total. Os *tweets* nesse caso foram desconsiderados e como também excluímos o único *tweet* anotado somente com racismo, a base final ficou com 939 *tweets* anotados.

4.3 PRÉ-PROCESSAMENTO

Com os *tweets* anotados, avançamos para a formação da base de dados para treino do modelo e teste. Para isso, precisamos transformar os textos dos *tweets* em uma sequência de *features*, uma representação que o modelo possa usar no aprendizado. Utilizamos a abordagem de tokenização do texto, mas antes, normalizamos as mensagens ao incluirmos no tratamento delas as diversas etapas que apresentamos a seguir.

Utilizando um conjunto de expressões regulares, retiramos endereços de *e-mail*, *links* e menções a outros perfis do Twitter. Decidimos aplicar estas remoções porque algumas mensagens poderiam se diferenciar apenas por estes elementos, sendo essencialmente duplicatas, que deveriam ser desconsideradas. Também, para simplificar os textos, removemos todas as aspas simples do texto.

Ainda no pré-processamento, substituímos quaisquer sequências de espaços em branco por apenas um caractere de espaço simples. Também, notamos em nossa base de *tweets*, diversos textos contendo risadas escritas com uma sequência de caracteres “k”. Nesses casos, uma simples diferença de um caractere a mais ou a menos seria suficiente para a geração de diferentes *tokens*, que no nosso caso, enxergamos como um efeito ruim, pois todos eles remeteriam à ideia de risada. Portanto, decidimos substituir quaisquer ocorrências de três “k” ou mais pela forma com apenas três (“kkk”). A Tabela 1, mostra as *strings* de expressões regulares usadas e pelas quais, respectivamente, seus *matches* são substituídos.

Tabela 1: Expressões regulares usadas no pré-processamento dos textos e as substituições realizadas por elas.

Caso	**Expressão Regular**	***Matches* substituídos por**
E-mails e menções	“\S*@\S*”	“” (String vazia)
Links	"(?i)\b((?:https?://|www\d{0,3}[.]|[a-z0-9.\-]+[.][a-z]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s()<>]+\)))*\))+(?:\(([^\s()<>]+|(\([^\s()<>]+\)))*\)|[^\s`!()\[\]{};:'\".,<>?«»“”‘’]))"	“” (String vazia)
Whitespaces	“\s+”	“ “ (Espaço simples)
Aspas simples	"\'"	“” (String vazia)
Risadas com 3+ “k”	"[kK]{3,}"	“kkk”

Após aplicação de todas as expressões regulares, fazemos a tokenização dos textos utilizando o método *simple_preprocess* do módulo Python gensim[23] (versão 4.2). Este método, além da tokenização, faz a conversão do texto para caixa baixa e possibilita retirada de acentos. Utilizamos ambas as funcionalidades como parte da normalização dos *tweets*. Os parâmetros usados no método foram *deacc=True*, *min_len=2* e *max_len=100*. O primeiro ativa a retirada de acentos, o segundo ignora palavras de apenas um caractere e o terceiro ignora palavras com tamanho maior do que cem caracteres para a formação de *tokens*. Escolhemos este valor para *max_len* porque estamos lidando com textos informais e gostaríamos de pegar diversas variações de termos e, além disso, porque *hashtags* podem ser uma junção de palavras que podem até mesmo formar frases sem os espaços, tornando-se uma "palavra" de grande tamanho.

Com os *tokens* formados, retiramos *stopwords* e geramos bigramas que unimos aos *tokens* padrão, de forma que o vocabulário do classificador é composto por unigramas e bigramas. Para as *stopwords,* utilizamos uma mistura de *stopwords* comuns em textos formais em português e mais algumas variações e abreviações que colocamos empiricamente, mais uma vez por estarmos lidando com textos informais na internet.

Após esta etapa do pré-processamento, todas as mensagens foram convertidas em uma sequência de *tokens*. Porém alguns *tweets* ficaram iguais após a tokenização. Eram casos de mensagens quase iguais que se diferenciavam apenas por menções a perfis diferentes, ou por caracteres discrepantes em links encurtados, por exemplo. E em alguns desses casos, os tweets originais foram anotados por trios de pessoas diferentes e que marcaram categorias distintas. Isso resultou em, na forma tokenizada, duplicatas de *tweets* com classificações discrepantes.

Para resolver este problema, as duplicatas que possuíam as mesmas categorias foram simplesmente removidas do *dataset*, permanecendo apenas as primeiras ocorrências de cada mensagem tokenizada. E nos casos onde houve discrepância nas categorias, além de descartar as duplicatas, refizemos a contagem de votos dos anotadores considerando todos os envolvidos nas duplicatas como avaliadores de um único *tweet*, refazendo a votação por maioria em cada categoria para gerar as novas *labels* dos *tweets* em questão.

Com os conflitos acima resolvidos, para a última etapa do pré-processamento dos dados, fizemos a contagem *TermFrequency-InverseDocumentFrequency* (TF-IDF) para os *tokens* no corpus de *tweets*, com o intuito de termos vetores de *features* onde o valor atribuído a cada *token* melhor reflete a sua importância no corpus (palavras muito comuns terão valores TF-IDF menor do que palavras mais específicas). Com a contagem TF-IDF

[23] https://radimrehurek.com/gensim/, acessado em 05/09/2022

realizada, formamos a nossa matriz de *features* para o treino dos modelos classificadores, onde cada linha é um *tweet* e as colunas são as contagens TF-IDF de cada *token*. O processo de contagem TF-IDF foi realizado através do módulo Python scikit-learn (sklearn, versão 1.1), gerando, além da matriz de *features*, o objeto de tokenização que será utilizado novamente durante o pré-processamento de novos *tweets* a serem classificados.

4.4 TREINO DOS MODELOS CLASSIFICADORES

Nesta etapa, utilizamos a nossa matriz de *features*, criada no passo anterior, para o treinamento dos modelos classificadores. Utilizamos três algoritmos de aprendizado de máquina supervisionado: Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), ambos comumente utilizados como *baselines* para performances de aplicações de aprendizado de máquina, e o modelo NBSVM, que é uma combinação dos dois primeiros e que, de acordo com seus autores, mostra boas performances para textos curtos e longos (ver seção 2.2.3).

Como um *tweet* pode ter mais de uma categoria, fizemos uma abordagem de classificação *One vs Rest* onde, para cada um dos algoritmos de aprendizado, treinamos um grupo de quatro classificadores binários (de resposta sim/não), um para cada *label*: “Transfobia”, “Outro discurso de ódio”, “Linguagem puramente ofensiva” e “Nenhuma das outras”. Cada um desses modelos sendo responsável por classificar os *tweets* com suas respectivas classes com que foram treinados.

Todos os modelos classificadores receberam a mesma matriz de *features* para treino e teste. Para a seleção de seus hiperparâmetros, fizemos uma busca exaustiva passando por valores dentro de um intervalo. No caso da necessidade de escolha de mais de um hiperparâmetro, fizemos a busca em todas as combinações possíveis de valores dentro de suas respectivas faixas. Cada instância dessa busca nos dá um conjunto de hiperparâmetros que serão utilizados no treino de um modelo candidato. Após treino e teste de todos os modelos candidatos, mensuramos suas performances e com base nesses valores podemos escolher o modelo que retorna os melhores resultados. Esta abordagem de busca exaustiva também é conhecida como *gridsearch*.

O processo de treino e teste dos modelos foi feito através de *k-fold cross-validation*, com k=5 (ver seção 2.2). O módulo Python sklearn[24] possibilita ambas as abordagens de *gridsearch* e *cross-validation* através da funcionalidade *GridSearchCV* e,

[24] https://scikit-learn.org/dev/index.html, acessado em 06/09/2022

por atender as necessidades do projeto, foi a forma escolhida para esta etapa. Esta funcionalidade requer a definição de uma métrica para referência na escolha dos melhores modelos (e seus hiperparâmetros) e escolhemos a métrica *F_1-score* (ver seção 2.3) por ser uma métrica comumente utilizada na avaliação de modelos de Aprendizado de máquina e que leva tanto em consideração o desempenho dos modelos em *precision* e *recall* (ver seção 2.3).

