Técnicas de Machine Learning Aplicada a Mineração de Dados e Análise de Sentimentos para Predição de Homofobia no Twitter

Mayk Brendon Almeida Antunes
Instituto Federal de Educação, Ciência e Tecnologia do Norte de Minas Gerais
Brasil
maykbrendon.antunes@gmail.com

Matheus de Freitas Issa
Instituto Federal de Educação, Ciência e Tecnologia do Norte de Minas Gerais
Brasil
matheusdefreitask9@hotmail.com

Raphael Magalhães Hoed
Instituto Federal de Educação, Ciência e Tecnologia do Norte de Minas Gerais
Brasil
raphael.hoed@ifnmg.edu.br

ABSTRACT
This project approaches the identification of homophobic tweets, using a natural language processing and machine learning approach. The data mining methodology applied in this project is CRISP-DM. The goal is to build a predictive model that can detect, with reasonable accuracy, whether a Tweet contains content that is offensive to individuals in the LGBTQIA+ community or not. The database used to train the predictive models was built from several tweets collected. Over 3000 tweets were collected amidst project development, obtaining results with 86% accuracy.

Keywords
Sentiment Analysis; Machine Learning; Supervised Learning; Homophobia; Twitter; Tweets.

RESUMO
Este trabalho estuda a identificação de tweets homofóbicos, utilizando uma abordagem de processamento de linguagem natural e aprendizado de máquina. A metodologia de mineração de dados aplicada neste trabalho foi a CRISP-DM. O objetivo é construir um modelo preditivo que possa detectar, com razoável precisão, se um Tweet contém conteúdo ofensivo a indivíduos da comunidade LGBTQIA+ ou não. O banco de dados utilizado para treinar os modelos preditivos foi construído a partir de tweets diversos coletados. Foram coletados mais de 3000 tweets para desenvolvimento do nosso trabalho, obtendo resultados com 86% de precisão.

Palavras-Chave
Análise de Sentimentos; Aprendizagem de Máquina; Aprendizagem Supervisionada; Homofobia; Twitter; Tweets.

ACM Classification Keywords
Twitter, Aprendizado de Máquina.

INTRODUÇÃO

As redes sociais são um lugar onde a comunidade LGBTQIA+ (lésbica, gay, bissexual, transgênero, queer, intersexual, assexual, entre outras expressões de gênero e sexualidade) pode se expressar livremente e se conectar com outras pessoas que compartilham os mesmos interesses. Sendo uma das mais ativas nas redes sociais, a comunidade LBTQIA+ usam as hashtags *#LGBTQIA* e *#Pride* para promover a causa e aumentar a visibilidade, criando um lugar onde as pessoas possam se apoiar, compartilhar histórias, experiências e se sentir parte de uma comunidade.

Muitos usuários usam as mídias sociais para assediar, ameaçar e atacar pessoas da comunidade LGBTQIA+, levando a questões como assédio, discriminação e falta de representação. Diante do cenário, foi levantada a seguinte questão: postagens de cunho homofóbico ganham engajamento e repercussão na rede social do twitter?

A proposta deste trabalho é aplicar técnicas de aprendizagem de máquina, processamento de linguagem natural e análise de sentimentos a fim de identificar discursos de ódio e conteúdos prejudiciais à comunidade LGBTQIA+ na rede social twitter, realizando uma análise sobre a relevância e o alcance que esses determinados textos classificados têm na rede, resultando em dados que serão inseridos em dashboards para melhores visualizações. Após a coleta e classificação, o objetivo será construir um modelo preditivo a fim de automatizar a identificação e denúncia dos textos em questão.

REFERENCIAL TEÓRICO

Esta seção está organizada em 4 subseções: **Rede Social Twitter** - discorre sobre a rede social objeto deste estudo; **Homofobia no Brasil** - Traz dados sobre práticas homofóbicas no âmbito nacional; **Mineração de dados** - Apresenta conceitos e usos da mineração de dados; **SGDClassifier** - Apresenta o classificador SGD utilizado no âmbito deste estudo; **Classificação de Texto** - apresenta a técnica de classificação de texto utilizada no âmbito deste estudo.