Na seção 5 poderemos ver que, mesmo nas melhores versões de cada classificador, apenas as categorias “Transfobia” e “Nenhuma das outras” obtiveram resultados acima de 70% de *f-score*. Como a categoria mais importante neste projeto é a “Transfobia”, decidimos continuar para o próximo passo focando apenas no melhor classificador desta classe. Melhorar a performance geral dos classificadores ficando para trabalhos futuros.

4.5 ANÁLISE DAS INFLUÊNCIAS DE TOKENS NA DECISÃO DO CLASSIFICADOR

Como próximo passo, realizamos uma análise das principais *features* do classificador de transfobia com valores SHAP (SHapley Additive exPlanations). Este método vem da teoria de jogos cooperativos e foi concebido para mensurar a contribuição de cada jogador de um time no resultado final de uma partida. Este método posteriormente encontrou aplicação na área de transparência e interpretabilidade de modelos de aprendizado de máquina (ver seção 2.4).

Ao tratarmos cada *feature* como um jogador e cada instância a ser classificada como uma partida, podemos utilizar o método para avaliar a contribuição de cada *feature* na tomada de decisão do classificador na instância em questão. Estes valores são calculados localmente, por instância (no nosso caso, por *tweet*) e, portanto, as mesmas *features* podem contribuir com valores diferentes em cada instância. Porém, além de ajudar na explicação de respostas para um determinado exemplo, podemos criar visualizações com os resultados de valores SHAP para todo o conjunto de dados de forma a ter uma visão geral da contribuição de cada *feature*, como veremos adiante na seção 5.

5 EXPERIMENTO E RESULTADOS

Nesta seção, começamos explicando a implementação de cada um dos algoritmos de aprendizado de máquina e seus resultados. Lembrando que utilizamos a métrica *F_1-score* como parâmetro de avaliação de modelos na funcionalidade *GridSearchCV* (ver seção 4.4).

Em seguida mostramos os resultados do cálculo de valores SHAP e nossa interpretação dos mesmos.

5.1 CLASSIFICADORES MULTINOMIAL NAIVE BAYES

Para a implementação dos classificadores MNB, a variação de hiperparâmetros considerada para os modelos foi seu parâmetro alfa, com os seguintes valores: de 0.1 a 5.0 em passos de 0.1. A tabela 2 mostra os valores de precisão, *recall* e *F_1-score* para os classificadores de cada categoria.

Tabela 2: Performance do melhor modelo treinado com MNB para cada categoria

Categoria	*Accuracy*	*Precision*	*Recall*	***F_1-score***
Transfobia	0.797756	0.537323	0.617857	**0.569676**
Outro discurso de ódio	0.523831	0.263372	0.736842	**0.387230**
Linguagem puramente ofensiva	0.591971	0.346836	0.718297	**0.467504**
Nenhuma das outras	0.757468	0.757116	0.724207	**0.738495**

5.2 CLASSIFICADORES SUPPORT VECTOR MACHINE

Para os modelos SVM, escolhemos *Kernel* linear e o hiperparâmetro variado foi o C. Os valores testados foram as potências de dez, de 10^{-4} até 10^{4}. A tabela 3 mostra os *scores* obtidos para os classificadores de cada categoria.

Tabela 3: Performance do melhor modelo treinado com SVM para cada categoria

Categoria	*Accuracy*	*Precision*	*Recall*	***F_1-score***
Transfobia	0.843081	0.830934	0.358571	**0.496278**
Outro discurso de ódio	0.810659	0.690000	0.147368	**0.231441**
Linguagem puramente ofensiva	0.766351	0.582554	0.195109	**0.287516**
Nenhuma das outras	0.752115	0.740241	0.740169	**0.738961**

5.3 CLASSIFICADORES NBSVM

Como o modelo NBSVM é uma combinação do Multinomial Naive Bayes com o Support Vector Machine, temos os hiperparâmetros de ambas as formas isoladas. Isto é, os valores alfa e C. Além disso, também temos um terceiro hiperparâmetro: beta, que serve como uma interpolação entre os modelos MNB e o SVM adaptado com *features* provenientes de logs de probabilidades, conforme mencionado na seção 2.2.3.

Como temos agora três hiperparâmetros, a combinação de valores destes para o *gridsearch*, juntamente com as repetições do 5-fold *cross-validation*, pode facilmente gerar o treinamento de um número muito alto de classificadores candidatos. Portanto, aumentamos o passo do intervalo de valores do alpha. As faixas de valores utilizadas no *gridsearch* foram de acordo com os seguintes intervalos: alfa de 0.1 a 4.85 em passos de 0.25, valores beta de 0.1 a 0.9 em passos de 0.2 e valores para C nas mesmas potências de dez utilizadas para o SVM, de 10^{-4} a 10^{4}. A tabela 4 mostra os resultados e hiperparâmetros dos melhores classificadores para cada categoria.

Tabela 4: Performance do melhor modelo treinado com NBSVM para cada categoria

Categoria	*Accuracy*	*Precision*	*Recall*	***F_1-score***
Transfobia	0.873408	0.691490	0.766463	**0.726262**
Outro discurso de ódio	0.766251	0.444811	0.426316	**0.427454**
Linguagem puramente ofensiva	0.521698	0.332843	0.821739	**0.468862**
Nenhuma das outras	0.752168	0.685756	0.902090	**0.776963**

5.4 RESULTADOS DO APRENDIZADO

A tabela 5 mostra as métricas F_1-*score* para todas as melhores versões dos modelos treinados. Um fator que achamos curioso, foi a performance do SVM ter sido pior que a do MNB em todas as classes menos “Nenhuma das outras”. E mesmo nesse caso a diferença de *f-score* foi baixa. Isso pode ter sido decorrente da baixa quantidade de dados de treino disponíveis, ou da necessidade de outras tentativas de *kernels* e métodos para melhorar os demais hiperparâmetros dos classificadores.

Na tabela 5 podemos ver que o melhor modelo, para todas as categorias, foi o NBSVM. Porém, mesmo na melhor combinação de hiperparâmetros do *gridsearch*, apenas os seus classificadores das classes “Transfobia” e “Nenhuma das outras” obtiveram resultados acima de 70%.

Outro ponto que vale a pena ser mencionado é que, devido à natureza do nosso trabalho, a métrica de *recall* (ver seção 2.3) pode ser considerada a mais importante para o classificador de transfobia, do ponto de vista de que é pior deixar de detectar casos de transfobia (falsos negativos) do que acusar um *tweet* de conter transfobia erroneamente (falsos positivos), pois no primeiro tipo de erro de classificação, crimes estão deixando de ser detectados. Conforme podemos ver nos resultados de *recall* das tabelas 2, 3 e 4, o *recall* do modelo NBSVM para a classe transfobia foi de 76.6%, sendo melhor do que os outros modelos testados também neste quesito, onde o SVM teve *recall* de 35.8% e o MNB de 61.8%.

Decidimos, continuar para a parte de cálculo dos valores SHAP analisando apenas o classificador de transfobia, dado que é a classe mais importante neste trabalho e foi uma que obteve resultados melhores.

Tabela 5: Métricas F_1-*score* de cada modelo para cada categoria

Modelo	Transfobia	Outro discurso de ódio	Linguagem puramente ofensiva	Nenhuma das outras
MNB	0.569676	0.387230	0.467504	0.738495
SVM	0.496278	0.231441	0.287516	0.738961
NBSVM	**0.726262**	**0.427454**	**0.468862**	**0.776963**

5.5 VALORES SHAP PARA O CLASSIFICADOR DE TRANSFOBIA

Para o cálculo de valores SHAP utilizamos o módulo Python shap[25] versão 0.41. Este é um módulo python muito utilizado para tal propósito, que já implementa todo o processo de cálculo de valores SHAP no estado da arte. Ele possibilita diversos métodos de cálculo através de seus diferentes *explainers* (objetos próprios do módulo que são os responsáveis por calcular os valores SHAP). Dentre os *explainers* disponíveis, utilizamos o *Permutation Explainer* por ser um dos que funciona para qualquer modelo de aprendizado de máquina. Outros *explainers* podem ter melhor performance e precisão nos cálculos, porém só funcionam para modelos específicos como, por exemplo, modelos baseados em árvores de decisão.

Os valores SHAP de todas as *features* são calculados por instância. Utilizamos o modelo treinado para detecção de transfobia com melhor performance em termos de F_1*-score* para detecção da classe em todos os 7358 *tweets* únicos de nossa base que não foram utilizados durante o treino/teste (por não termos conseguido participantes suficientes para anotá-los, enquanto mantendo nosso critério de anotação representativa), e calculamos os valores SHAP para todas estas instâncias.

Com tais valores calculados, plotamos um gráfico no estilo *beeswarm*. Neste gráfico os *tokens* são dispostos por linhas no eixo Y e o eixo X corresponde aos valores SHAP da instância em questão. Os pontos plotados no gráfico são as instâncias do *dataset*, de forma que para cada linha (cada *feature*/*token*), se o conjunto de dados é composto por P instâncias, P pontos serão plotados. A variação da posição destes pontos no eixo Y (mas ainda dentro do espaço dedicado a cada *token* na visualização) também indica a densidade de pontos para os valores SHAP. Por fim, ainda há mais uma informação presente numa visualização de *beeswarm*, a cor de cada ponto representa o valor da *feature*/*token* para esta instância.

A seguir apresentamos dois plots *beeswarm* e os analisamos.

5.5.1 Análise dos termos mais relevantes para a classificação

A figura 9 apresenta um gráfico do tipo *beeswarm* com os top 10 *tokens* para o classificador de transfobia*,* ordenados pela média dos valores absolutos de SHAP por

[25] https://shap.readthedocs.io/en/latest/index.html, acessado em 06/09/2022

todo o *dataset*. Em outras palavras, o plot é composto pelos 10 *tokens* mais importantes para o classificador de transfobia, quando considerado todo o conjunto de dados. São os *tokens* que, em média, tendem a ser mais relevantes para o classificador, seja para a classe positiva quanto para a negativa (dependente do sinal do valor SHAP).

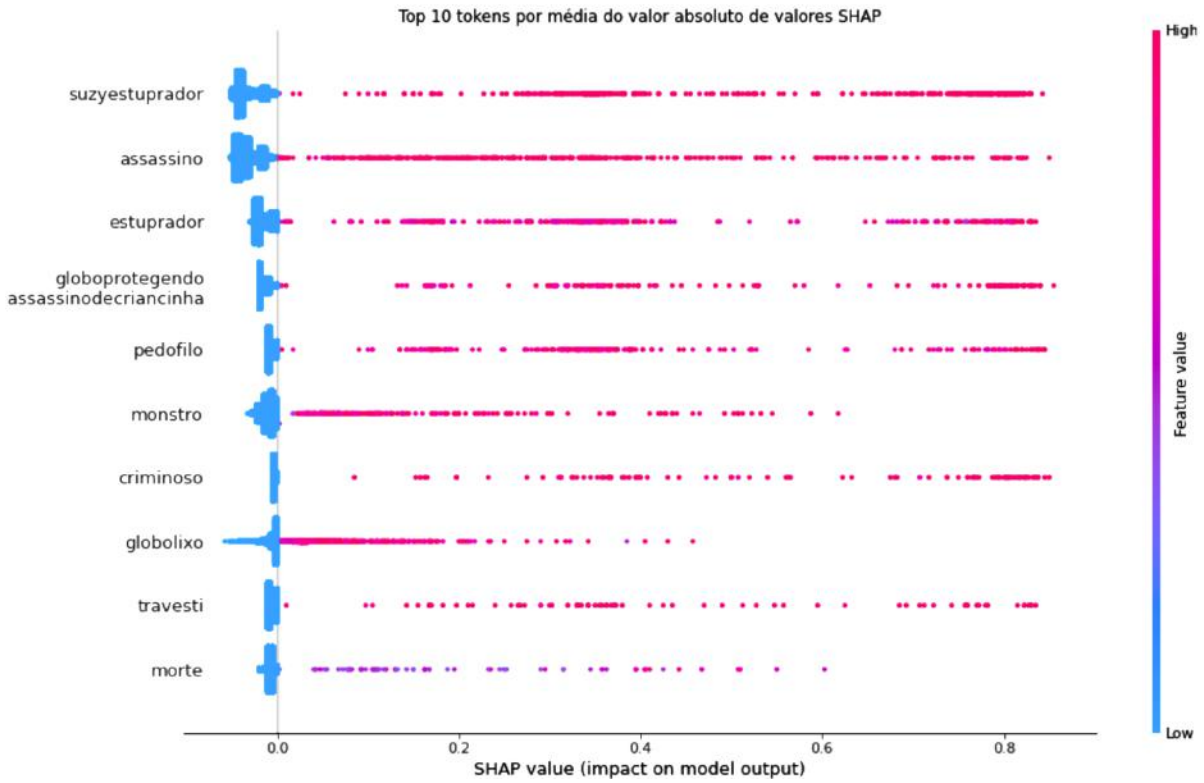

Figura 9: Top 10 tokens mais relevantes para o classificador de transfobia

As cores e os eixos no gráfico nos auxiliam a entender a relevância de cada um dos 10 termos para o classificador. Para cada termo listado, temos os 7358 *tweets* do nosso *dataset* plotados como pontos. A cor destes pontos indica o valor TF-IDF dos termos para cada *tweet* e a posição do ponto no eixo X nos dá o valor SHAP para o termo. Por exemplo, podemos ver que a *hashtag* "#SuzyEstuprador" (representada no gráfico em sua forma tokenizada), quando presente de forma significativa nos *tweets* (pontos de cor vermelha, alto valor TF-IDF), recebe valores SHAP positivos, influenciando o classificador a dar a resposta "transfobia", e quando ausente nos *tweets* (pontos de cor azul, baixo valor TF-IDF), recebe valores shap negativos, influenciando o classificador a responder a classe "não transfobia".

Podemos observar que os termos presentes na figura 9, que são os que mais influenciam o classificador de transfobia, são termos pejorativos, xingamentos e acusações, além de duas hashtags atacando Suzy (representadas em suas formas tokenizadas: "suzyestuprador" e "globoprotegendoassassinodecriancinha") e o termo

“travesti”. Todos eles apresentam um comportamento similar: a presença de forma significativa no texto (cor do roxo para vermelho) destes elementos influencia o classificador a dar a resposta positiva (*tweet* transfóbico). Já a ausência dos termos nos textos (cor azul mais clara), indica ao classificador para dar a resposta negativa (*tweet* não transfóbico). Ou seja, durante o treino, o classificador aprendeu que a presença dos termos listados na figura aumenta a probabilidade do *tweet* em questão ser transfóbico. E, no caso contrário, a simples ausência deles, ao invés de não contribuir em nada (valor SHAP 0.0), contribui para a decisão do classificador de que o texto não é transfóbico, mesmo que em uma escala menor (não chegando a -0.1, mas ainda sendo um valor negativo de SHAP).

Outro comportamento pode ser notado neste plot. A maioria dos pontos azul claro se encontra mais próximo do valor 0.0. Isto é esperado devido ao fato de que nossa matriz de *features* é esparsa, a maioria dos *tweets* não vai conter a maior parte das palavras do vocabulário composto por todos os termos que aparecem em todo o conjunto de dados, principalmente pelo fato de *tweets* serem textos pequenos. Isso, porém, ressalta ainda mais as observações do parágrafo anterior, pois a mera ausência dos termos listados na figura 9, mesmo esse fato sendo algo comum e esperado num *dataset* esparso, ajuda a inclinar o classificador a determinar que o *tweet* em questão não é transfóbico.

Outro fato que se vale notar é que todos dos termos listados no plot, todos os que podem sofrer flexão de gênero estão no gênero masculino, o que, de acordo com as definições usadas neste trabalho (vide seção 2.1), configura transfobia nos casos em que tais termos são usados para se tratar de Suzy.