Rede Social Twitter

O Twitter foi criado pela empresa Obvious, em 2006, e é caracterizado como microblog que permite aos usuários enviar ou receber atualizações que contenham, no máximo, 280 caracteres. De acordo com o próprio site, ao completar 7 anos de lançamento no início de 2013, o número total de usuários ativos era de 200 milhões, gerando mais de 400 milhões de tweets por dia e sendo aproximadamente 33 milhões de brasileiros. [18]

Pesquisas estatísticas recentes, há um total de 1.3 bilhão de contas no Twitter, mas apenas 330 milhões são usuários ativos, onde 500 milhões de tweets são publicados a cada dia, sendo 350,000 tweets postados a cada minuto. [28]

Sendo uma rede social muito utilizada como meio publicitário e promoção de marcas, ao longo dos anos, a mobilização dos movimentos feminista, negro e LGBTQIA+ vem acionando questões de gênero, raça e sexualidade que hoje mobilizam algumas empresas a investir na construção de uma marca socialmente comprometida, ao mesmo tempo em que gera nos consumidores a consciência de cobrar das marcas valores e direitos humanos básicos. O uso do Twitter, entre outras redes sociais, tem sido feito como meio para divulgação de seus posicionamentos acerca de temas relacionados à diversidade de gênero, raça e sexualidade, além de seus programas de responsabilidade social relacionados a esses temas. [3]

Foi realizada uma pesquisa onde os autores abordaram a apropriação da *hashtag* "Quem Lacra Não Lucra" (*#quemlacranãolucra*) da campanha publicitária da marca *Burger King* relativa ao Dia Internacional do Orgulho LGBTQIA+ de 2020, onde interessa identificar o contexto da ação da marca e o subsequente engajamento dos usuários da rede social Twitter. O que se evidenciou com o termo "Lacrou!" e seu desdobramento na expressão "Quem Lacra Não Lucra" demonstra as disputas de poder que ora evidenciam uma subversão pelas minorias e, em outro momento, são cooptadas pelas marcas, como estratégia de posicionamento de mercado. [1].

Pesquisas realizadas pelo *Center for Countering Digital Hate* (CCDH) [27], mostram 989,547 tuítes postados entre os meses de janeiro e julho no ano 2022, que mencionam a comunidade LGBTQIA+ junto com insultos como *groomer*, predador e pedófilo. Uma auditoria concluiu que o Twitter falhou em agir frente a 99% das 100 denúncias sobre tuítes com discurso de ódio feitas anonimamente pelos pesquisadores do CCDH.

O Twitter é majoritariamente empregado para cobertura de eventos políticos, econômicos e sociais. A replicação por parte dos usuários, seja na forma de "retuítes", seja na forma de "curtidas", atua de forma complementar e ortogonal à circulação das notícias pelos canais tradicionais, uma vez que pode emprestar destaque especial a determinados temas ou obstruir a difusão de eventos específicos.[22]

Segundo pesquisadores [18], o Twitter se tornou uma ferramenta muito importante no meio político, pois é interessante por se mostrar como um local de ressonância de temas e discussões políticas que são divulgadas pelos mais diversos meios de comunicação. Parece ser nas mídias sociais que as questões políticas repercutem e ganham diferentes desdobramentos. Os autores concluíram que o uso político do Twitter aumentou o ativismo, tornou os usuários mais questionadores, curiosos e informados, forçando todos os usuários a serem mais concisos e claros, mudando o modo como os negócios políticos interagem com os cidadãos/eleitores e influenciando como as notícias passam por outras plataformas midiáticas.

O Twitter no meio político consiste em basicamente três ideais: (1) É uma maneira rápida de obter informações políticas sem filtragem; (2) Satisfaz o desejo dos usuários que querem fazer parte do processo político; (3) É uma ótima ferramenta para quem faz trabalho político ou reportagem sobre política;[13]

Homofobia no Brasil

Com base em preconceitos e estereótipos, os números da violência LGBT são subestimados, pois as investigações são conduzidas a partir desse viés, mesmo em países com leis que punem esse tipo de agressão. Em alguns países, a falta de estatísticas governamentais torna a violência contra grupos minoritários invisível. Assim, organizações não governamentais como o Grupo Gay da Bahia fazem suas próprias pesquisas para denunciar e tornar visível a violência contra LGBT. De acordo com o relatório de 2016 do Grupo Gay da Bahia, 343 LGBTs foram assassinados no Brasil em 2016, tornando o Brasil o campeão mundial de crimes contra as minorias sexuais [12]. Segundo o antropólogo Luiz Mott [16], responsável pelo site "Quem a Homofobia matou hoje": "números tão alarmantes são apenas a ponta de um iceberg de violência e sangue, já que não há estatísticas governamentais sobre crimes de ódio, esses números são sempre subnotificados, pois nosso banco de dados é baseado em notícias publicados na mídia, na internet e informações pessoais".