Das diversas mensagens totais em nossa base de dados, listamos e analisamos a seguir alguns exemplos de textos dos *tweets* da base de dados usada nesta etapa, que ilustram o contexto de uso dos termos. Selecionamos *tweets* manualmente buscando exemplos que continham os termos em questão e que ilustrassem os diversos temas discutidos no caso da Suzy. Nestes, quando presentes, removemos apenas links e menções a outros perfis da plataforma Twitter, que não sejam pessoas públicas, para fins de anonimização. Para facilitar a leitura, também removemos sequências de espaços em branco desnecessários (diversas quebras de linha ou espaços em sequência) e destacamos as ocorrências dos termos listados na figura 9.

Dos diversos *tweets* que contêm os termos em questão, alguns temas em comum podem ser encontrados. O primeiro, que demonstramos a seguir, é o de ataque direto à Suzy:

Claro que os arco-íris iriam defender o **pedófilo**, **estuprador**. Vocês são nojentos. #Boicoteaglobo #**SuzyEstuprador**
Todos por estupro, ser humano escroto e abominável, só porque é transexual não te faz vitima! **Criminoso** e tá pagando sentença como tal! #**SuzyEstuprador** #GloboApoiaPedofilia #**GloboLixo** #BOLSONARO #DrauzioVarellaLixo
Estão tentando humanizar um **pedófilo assassino**. Por mim "a coitadinha que não recebe cartas" já estaria mortO, ou no mínimo em prisão perpétua com trabalhos forçados + castração. + #**GloboLixo** #GloboApoiaPedofilia #**SuzyEstuprador** #drauziovarella
Cada vez que eu vejo a cara desse **pedófilo**, **estuprador** e **assassino**, só me vem à mente a frase do @jairbolsonaro : "É só você não estuprar, não matar, que tu não vai pra lá, p..." Cada vez mais sinto orgulho de ter metido o dedo com força no 17. Vem 2022. #DrauzioVarellaLixo
@jairbolsonaro Vão te acusar de atacar um "inocente" pq essa coisa é "trans". 😡😡😡 #DrauzioVarellaLixo #GloboApoiaPedofilia #**GloboProtegendoAssassinoDeCriancinha** #**GloboLixo**
Em um país justo esse tal Suzy já estaria no corredor da **morte** #DrauzioVarellaLixo
@monicabergamo Sou a favor da pena de **morte** para crimes hediondos, #**SuzyEstuprador** merece a injeção letal, as famílias das vítimas desse montro que merecem nossa compaixão.
Eu também desejo. Que arda no fogo do inferno. Enquanto uma família chora pela **MORTE** ABSURDA DE UMA CRIANÇA, monstros abraçam o seu **assassino**. Que ele vire pó no inferno. #DrauzioVarellaLixo

Podemos observar que, em todos esses exemplos de tweets com ataques diretos à Suzy, a referência a ela é feita com flexão de gênero no masculino, principalmente no terceiro exemplo, onde caixa alta é utilizada especificamente na palavra “mortO” de forma a reforçar a negação da identidade de gênero de Suzy. Este é um exemplo importante de transfobia, que aparece bastante na base de dados anotada. Além disso, em muitos dos ataques podemos ver tentativas de desumanização de Suzy e desejos nocivos a ela, evidenciando um ponto de vista que não acredita na ressocialização de pessoas detentas, chegando ao ponto de desejar pena de morte para certos crimes. Neste caso, um cenário onde mostrar empatia a uma mulher trans que já cumpre pena e está sendo punida por seus crimes é algo inconcebível. Frequentemente, a identidade de gênero de Suzy (e sua negação) também é inserida nas acusações e desejos.

Também podemos ver comentários em apoio a Jair Bolsonaro que, na época, comentou também na plataforma Twitter sobre o assunto, criticando a Rede Globo pela

forma que conduziu a matéria do programa[26]. Em seu tweet, Bolsonaro também se refere à Suzy com pronomes masculinos[27].

Outro tema comum nas mensagens contendo os termos da figura 9, é o ataque ao médico Drauzio Varella, que entrevistou a detenta na matéria do programa Fantástico:

"Esse é o meu jeito". Vabagundo! Se comove primeiro com um **travesti assassino** do que com a vítima. Antes de dar um abraço no algoz, deveria se sensibilizar com o que a vítima sofreu nas mãos deste. Babaca! #DrauzioVarellaLixo #GloboApoiaPedofilia #GloboMentirosa #Globosta
Puxa saco de **pedófilo assassino** #DrauzioVarellaLixo
#DrauzioVarellaLixo Dráuzio Varella é um mentiroso canalha, ele sabia que aquele **travesti** é um monstro **estuprador** de crianças, **assassino**

Nestes exemplos de tweets, vemos novamente que os termos listados no gráfico da figura 9 são usados para fazer referência à Suzy, novamente utilizando o gênero masculino. As mensagens se utilizam do ódio à Suzy e o crime cometido por ela para atacar o médico.

Muitos ataques à Rede Globo também foram encontrados na base. Em algumas das mensagens já demonstradas anteriormente já podíamos notar hashtags contra a emissora, mas também existem tweets de ataque direto à emissora, como os abaixo.

Não acho #DrauzioVarellaLixo. Acho-o ingênuo ou cego. Se deixa usar por uma emissora sem escrúpulos, a fim de apologia política. #RedeGlobo #**GloboLixo** #drauziovarella #drauzio #SUZYLIVRE @RedeGlobo #**SuzyEstuprador** #**GloboProtegendoAssassinoDeCriancinha** #Globo #GloboNews
#DrauzioVarellaLixo #**GloboLixo** Romantizando a historia de um **estuprador** e **assassino** de criança ? a que ponto chega a vontade de defender um bandido ? isso é doença ou é mau caráter mesmo.

Nos *tweets* de ataque à Globo, podemos novamente notar o uso dos termos listados na figura 9, dentre outros também com flexão de gênero masculina, para mencionar à Suzy.

Como podemos ver pelos exemplos de *tweets* contendo termos destacados pela figura 9, muitos deles são de ataques direcionados à Suzy, ao médico Drauzio Varella que a entrevistou e à Rede Globo. Ataques esses, muitas vezes agressivos. Também podemos confirmar que os termos flexionados no gênero masculino são predominantemente referentes à Suzy. Isso, mesmo sem levar em consideração a

[26] https://www.terra.com.br/diversao/tv/blog-sala-de-tv/caso-suzy-ao-se-desculpar-globo-critica-bolsonaro-e-cia,c346e8686cddf584a5be72e91c0b39f1f28g4st8.html, acessado em 04/09/2022
[27] https://twitter.com/jairbolsonaro/status/1237120872676237312, acessado em 04/09/2022

presença ou não de desejos de morte e outras agressividades, já configura discurso de ódio transfóbico por si só, pois a identidade de gênero de Suzy não é respeitada.

Também vale ressaltar que o caso da reportagem da emissora foi à justiça, quando familiares da vítima do crime cometido por Suzy acionaram a justiça para processar a Rede Globo e Drauzio Varella[28]. A defesa da emissora recorreu e afirmou que o propósito da reportagem era divulgar, nas suas palavras, a “precariedade do sistema penitenciário brasileiro e o preconceito contra as pessoas transexuais” e não os crimes cometidos pelas detentas. A sentença foi, então, derrubada por unanimidade. Podemos observar pelo tema de alguns dos *tweets* que este episódio também inflamou a discussão online.

Vemos, através dos exemplos demonstrados, que a transfobia de fato tende a acontecer no cenário estudado, quando os termos listados pela figura 9 estão presentes nas mensagens. Os altos valores SHAP para a classe positiva, em relação aos valores dos demais termos que não chegam ao Top 10 de influência, nos indicam que o modelo de aprendizado de máquina treinado foi capaz de atribuir importâncias corretamente para a ocorrência desses termos na detecção de transfobia. O método SHAP também nos elucida as principais palavras e hashtags correlacionadas com a transfobia no cenário específico estudado.