A organização não governamental Safernet Brasil, que monitora e denuncia crimes e violações de direitos humanos na internet desde 2005, confirma o aumento de casos de denúncias de material tendencioso e racista na internet. Em 2006, quando a ONG começou a contá-los, havia 11.444 casos de denúncias de conteúdo ofensivo com homofóbicos, xenófobos, intolerância racista, nazista e religiosa. Em 2012, o número subiu para 21.033 casos, um aumento de cerca de 45%. Em 2016, a SaferNet Brasil processou 2.891 denúncias anônimas de homofobia envolvendo 1.436 páginas diferentes. 10% deles foram compartilhados em perfis no Twitter e 63% em publicações no Facebook [12].

O trabalho "*Using supervised machine learning and sentiment analysis techniques to predict homophobia in portuguese tweets*" estuda a previsão de conteúdo

homofóbico, em tweets em língua portuguesa. Tweets com temas LGBT foram filtrados e classificados manualmente.Uma ampla gama de técnicas de aprendizado de máquina foi empregada na tarefa de classificação, após estimar e ajustar cada modelo, eles foram combinados usando *voting* e *stacking*. Utilizando 10 modelos, o classificador de *voting* obteve 89,4% de precisão. [12]

Mineração de dados

Provost e Fawcett [8], ressaltam que a Ciência de Dados envolve princípios, técnicas e processos para entender fenômenos de diversas áreas através da análise de dados. Assim, a ciência de dados apoia a tomada de decisão orientada por dados que pode ser realizada manualmente ou de forma automatizada.

Destaca-se que a aprendizagem de máquina (machine learning) consiste em empregar técnicas para que os computadores aprendam a partir de dados históricos. Géron (2019) afirma que isso é possível através da disponibilização de dados e emprego de algoritmos específicos para que haja um aprendizado dessas informações e a criação de modelos de decisões que dessa forma serão utilizados como ferramenta de predição e também de análise.

De acordo com os autores [20], com o advento da área de mineração de dados, e posteriormente ciência de dados, tornou-se necessária a padronização de projetos de mineração de dados para propiciar um maior controle e previsibilidade das atividades, além de prover meios para a avaliação da efetividade da análise. Uma dessas metodologias é o CRISP-DM (*Cross Industry Standard Process for Data Mining*).

Com o advento da ciência de dados e com a disponibilização de ferramentas cada vez mais acessíveis no mercado para a realização das suas diversas etapas, diversas áreas do conhecimento têm aproveitado dos seus benefícios.

A linguagem de programação Python está se estabelecendo como uma das linguagens mais populares para computação científica. Graças à sua natureza interativa de alto nível e seu ecossistema maduro de bibliotecas científicas, é uma opção atraente para desenvolvimento algorítmico e análise exploratória de dados (Dubois, 2007; Milmann e Avaizis, 2011).

A biblioteca Scikit-learn [11] é uma das mais utilizadas para a resolução de problemas de aprendizado de máquinas devido sua enorme gama de modelos de aprendizado disponíveis e das demais ferramentas que são utilizadas para manipular os dados e deixá-los preparados para os modelos. Ela foi criada em cima de estruturas do Numpy para que o custo computacional seja menor, tornando-a extremamente atraente para cientistas e pesquisadores.

A análise de sentimentos em redes sociais evoluiu não só para detectar a polaridade das mensagens, mas também para detectar discursos de ódio. Segundo o dicionário [7], "discurso de ódio é o discurso que ataca, ameaça ou insulta uma pessoa ou grupo com base na origem nacional, etnia, cor, religião, gênero, identidade de gênero, orientação sexual ou deficiência".

O reconhecimento de emoções se concentra na extração de um conjunto de rótulos de emoções e a detecção de polaridade é geralmente uma tarefa de classificação binária com saídas como "positivo" versus "negativo"[2]. Além disso, a análise de sentimentos usa métodos de PNL, estatísticas ou aprendizado de máquina para extrair, identificar ou caso contrário, caracterize o conteúdo de sentimento de uma unidade de texto. Às vezes é referido como mineração de opinião, o que foi discutido em Dave et al. [4].

Em [25], os autores apresentam uma abordagem para detectar discurso de ódio (anti-semita, antinegro, anti-asiático, anti-mulher, anti-muçulmano, anti-imigrante ou outro-ódio), em texto online e observaram que o ódio contra cada grupo diferente é tipicamente caracterizado pelo uso de um pequeno conjunto de palavras estereotipadas de alta frequência. Eles trabalharam em discurso de ódio antissemita, usando a representação BOW (bag-of-words) de comentários de usuários e treinaram um algoritmo de aprendizado SVM. No artigo [14], os autores aplicaram uma abordagem de aprendizado de máquina supervisionado, empregando dados rotulados adquiridos de forma barata de diversas contas do Twitter para aprender um classificador binário para os rótulos "racista" e "não racista", aplicando um modelo BOW.