5.5.2 Análise de termos relevantes para a decisão “não transfobia”

Conforme visto na seção anterior, todos os top 10 *tokens* de maior influência no classificador, quando presentes no texto com altos valores de TF-IDF, influenciavam o modelo a dar a resposta positiva. Nesta nova visualização, buscamos explorar os termos que, quando em altos valores de TF-IDF, influenciam o classificador na resposta negativa. Para isso, precisamos ordenar os *tokens* de outra maneira. Fizemos então outro plot de *beeswarm* com uma ordenação baseada na assimetria (*skewness*, terceiro momento central) da distribuição de valores SHAP de cada *token*. Esta foi a forma que utilizamos para “isolar” tais termos, pois valores de assimetria negativos indicam uma cauda maior para o lado negativo do que para o positivo em uma distribuição normal. Isso também nos ajuda pelo fato de que para a maioria dos termos a média de valores SHAP é em torno de zero, de forma que se há uma cauda que se estende para valores abaixo da média, ela estará na parte negativa do plot.

Porém, mais um ajuste precisou ser feito, pois existem diversos termos que apenas estão presentes em um ou pouco mais de um *tweets*. Isso faz com que a magnitude da

[28] https://dol.com.br/entretenimento/fama/711399/justica-absolve-globo-e-drauzio-varella-no-caso-suzy?d=1, acessado em 07/09/2022

assimetria para esses casos seja muito mais alta do que nos casos de termos mais comuns, que podem até ter uma “cauda maior”. Para resolver isso, levamos em consideração o valor SHAP das *features* somente para textos que possuem o valor de TF-IDF correspondente maior do que zero. Isso muda a distribuição de valores SHAP, principalmente para os casos de termos que só aparecem em pouquíssimos *tweets*, onde a assimetria é praticamente anulada.

A figura 10 mostra o novo plot de *beeswarm* com os top 10 *tokens* ordenados pela assimetria negativa, como descrito nos parágrafos acima. Nela podemos ver os casos de termos que, quando presentes no *tweet*, indicam ao classificador que a instância em questão não se trata de um texto transfóbico. Podemos ver também, que a contrapartida não é tão marcante quanto antes. A ausência dos termos, em média, fica em torno do valor SHAP zero, não influenciando o classificador nem para a classe positiva, nem para a negativa. Apenas em algumas exceções, como na *hashtag* “drauziovarella”, vemos exemplos de *tweets* que tiveram sua classificação influenciada para a classe positiva (transfobia) pela ausência dos termos, mas ainda assim, a maioria dos casos (maior densidade de pontos) ocorre no valor SHAP zero.

Podemos ver que o teor das palavras e hashtags listadas na figura 10 é, com algumas exceções, diferente das da figura 9. Nos *tokens* ordenados por assimetria negativa, temos alguns que remetem ideias neutras, como o nome do médico entrevistador “Drauzio Varella” e palavras como “queria”, “deveria”. Também temos o *token* “humanidade” que pode ser visto como algo neutro ou mais positivo. Já nos *tokens* mais influentes na classificação (figura 9), temos majoritariamente a presença de termos associados com um sentimento negativo e de acusação como: “morte”, “globolixo” e “assassino”.

A seguir analisaremos com mais detalhes alguns dos *tweets* que contêm os termos listados na figura 10, com o mesmo tratamento das mensagens utilizado nos exemplos de mensagens com termos da figura 9.

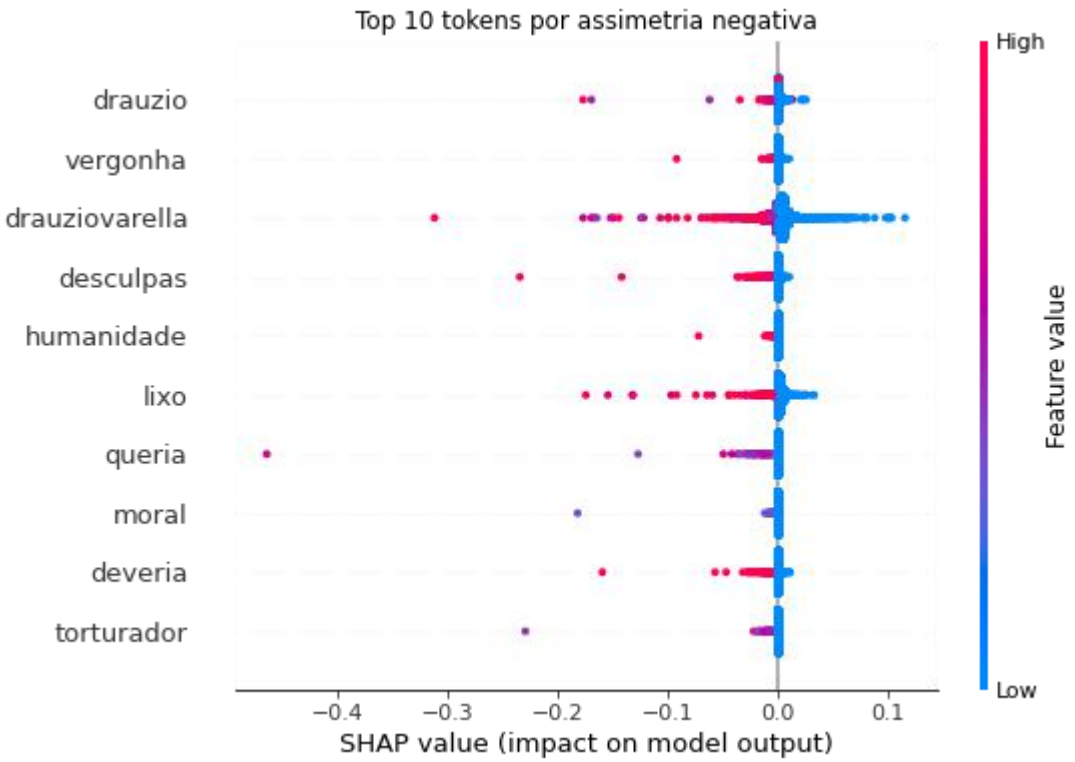

Figura 10: Top 10 tokens por assimetria negativa.

Nos *tweets* que contêm os termos da figura 10, podemos ver uma gama maior de temas. Há mensagens neutras, com o intuito de compartilhar acontecimentos, com níveis de formalidade diferentes.

Após ter o caso revelado por **Drauzio** Varella, presa trans já recebeu 234 cartas #suzyoliveira #**drauziovarella**
Augusto Nunes, José Maria Trindade e Silvio Navarro comentam o pedido de **desculpas** feito por **Drauzio** Varella à família da criança morta pela trans Suzy. #OsPingosNosIs #**DrauzioVarella** #CasoSuzy

Podemos notar que nestes exemplos as referências à Suzy são feitas com a flexão de gênero feminina, que é a correta.

Um grande número de mensagens são de defesa ao Dr. Drauzio Varella, são mensagens que rebatem as acusações e xingamentos direcionados ao médico.

Os direitistas subindo a tag: #DrauzioVarellaLixo pelo simples fato do **Drauzio** exercer a sua função sem olhar a quem, mas os mesmos enaltecem um ser que apoia um assassino, **torturador** que na época estupravam mulheres, homens e crianças de todas as idades. Seus lixos, covardes!

deve ser um **lixo** pra promover TAG galera tenham consciência que se a Suzy cometeu o crime que estão acusando ou não, não é culpa do **Drauzio**, ele não é juiz para julgar crimes de ninguém, eu não apoio ela se isso for verdade, mas o **Drauzio** não tem culpa #DrauzioVarellaLixo
Eu fico indignada com os ataques ao #**drauziovarella** !!! A função do cara é salvar vidas e não puni-las! Ele é tão humilde, ao ponto de gravar um vídeo pedindo **desculpas**! Esse mundo não te mereçe não **Drauzio**.
Emocionante.Treze minutos, uma infinidade de estereótipos quebrados, e a esperança na **humanidade** renovada por saber que existem pessoas como o senhor. Obrigado Dr.! #**drauziovarella** #drauziopresidente
#DrauzioVarellaLixo gente como assim???? **Drauzio** é um fofo, um amor de pessoa, e me impressiona cada vez mais. Muito inteligente, perfeito!! Agora só porque ele abraçou uma mulher ele é **lixo**. Gente, ele só tava fazendo a reportagem dele, não importa o que a mulher fez pra ele +
Queria eu ser uma médica tão humana como o seu Dráuzio #DrauzioVarellaLixo
ah vtm*nc vcs batem palma pro Ustra e acham que tem **moral** alguma pra criticar o Drauzio #DrauzioVarellaLixo
Coitado do **Dráuzio**, deu um abraço no **lixo** radioativo de nome Suzy, e agora está sendo massacrado por gente sem **moral**. Óbvio que ele não sabia que o ser que ele estava abraçando era na verdade uma "besta fera" como dizia minha vó. #DrauzioVarellaLixo
Impossível se decepcionar com esse Sr. e sua atitude humana. Quem critica **deveria** se envergonhar #**drauziovarella**
Vergonha de quem tá usando essa tag pra atacar o **Drauzio** #DrauzioVarellaLixo
Chamam o **Drauzio** de **lixo**, mas passam pano pro presidente que apoia **torturador** (torturava crianças e mulheres grávidas). #DrauzioVarellaLixo