Uma maneira clássica de se identificar o discurso de ódio é usar uma abordagem baseada em palavras-chave, conforme descrito em [23]. Para os autores, com base em uma ontologia ou dicionário, textos que contenham potencial discursos ódio podem ser identificados. Um exemplo considerado pelos autores é o vocabulário Hatebase. Eles ponderam que apenas o uso do termo presente no texto não é suficiente para definir a presença do discurso de ódio. Além disso, segundo eles, a utilização dessa abordagem exclusivamente não permite identificar discurso de ódio em textos que não contenham quaisquer palavras-chave de incitação ao ódio, como é o caso do uso de linguagem figurada.

SGDClassifier

SGD é um classificador linear, ou seja, assume que os dados se comportam linearmente. Este classificador utiliza da abordagem Stochastic Gradient Descent (SGC), que é uma técnica de otimização para treinar classificadores lineares e regressores. O Classificador SGD implementa modelos lineares regularizados com descida de gradiente estocástico. A descida do gradiente estocástico considera apenas 1 ponto aleatório enquanto muda pesos ao contrário da descida gradiente que considera todos os dados de treinamento. Como tal, a descida de gradiente estocástico é muito mais rápida do que a descida gradiente ao lidar com grandes conjuntos de dados.

Classificação de Texto

A classificação de texto, tem como objetivo classificar os textos de interesse em classes ou categorias. Um sistema de classificação de texto consiste principalmente em um mecanismo de extração de recursos que calcula informações numéricas de um documento de texto bruto e um classificador que executa um processo de classificação usando conhecimento prévio dos dados rotulados [24]. Considere D = {d1,d2,...,dn} como um conjunto de documentos e C = {c1,c2,...,cn} como o conjunto de categorias (classes), a tarefa de classificação de texto consiste em atribuir para cada par (ci,dj) de C x D um valor de 0 ou 1, ou seja, 1 se o documento dj for atribuído à classe cj e 0 em caso contrário [26].

É uma parte essencial de muitas aplicações de processamento de linguagem natural, incluindo análise de sentimento, resposta a perguntas e categorização e agrupamento de textos etc.

Uma estrutura típica de classificação de texto, consiste em pré-processamento, extração de recursos, seleção de recursos e estágios de classificação.

METODOLOGIA

Diversas ferramentas estão disponíveis no mercado para a realização de análises descritivas dos dados e aplicação de algoritmos de aprendizagem de máquina.

Para o desenvolvimento da mineração de dados, foram utilizadas as ferramentas *Python* por ser uma linguagem flexível e com enormes bibliotecas para manipulação de dados, e o software *Google Colab*. A metodologia aplicada foi a CRISP-DM que é utilizada justamente em projetos que envolvem grandes quantidades de dados e processamentos. Essa metodologia é composta por seis etapas: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem dos Dados, Avaliação e Implementação.

Seguindo este conceito, foram realizadas as seguintes etapas:

Entendimento do Negócio

Tendo em vista que as redes sociais são utilizadas por usuários mal intencionados como um ambiente para propagar ódio e violência contra a comunidade LGBTQIA+, é dado como objetivo a identificação e denúncia de textos e publicações feitas por estes usuários. O foco do estudo é na detecção de casos considerados homofóbicos.

Entendimento dos Dados

Nesta etapa é onde de fato se inicia o trabalho de mineração de dados, onde foram selecionados termos relacionados a indivíduos homossexuais de dentro da comunidade LGBTQIA+ ou que têm referências a esses indivíduos. Todos os termos foram considerados como o mesmo peso dentro da pesquisa, logo uma não influencia mais que a outra nesse cenário. Com a definição das palavras, foi iniciada a primeira coleta dos dados, sendo feita através da API do Twitter, que permite a busca por textos publicados por um perfil público/aberto, por meio de uma palavra chave. O tempo de coleta dos dados foi em um período de 15 dias, coletando 3243 tweets brasileiros. Por fim, esses tweets foram salvos juntamente com sua quantidade de curtidas e retweets para uma análise de alcance e engajamento.

Preparação dos Dados

Com os dados coletados inicia-se a fase que antecede a fase de criação do modelo, aqui é onde preparamos os dados para posteriormente treinarmos o modelo. Foi realizada a limpeza dos dados onde foi removido tudo o que possa prejudicar o entendimento da mensagem e focando no que é essencial para o desenvolvimento e seja executável para o sistema. Após limpeza e formatação, foram classificados 3243 tweets manualmente a partir de 3 diferentes opiniões, sendo uma delas de dentro da própria comunidade indicando 0 para tweet não ofensivo e 1 para ofensivo.

Modelagem

Logo após o processo de limpeza dos textos e extração do conteúdo significativo, inicia-se a fase de classificação, que por sua vez é feita com aprendizagem de máquina. O modelo para classificação dos tweets foi feito utilizando a linguagem Python que possui uma grande variedade de bibliotecas científicas para desenvolvimento de algoritmos de aprendizado de máquina e análise de dados. Para esta pesquisa utilizamos a API scikit-learn que fornece a classe *SGDClassifier* para implementar o método SGD (*Stochastic Gradient Descent*), que serve para a construção de um estimador para problemas de classificação. Nesta etapa utilizamos a função *train_test_split* para dividir os dados de aprendizado de máquina, sendo 70% para conjunto de treinamento e 30% para conjunto de testes.

Validação

Aqui é onde avaliamos os modelos para garantir que eles atendem às necessidades do negócio e a partir dos resultados decidimos se será necessário uma remodelagem.

Implantação

Esta é a etapa final onde apresentamos os resultados finais do projeto fazendo a implementação dos modelos validados.

RESULTADOS

Nas classificações dos tweets foram atribuídos dois tipos de classes, 0 e 1, onde 0 indica tweets não ofensivos e 1 para tweets considerados ofensivos.

Para visualização utilizou-se da matriz de confusão, que é uma tabela que mostra as frequências de classificação para cada classe do modelo.

- Verdadeiro positivo (VP): ocorre quando no conjunto real, a classe que estamos buscando foi prevista corretamente. Por exemplo, quando o texto é ofensivo e o modelo previu corretamente que o texto é ofensivo.
- Falso positivo (FP): ocorre quando no conjunto real, a classe que estamos buscando prever foi prevista

incorretamente. Exemplo: o texto não é ofensivo, mas o modelo previu que o texto é ofensivo.

- Verdadeiro negativo(VN): ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista corretamente. Exemplo: o texto não é ofensivo e o modelo previu corretamente que o texto não é ofensivo.
- Falso negativo (FN): ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista incorretamente. Por exemplo, o texto é ofensivo, mas o modelo previu que o texto não é ofensivo.

A partir da matriz de confusão foram analisadas 4 métricas de avaliação, sendo elas: Precisão, Recall, F1-Score e Acurácia. As fórmulas matemáticas para cada uma dessas métricas podem ser visualizadas na imagem 1.

- Acurácia: indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente;
- Precisão: dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas;
- Recall: dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas;
- F1-Score: média harmônica entre precisão e recall.

Para o nosso objetivo analisamos que a métrica que mais importa é a precisão, pois ela é usada em uma situação em que os Falsos Positivos são considerados mais prejudiciais que os Falsos Negativos.

Como resultado, obteve 93% de precisão na classe 0, no qual é um número bastante aceitável. Já na classe 1 obtivemos apenas 62% de precisão (veja na tabela 2), pois a quantidade de classe 1 no momento de treinamento e teste era demasiadamente menor à classe 0, ocorrendo um desbalanceamento. Apesar disso, o algoritmo obteve 86% de acurácia nos resultados finais e mesmo com a baixa precisão da classe 1, a alta precisão da classe 0 cumpre com o nosso objetivo que é diminuir ao máximo os Falsos Positivos (quando temos apenas 7%). Futuramente pretendemos coletar mais dados de classe 1 para balanceamento e provavelmente obtermos melhor precisão para a classe 1.

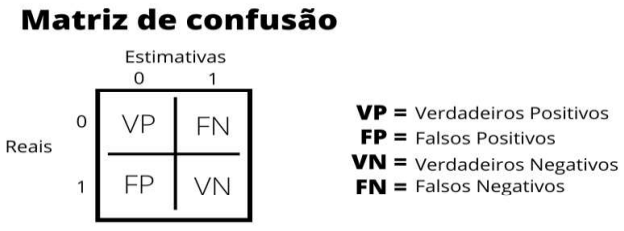

Imagem 1 - Fórmulas das métricas aplicadas no estudo

Classe	Precisão	Recall	F1-score
0	0.93	0.90	0.91
1	0.62	0.70	0.66
Acurácia			0.86

Tabela 2 - Resultados obtidos do modelo

CONCLUSÃO

De acordo com os estudos realizados foi possível criar um modelo preditivo com razoável precisão. O objetivo é que o modelo seja capaz de denunciar conteúdo de cunho homofobico no twitter, atualmente quando se realiza uma denúncia no twitter a mesma passa por uma avaliação antes que seja aceita e o conteúdo sofra alguma punição de fato. Contudo o modelo se mostrou bem promissor em eliminar Falsos Positivos, ou seja, erra pouco ao fazer denúncias em algum conteúdo que não seja o alvo de fato, e mesmo que aconteça, o erro não é tão grave levando em consideração que essa denuncia não passara pela análise do twitter, então usuários que publicarem algum conteudo com cunho não homofobico que acabe sendo julgado de forma precipitada pelo modelo, não será penalizado. A baixa precisão da classe 1 remete ao desbalanceamento entre as classes, algo a ser corrigido no futuro. Após a validação do modelo foi feita sua implantação e já sendo auxiliado pela classificação do modelo foi realizada a análise do engajamento de conteúdos. Os resultados podem ser observados como *reports* nas imagens 3 e 4 a seguir. Felizmente, conteúdos de cunho homofobico vem tendo menos engajamento, apesar de ainda possuírem alguns números relevantes, mas isso se deve a popularidade de certos perfis que publicam ou compartilham o conteúdo.