Podemos perceber, nos exemplos acima, que as pessoas que defendem Drauzio Varella não necessariamente defendem Suzy, algumas mensagens inclusive contém ataques à Suzy. Mas algo que podemos notar em todas elas, é que quando fazem referência à detenta trans, utilizam a flexão de gênero correta, a feminina, nem a excluem pelo fato de ser uma pessoa trans, o que não configura transfobia.

Outro dos temas das mensagens envolvendo os termos em questão, ainda é o de ataque ao Drauzio e a Rede Globo, como nos exemplos abaixo:

#DrauzioVarellaLixo **VERGONHA** Todo apoio à família da vítima
#**drauziovarella** Você é um **lixo**, escória! Nojo de você satanista!
Drauzio apoia o aborto e a legalização das drogas. **Drauzio** ocultou o crime. **Drauzio** vitimizou um monstro. **Drauzio** enganou o povo. Pediu **desculpas** só por pressão ,porque aquele abraço escondia um cadáver: real motivo da solidão do assassino. #DrauzioVarellaLixo #EnfermagemPorAmor

Dignidade, hombridade, decência, **humanidade** (a verdadeira), a gente não vê por aqui. #DrauzioVarellaLixo
#DrauzioVarellaLixo **lixo** e 1000 vezes **lixo**. Qualquer um com um pingo de **humanidade** sabe do monstro que ele ajudou a dar visão. Esses seres são pra ficar esquecidos no buraco mais escuro. Imaginem as famílias das vitimas vendo o povo passar pano pra um monstro desses.
Queria fazer reportagem romântica de caridade mais contribuiu com pedofilia fazer um perdofilo virar estrela e o fundo do poço Globo **Lixo** Globo **Lixo**. #DrauzioVarellaLixo

Nos *tweets* acima, vemos novamente exemplos do uso de flexões de gênero masculinas nas referências à Suzy, inclusive com a presença da palavra “assassino” (pertencente às palavras da figura 9) no terceiro exemplo, configurando transfobia. Também, mais uma vez podemos ver o uso do ataque e críticas à Suzy como meio de atacar a emissora e o médico entrevistador.

Por fim, das mensagens que se utilizam dos termos listados na figura 10, temos ainda uma menor parcela de *tweets* de ataque direto à Suzy. Demonstramos um exemplo a seguir:

Suzy. Transsexual. Pobre. Negra. Estupradora. Assassina. Escória. Lixo. Morra. Monstro. Humana. Menos que humana. #DrauzioEnsinaAmor #**drauziovarella** #**humanidade**

Neste exemplo, é feito um ataque direto à Suzy, mas se utiliza da flexão de gênero correta para ela. Porém, ocorre a desumanização, depreciação e desejos de morte à detenta logo após caracterização da mesma com as palavras “Transsexual”, “Pobre” e “Negra”, o que pode configurar transfobia e, até mesmo, classismo e racismo. Também há o uso de *hashtags* de teor contrário ao imediatamente passado pelo corpo da mensagem. Isso pode ter acontecido por motivos de inclusão da mensagem na discussão dos assuntos pertinentes às *hashtags*, mas não temos informações para dar certeza quanto a isso.

De forma geral, em todos os exemplos de *tweets* contendo alguns dos termos listados na figura 10, vemos que o tema é mais focado no Dr. Drauzio Varella. Há alguns *tweets* em defesa do médico e outros continuam sendo de ataque, tanto a ele quanto à Rede Globo. Os ataques à Suzy ainda estão presentes, mas pode-se notar que o grau de referências à ela com pronomes e flexões de gênero erradas diminui, embora ainda encontramos alguns casos.

Os *tweets* deste recorte ainda podem ser agressivos e até mesmo conter discursos de ódio de outra natureza, como classismo e/ou racismo, porém a quantidade dos que configuram especificamente um discurso transfóbico diminuiu. Esse resultado bate com o indicado pela figura 10, onde a presença dos termos listados nos *tweets* (cores roxa e

vermelha, valores TF-IDF mais altos) aumentam a chance do *tweet* em questão não ser transfóbico. Ainda quanto ao método SHAP, vale lembrar que embora o valor de uma *feature*/*token* isolada possa ser alto para algum dos lados (classe positiva ou negativa, em nosso classificador binário de transfobia), isso não significa que o classificador responderá a classe correspondente como resultado. Isso pode ocorrer tanto por influência dos demais *tokens*, que contribuirão com suas respectivas influências no processo, quanto por um eventual erro de classificação.

6 CONSIDERAÇÕES FINAIS

Neste trabalho criamos um detector de transfobia para o caso específico da repercussão, na plataforma do Twitter, acerca da entrevista do Dr. Drauzio Varella à detenta trans Suzy Oliveira. Para a preparação dos dados de treino, utilizamos um método rigoroso de anotação da base, priorizando a participação de anotadores trans, mesmo que tal abordagem traga desafios de disponibilidade de recursos humanos para a anotação. Mesmo assim conseguimos bons resultados, atingindo, por exemplo, uma performance acima de 72% de F_1-*score* para as classes “Transfobia” e “Nenhuma das outras”.

Uma grande contribuição do nosso trabalho foi analisar os termos mais relevantes utilizados pelo classificador na tarefa de detectar discursos transfóbicos. Utilizamos um método da área de Interpretabilidade de Aprendizado de Máquina (método SHAP), que tem sua origem na teoria de jogos cooperativos. O método SHAP nos permitiu calcular as influências que cada *feature* (*tokens* de palavras no nosso caso) exerce sobre o modelo treinado. Com as informações trazidas pelos valores SHAP em gráficos de *beeswarm*, conseguimos averiguar que para um modelo detector de transfobia com uma boa performance, o modelo conseguiu aprender que palavras agressivas com flexão de gênero masculina estão associadas a discursos transfóbicos, dado o contexto dos *tweets*. E palavras comuns ao tema, mas sem tal flexão de gênero, como “Drauzio” e “lixo”, podem estar ligadas a agressividade ou não, mas com menor probabilidade do *tweet* em questão ser transfóbico, pois, nos casos que vimos, os tópicos dos textos com maior presença de termos como estes eram mais distante da Suzy em si, ou pelo menos a atacavam se utilizando menos de transfobia ou deixando este discurso menos evidente à primeira vista. Os valores SHAP para o classificador de transfobia pareceram ser condizentes com o esperado do tema restrito que abordamos.

Quanto à análise do discurso transfóbico no cenário estudado, pudemos concluir que, para este cenário, a transfobia aparece com mais frequência na forma de termos com flexão de gênero masculina direcionados à Suzy. As mensagens contendo este tipo de termos muitas vezes continham desejos nocivos à Suzy e se utilizavam dela como meio de ataque e repúdio ao Dr. Drauzio Varella e a Rede Globo, enunciando um ponto de vista onde mostrar empatia a uma mulher trans que já cumpre pena por seus crimes é algo intolerável.

O nosso trabalho foi pioneiro na investigação de métodos de aprendizado de máquina no combate ao discurso de ódio transfóbico em português e que serve como referência a mais trabalhos tratando do tema da transfobia e análise de discursos transfóbicos, nas áreas da computação, social e no encontro das áreas. Rumo a um futuro e presente mais inclusivo.