Imagem 2 - Legenda das imagens a seguir

Imagem 3 - ***Favorites report***

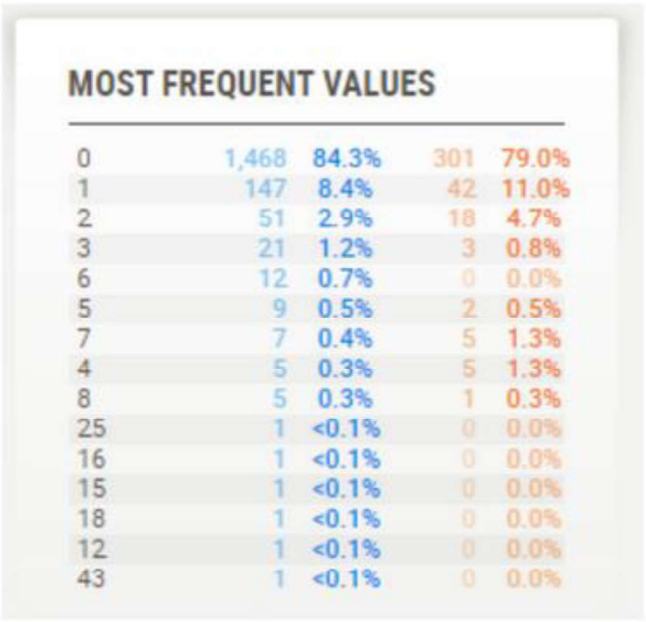

Imagem 3.1

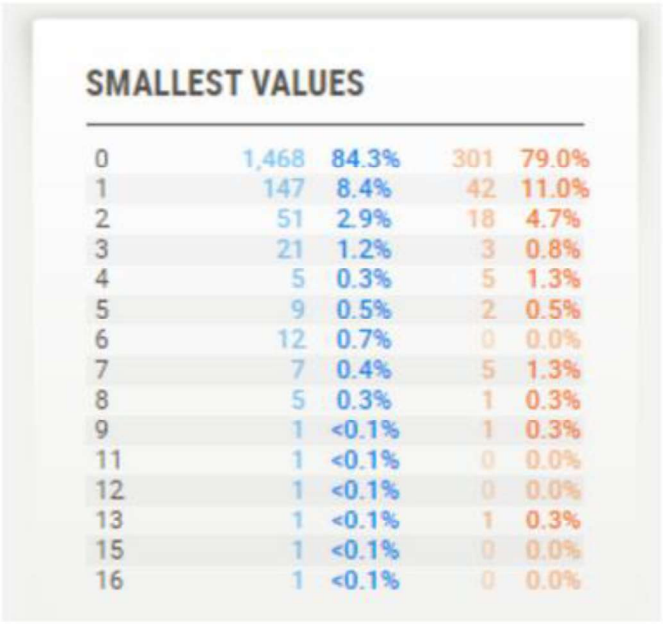

Imagem 3.2

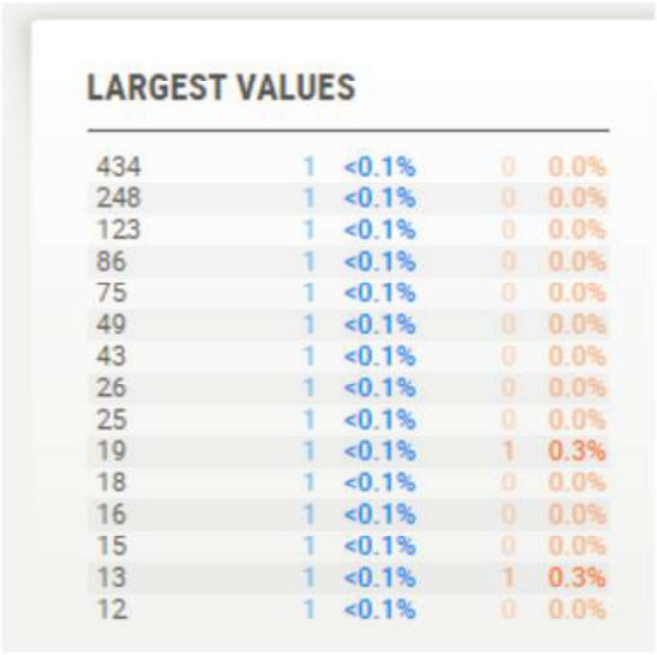

Imagem 3.3

RETWEETS report

MAX	7,055	7,404
95%	5,368	2,989
Q3	5,350	169
AVG	1,431	459
MEDIAN	6	0
Q1	0	0
5%	0	0
MIN	0	0

Imagem 4 - *Retweets report*

SMALLEST VALUES

0	743	42.7%	222	58.3%
1	77	4.4%	25	6.6%
2	15	0.9%	11	2.9%
3	18	1.0%	1	0.3%
4	9	0.5%	6	1.6%
5	5	0.3%	6	1.6%
6	21	1.2%	0	0.0%
8	10	0.6%	0	0.0%
9	3	0.2%	0	0.0%
10	14	0.8%	2	0.5%
11	1	<0.1%	0	0.0%
13	14	0.8%	0	0.0%
14	2	0.1%	0	0.0%
17	19	1.1%	0	0.0%
18	5	0.3%	4	1.0%

Imagem 4.2

MOST FREQUENT VALUES

0	743	42.7%	222	58.3%
1	77	4.4%	25	6.6%
371	53	3.0%	0	0.0%
5368	52	3.0%	0	0.0%
5354	51	2.9%	0	0.0%
5350	51	2.9%	0	0.0%
5369	48	2.8%	0	0.0%
50	45	2.6%	0	0.0%
45	42	2.4%	0	0.0%
61	42	2.4%	0	0.0%
5366	41	2.4%	0	0.0%
5363	41	2.4%	0	0.0%