7 TRABALHOS FUTUROS

Como trabalhos futuros, deixamos a exploração de métodos mais sofisticados de aprendizado de máquina, como diversos tipos de redes neurais, métodos que combinam árvores de formas mais sofisticadas e robustas, e a consequente exploração do método SHAP nos modelos criados com tais algoritmos. Com o intuito de ver o comportamento desses modelos quanto às *features* textuais em modelos mais precisos e possivelmente mais “caixa preta” (o que justifica ainda mais a necessidade de interpretabilidade destes através de valores SHAP ou outras alternativas).

Também deixamos como trabalho futuro a tarefa de criação de uma base maior de textos transfóbicos e não transfóbicos, e com maior diversidade de temas e cenários, anotada de forma inclusiva. Visando expandir o escopo de detectores de transfobia para além de apenas um cenário restrito, explorando mudanças no comportamento dos modelos treinados e diferenças de performance.

REFERÊNCIAS

ALMEIDA, Silvio Luiz de. **O que é racismo estrutural?**. Belo Horizonte: Letramento, 2018.

BENEVIDES, Bruna (org.). **Dossiê assassinatos e violências contra travestis e transexuais brasileiras em 2021**. Brasília: Distrito Drag, ANTRA, 2022.

BISHOP, Christopher M.; NASRABADI, Nasser M. **Pattern recognition and machine learning**. New York: springer, 2006.

BURKOV, Andriy. **The hundred-page machine learning book**. Quebec City, QC, Canada: Andriy Burkov, 2019.

CAMELO, Fábio Assunção Berlim. **Detecção automática de discursos de ódio em comentários de jornais online**. 2018.

DAVIDSON, Thomas et al. **Automated hate speech detection and the problem of offensive language**. In: Proceedings of the international AAAI conference on web and social media. 2017. p. 512-515.

FORTUNA, Paula; NUNES, Sérgio. **A survey on automatic detection of hate speech in text**. ACM Computing Surveys (CSUR), v. 51, n. 4, p. 1-30, 2018.

JESUS, Jaqueline Gomes de. **Orientações sobre a população transgênero: conceitos e termos**. Brasília: Autor, p. 1-30, 2012.

LUNDBERG, Scott M.; LEE, Su-In. **A unified approach to interpreting model predictions**. Advances in neural information processing systems, v. 30, 2017.

FREITAS MELO, Philipe de et al. **Can WhatsApp counter misinformation by limiting message forwarding?**. In: International conference on complex networks and their applications. Springer, Cham, 2019. p. 372-384.

GAMBÄCK, Björn; SIKDAR, Utpal Kumar. **Using convolutional neural networks to classify hate-speech**. In: Proceedings of the first workshop on abusive language online. 2017. p. 85-90.

MANNING, Christopher D et al. **Introduction to information retrieval**. Syngress Publishing, 2008.

MARCÍLIO, Wilson E.; ELER, Danilo M. **From explanations to feature selection: assessing shap values as feature selection mechanism**. In: 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI). Ieee, 2020. p. 340-347.

MENG, Yuan et al. **What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values**. Journal of Theoretical and Applied Electronic Commerce Research, v. 16, n. 3, p. 466-490, 2020.

NAKAYAMA, H et al. **doccano: Text Annotation Tool for Human**. 2018. Disponível em: <https://github.com/doccano/doccano>, acessado em 06/09/2022

RUBACK, Lívia; OLIVEIRA, Jonice. **Analyzing polarization in Twitter: The murder of Brazilian councilwoman and activist Marielle Franco**. 2018.

RUBACK, Lívia; CARVALHO, Denise; AVILA, Sandra. **Mitigando Vieses no Aprendizado de Máquina: Uma Análise Sociotécnica**. iSys-Brazilian Journal of Information Systems, 2022. Disponível em: <https://sol.sbc.org.br/journals/index.php/isys/article/view/2396>, acessado em 07/09/2022

SALVADOR, Nayara Cunha et al. **Fracasso, evasão e abandono escolar de pessoas trans: algumas reflexões necessárias**. Revista de Educação Pública, v. 30, p. 1-18, 2021.

TUFEKCI, Zeynep. **Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency**. Colo. Tech. LJ, v. 13, p. 203, 2015.

WANG, Sida I.; MANNING, Christopher D. **Baselines and bigrams: Simple, good sentiment and topic classification**. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2012. p. 90-94.

Anexo A - Estudo da UFRJ para classificar transfobia no Twitter (Caso Suzy)

Este guia serve como base para classificação manual de tweets coletados a partir de hashtags encontradas na repercussão no Twitter do caso da entrevista de Drauzio Varella com a detenta Suzy Oliveira no programa Fantástico de 1° de Março de 2020[29]. Além da detecção de transfobia, também incluiremos no trabalho a detecção de racismo, devido à particularidade do caso Suzy, que além de trans é preta. E também levando em consideração que a lei do racismo atualmente é a que cobre os casos de transfobia.

- Cada participante terá um conjunto de aproximadamente 150 tweets para classificar.
- Os tweets deverão ser classificados em uma das categorias:
    - Transfobia
    - Racismo
    - Outro discurso de ódio
    - Linguagem puramente ofensiva
    - Nenhuma das outras

Vale ressaltar que transfobia, racismo e outros tipos de ódio podem se manifestar em uma mesma mensagem - ou no mesmo tweet. Neste caso, deve-se marcar todas as opções válidas. Abaixo daremos mais detalhes sobre cada uma das categorias e instruções de uso da plataforma de classificação dos textos.

Categorias:

As definições e exemplos a seguir são sugestões baseadas em teóricas e teóricos de seus respectivos campos.

1) Transfobia

De acordo com a teórica e mulher trans Jaqueline Gomes de Jesus, em seu guia técnico “Orientações sobre identidade de gênero: conceitos e termos”, transfobia é: “Preconceito e/ou discriminação em função da identidade de gênero de pessoas transexuais ou travestis.” (DE JESUS, 2012)

[29] https://g1.globo.com/sp/sao-paulo/noticia/2020/03/07/detenta-trans-suzy-ja-recebeu-234-cartas-apos-reportagem-do-fantastico-diz-secretaria-de-sp.ghtml

No caso de textos online, são exemplos quaisquer ataques, ameaças ou desrespeito ao corpo trans, como por exemplo: negar a identidade de gênero de uma pessoa (chamar, tratar ou se referir a uma mulher trans, como Suzy, com pronomes direcionados a homens. E o contrário para homens trans).

2) Racismo

Silvio Luiz de Almeida, em seu livro “O que é racismo estrutural?”, define racismo como “uma forma sistemática de discriminação que tem a raça como fundamento, e que se manifesta por meio de práticas conscientes ou inconscientes que culminam em desvantagens ou privilégios para indivíduos, a depender do grupo racial ao qual pertençam.” (DE ALMEIDA, 2018)

3) Outro discurso de ódio

Neste trabalho, focaremos na detecção de transfobia e racismo. Porém, todas as manifestações de ódio devem ser combatidas. Para os demais tipos de discurso de ódio que eventualmente podem ser encontrados nos textos usaremos esta categoria.

Para uma definição geral, usaremos a de Paula Fortuna e Sérgio Nunes em sua revisão sistemática da literatura acerca do tema de detecção automática de discurso de ódio:

> "Discurso de ódio é linguagem que ataca ou diminui, que incita violência ou ódio contra grupos, baseado em características específicas como aparência física, religião, descendência, origem nacional ou étnica, orientação sexual, identidade de gênero ou outras, e pode ocorrer em diferentes estilos linguísticos, até mesmo de formas sutis ou quando humor é utilizado."(FORTUNA e NUNES, 2018. Tradução nossa)

4) Linguagem puramente ofensiva

Esta categoria se refere a textos que contém linguagem ofensiva, mas que (ao contrário das categorias acima) não se utilizam de discriminação.

É importante mencionar uma particularidade importante neste contexto: Muitas pessoas usam termos que podem ser considerados ofensivos de forma geral, mas que se são manifestados por alguém da própria comunidade que seria atacada, não seriam ofensivos. Como por exemplo, alguns perfis de drag queens que utilizam linguagem comumente vista como mais agressiva e até podendo conter xingamentos de origem LGBTQfóbica, mas que são utilizados pela comunidade, muitas vezes num tom de protesto ou de ressignificar os termos[30].

[30] https://tab.uol.com.br/noticias/redacao/2020/03/06/bom-dia-gay-guerra-ao-discurso-de-odio-nas-redes-derruba-posts-lgbtq.htm

5) Nenhuma das outras

Esta categoria inclui textos que contém mensagens que não se encaixam em nenhuma das situações acima, ou seja, sem nenhum tipo de discurso de ódio e sem utilizar linguagem ofensiva.

Plataforma de Classificação dos Textos:

A plataforma que utilizaremos se chama Doccano[31]. Você receberá uma conta com login e senha, além de um link para o acesso da plataforma. O botão de login fica no canto superior direito da tela, conforme mostra a figura 1.

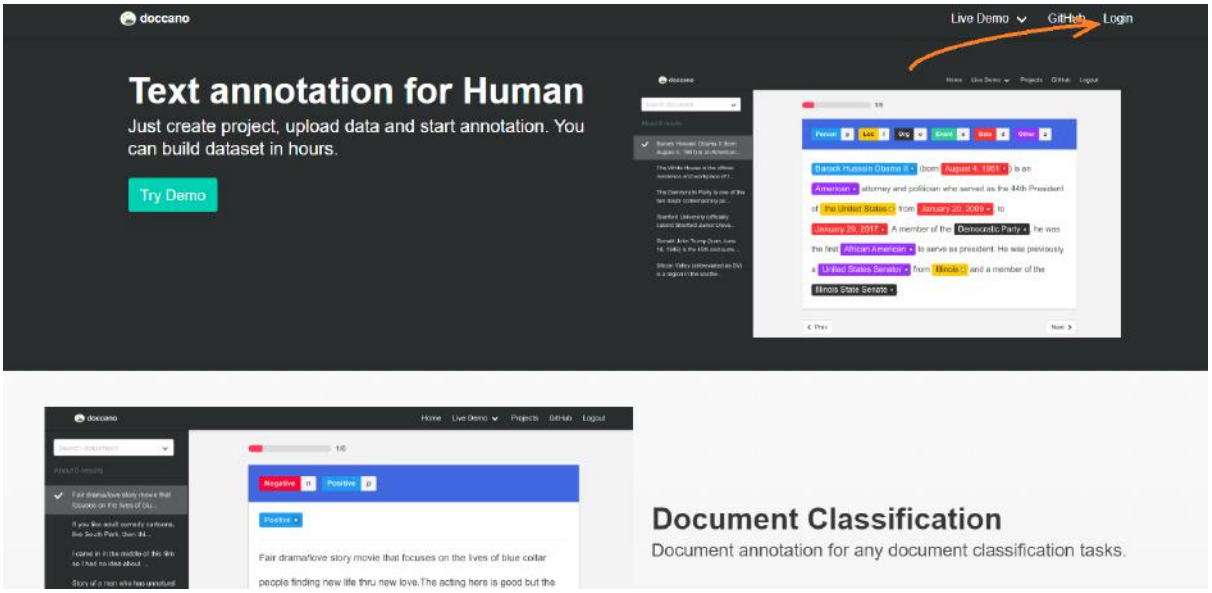

Figura 1: Tela inicial da plataforma

Após logar na plataforma, você verá uma tela como na figura 2, que lista os “projetos” que você estará participando. No nosso caso, os “projetos” serão grupos de aproximadamente 150 tweets a serem classificados. Ao selecionar um projeto para trabalhar, a próxima tela será como a da figura 3, onde se realizará a classificação dos textos.

[31] https://github.com/doccano/doccano

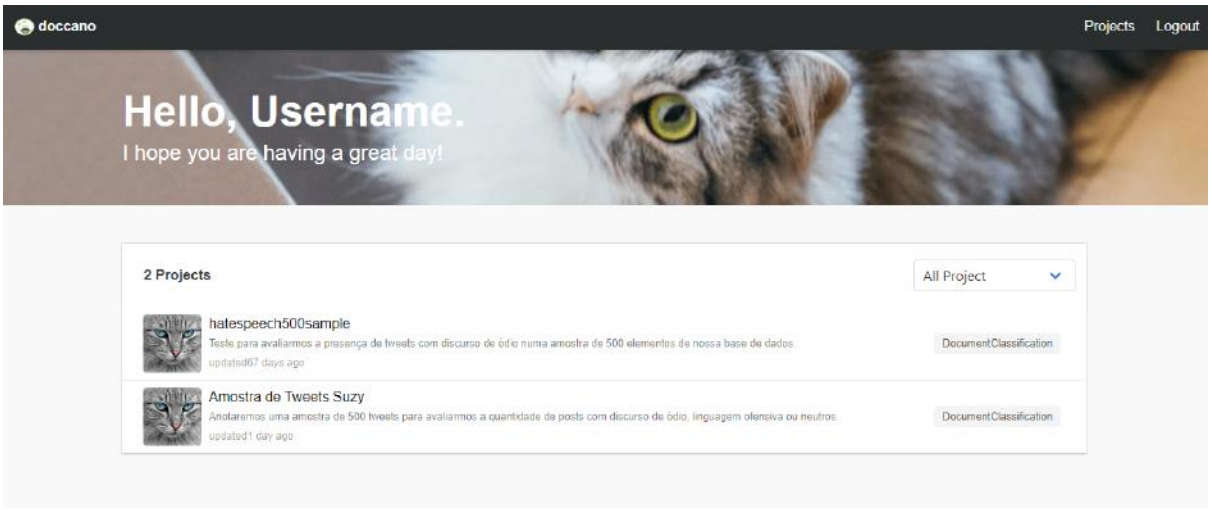

Figura 2: Tela de seleção de projetos

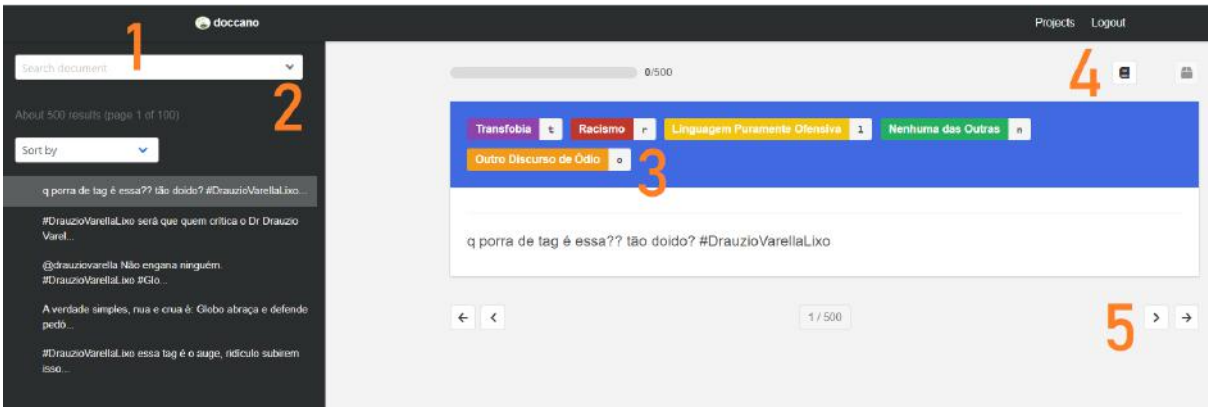

Figura 3: Tela de classificação dos textos. Legendas abaixo.

1 - Busca textual; 2 - Filtrar tweets (Todos, Respondidos, Restantes)

3 - Categorias e seus respectivos atalhos no teclado; 4 - Resumo do guia de classificação

5 - Próximo tweet/página (atalho: setas do teclado)

Referências:

DE ALMEIDA, S.; *O que é racismo estrutural?* Letramento, 2018.

DE JESUS, J. *ORIENTAÇÕES SOBRE A POPULAÇÃO TRANSGÊNERO: CONCEITOS E TERMOS*. Abril, 2012.

FORTUNA, P.; NUNES, S. *A survey on automatic detection of hate speech in text.ACMComputing Surveys (CSUR)*, v. 51, n. 4, p. 1–30, 2018.