5357	40	2.3%	0	0.0%
5360	39	2.2%	0	0.0%
5365	38	2.2%	0	0.0%

Imagem 4.1

LARGEST VALUES

7055	1	<0.1%	0	0.0%
5609	1	<0.1%	0	0.0%
5369	48	2.8%	0	0.0%
5368	52	3.0%	0	0.0%
5366	41	2.4%	0	0.0%
5365	38	2.2%	0	0.0%
5363	41	2.4%	0	0.0%
5360	39	2.2%	0	0.0%
5357	40	2.3%	0	0.0%
5354	51	2.9%	0	0.0%
5352	37	2.1%	0	0.0%
5350	51	2.9%	0	0.0%
4421	1	<0.1%	0	0.0%
2989	11	0.6%	8	2.1%
2476	1	<0.1%	0	0.0%

Imagem 4.3

TRABALHOS FUTUROS

Existe a pretensão de continuar com a coleta de tweets, buscar uma alta gama de dados para tornar o modelo cada vez mais preciso, inicialmente focar no balanceamento das classes, buscar textos a serem rotulados com a classe minoritária. Experimentar técnicas de *VotingClassifier* onde se combina vários modelos treinados, avalia a saída de cada um e se estima a melhor saída possível, é uma técnica válida para melhorar o desempenho do modelo, e com isso se obter maior precisão nas denúncias.

REFERÊNCIAS

1. Iribure, A. e Gonçalves Jardim, P. 2021. #QUEMLACRANÃOLUCRA (MESMO): DISPUTAS ENTRE A MARCA BURGER KING E USUÁRIOS DO TWITTER. TROPOS: COMUNICAÇÃO, SOCIEDADE E CULTURA (ISSN: 2358-212X). 10, 2 (dez. 2021).
2. E. Cambria, "Affective computing and sentiment analysis," IEEE Intelligent Systems, vol. 31, pp. 102–107, Mar 2016.
3. Aquino, M.C. e Schuch, C. 2022. "E olha que eu sou virtual!": Um estudo sobre as marcas Magazine Luiza e Natura e seus posicionamentos sobre questões de gênero no Twitter . Culturas Midiáticas. 16, (maio 2022), 22. DOI:https://doi.org/10.22478/ufpb.2763-9398.2022v16n.61889.
4. K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of the 12th International Conference on World Wide Web, WWW '03, (New York, NY, USA), pp. 519–528, ACM, 2003.
5. DataTechNotes. (2020) SGD Classification Example with SGDClassifier in Python. Retrieved September 19, 2022 from https://www.datatechnotes.com/2020/09/sgd-classification-example-with-sgdclassifier-in-python.html
6. PALMA, Rodrigo Demetrio. Análise comparativa entre aprendizado supervisionado e aprendizado por transferência aplicados a análise de sentimentos em textos. 2020. 48 f., il. Trabalho de Conclusão de Curso (Bacharelado em Engenharia da Computação)—Universidade de Brasília, Brasília, 2020.
7. Dictionary.com. Retrieved September 28, 2022 from http://www.dictionary.com/browse/hate-speech? s=t.
8. Provost, F.; Fawcett, T. Data science and its relationship to big data and data-driven decision making. Big Data, v. 1, n. 1, p. 51–59, 2013. PMID: 27447038. Retrieved September 10, 2022 from https://www.liebertpub.com/doi/10.1089/big.2013.1508.
9. Alessandra, G., Dennys, A. e Thiago, O. (2019) "Drag queens e Inteligência Artificial: computadores devem decidir o que é 'tóxico' na internet?". Retrieved September 02, 2022 from https://internetlab.org.br/pt/noticias/drag-queens-e-inteligencia-artificial-computadores-devem-decidir-o-que-e-toxico-na-internet/
10. Leandro, G. Você sabe o que é metodologia CRISP-DM? Descubra aqui. Retrieved September 13, 2022 from https://www.knowsolution.com.br/voce-sabe-o-que-e-metodologia-crisp-dm-descubra-aqui
11. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825-2830, 2011.
12. Pereira, V. Gomes. "Using supervised machine learning and sentiment analysis techniques to predict homophobia in portuguese tweets." PublishedVersion, reponame:Repositório Institucional do FGV, 2018. http://hdl.handle.net/10438/24301.
13. PARMELEE, John H.; BICHARD, Shannon L. Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public. Maryland: Lexington Books, 2012.
14. Irene Kwok and Yuzhou Wang. 2013. Locate the hate: detecting tweets against blacks. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI'13). AAAI Press, 1621–1622.
15. Lil, H, Mess. (2017) "Facebook's Hate Speech Policies Censor Marginalized Users", retrieved September 11, 2022 from https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/
16. Eduardo, M. "Quem a homotransfobia matou hoje?". Retrieved September 20, 2022 from https://homofobiamata.wordpress.com/quem-somos-3/homofobia-e-crime/.
17. Vinay, Patlolla. (2017) "How to make SGD Classifier perform as well as Logistic Regression using parfit", retrieved October 4, 2022 from https://towardsdatascience.com/how-to-make-sgd-classifier-perform-as-well-as-logistic-regression-using-parfit-cc10bca2d3c4
18. Rossetto, G.P., Carreiro, R. e Almada, M.P. 2013. Twitter e comunicação política: limites e possibilidades. Compolítica. 3, 2 (dez. 2013), 189-216. DOI:https://doi.org/https://doi.org/10.21878/compolitica.2013.3.2.49.
19. Boulic, R. and Renault, O. (1991) "3D Hierarchies for Animation", In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons ltd., England.
20. Wirth, R.; Hipp, J. Crisp-dm: Towards a standard process model for data mining. In: CITESEER.

Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. [S.l.], 2000. p. 29–39.

21. Vinícius Santos, Felipe Henriques, and Gustavo Guedes. 2022. O Discurso de Ódio Homofóbico no Twitter a partir da Análise de Dados. In Anais do XI Brazilian Workshop on Social Network Analysis and Mining, July 31, 2022, Niterói, Brasil. SBC, Porto Alegre, Brasil, 109-120.
22. Zago, G. S. and Bastos, M.T. 2013. Visibilidade de Notícias no Twitter e no Facebook: Análise Comparativa das Notícias mais Repercutidas na Europa e nas Américas. Brazilian journalism research. 9, 1 (Jun. 2013), 116–133.
23. MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019).Hate speech detection: Challenges and solutions.PLOS ONE, 14(8):1–16. Retrieved September 28, 2022 from https://doi.org/10.1371/journal.pone.0221152.
24. GÜNAL, S. Hybrid feature selection for text classification. Turkish Journal of Electrical Engineering and Computer Sciences, v. 20, n. SUPPL.2, p. 1296–1311, 2012.
25. W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the Second Workshop on Language in Social Media, LSM '12, (Stroudsburg, PA, USA), pp. 19–26, Association for Computational Linguistics, 2012.
26. Ikonomakis, M., Sotiris B. Kotsiantis and Vasilis T. Tampakas. "Text Classification Using Machine Learning Techniques." (2005).
27. Imran, A. "Social Media's Role in Amplifying Dangerous Lies About LGBTQ+ People". August, 2022. Center for Countering Digital Hate Inc.
28. Matt, A. (2022) "50 + TWITTER ESTATÍSTICAS E FATOS PARA 2022". Retrieved October 07, 2022 from https://www.websiterating.com/pt/research/twitter-statistics/

.