Classificação de Insultos em Mensagens LGBTQfóbicas no Twitter

Vinícius Matheus M. S[1]. Coutinho, Yuri Malheiros

Departamento de Ciências Exatas - Universidade Federal da Paraíba (UFPB)
CEP - Rio Tinto - PB - Brasil

{vinicius.matheus,yuri}@dcx.ufpb.br

Abstract. *With the popularization of social media in the last decade the way and structure of means of communication in society has been strongly changed. Therefore, users feel free to share and give their opinions autonomously. However this freedom of speech can be changed and end up becoming hate speech to social minorities. The purpose of this paper is to classify the sentiment of messages from Twitter in order to classify them as LGBTQphobic or non LGBTQphobic.*

Resumo. *Com a popularização das redes sociais na última década, a forma e estrutura da comunicação na sociedade foi fortemente alterada. Neste sentido, os usuários se sentem livres para compartilhar e disseminar suas opiniões de forma autônoma. Entretanto essa liberdade de expressão pode ser distorcida e acabar se tornando discurso de ódio direcionado a minorias sociais. Este trabalho tem como propósito classificar o sentimento de mensagens retiradas do Twitter, como LGBTQfóbicas ou Não LGBTQfóbicas.*

.

Palavras-Chave: Análise de Sentimentos; Homofobia; Inteligência Artificial; Aprendizagem de Máquina; Twitter.

1. Introdução

Com o advento da Internet na virada do milênio e a popularização das redes sociais na última década, mudanças significativas no comportamento da sociedade foram percebidas, principalmente no tocante às relações de comunicação entre as pessoas. Essa complexa engrenagem modificou permanentemente a maneira de se comunicar entre as pessoas, saindo do estático e dispendioso para a rápida fluidez e interação dos meios digitais, principalmente no que se conveniou a denominar-se de rede social.

A consolidação das redes sociais trilhou um caminho de evolução e aperfeiçoamento por alguns projetos exitosos, como por exemplo, MSN, Orkut e

[1] Trabalho de Conclusão de Curso (TCC) na modalidade Artigo apresentado como parte dos pré-requisitos para a obtenção do título de Bacharel em Sistemas de Informação pelo curso de Bacharelado em Sistemas de Informação do Centro de Ciências Aplicadas e Educação (CCAE), Campus IV da Universidade Federal da Paraíba, sob a orientação do professor Yuri Malheiros.

Fotolog, que deram sua contribuição para a idealização deste novo universo, porém foram ultrapassados por outras iniciativas mais ousadas, a saber: Instagram, WhatsApp, Twitter e Facebook. Dessa forma, as redes sociais são aplicações que suportam um espaço compartilhado de interesses, necessidades e metas comuns para a colaboração, partilha de conhecimento, interação e a comunicação (PETTENATI et al.,2006, BRANDTZAEG et al., 2007).

Através dessas plataformas digitais a sociedade passou a ter acesso a um universo de informações constantes, que em sua maioria não passam por nenhum crivo ou filtro social; portanto os ambientes das mídias sociais podem ser facilmente distorcidos e serem transformados em púlpitos para a disseminação do discurso e expressão de ódio.

A Internet transmite a sensação de liberdade, onde os usuários acreditam não existir barreiras ou punições, e sendo até certo ponto uma replicação do comportamento preconceituoso da realidade social em um ambiente virtual, como definem Andrade e Pischetola (2016) .

Dentro do amplo universo de opiniões e manifestações das redes sociais encontra-se de forma sistemática a disseminação dos discursos de preconceito, racismo e ódio inerente à sociedade real que se reproduzem no mundo virtual.

Meyer-pflug (2009, p. 97) define o discurso de ódio como a manifestação de “Ideias que incitem a discriminação racial, social ou religiosa em determinados grupos, na maioria das vezes, as minorias”. Além disso, Cohen-Almagor (2000) pontua que o discurso de ódio vai além da ofensa, mas que este tem o poder de silenciar os grupos minoritários alvos, subjugando-os ou rebaixando-os em relação ao seu direito de igualdade.

Entre as maiores e mais acessadas redes sociais em atividade no Brasil está o *Twitter* que surgiu em 2006 com o conceito de *microblogging,* o qual possui o objetivo de que seus usuários divulguem o que estão fazendo, comentando ou discutindo. Restringindo as postagens em 280 caracteres, esta rede social proporciona a divulgação em tempo real de informações entre seus interlocutores (NASCIMENTO et al., 2013).

A facilidade de compartilhamento e divulgação de informações e opiniões é um dos fatores determinantes para o crescimento do Twitter. Pela proposta do Twitter em incentivar o debate direto e objetivo, permitindo aos usuários emitir, acessar e confrontar suas ideias, seus posicionamentos e seu juízo, em relação a cada tema proposto ou debatido entre os usuários.

Um dos grandes equívocos dos usuários das redes sociais é acreditar que ao adentrar ao mundo virtual o mesmo estará livre das conveniências sociais que norteiam as relações entre as pessoas. O senso comum criou no imaginário popular que os usuários de redes sociais a ideia de anonimato, esta falsa ideia propiciou a intensificação de postagens com conteúdo que evidencia preconceito, racismo e ódio direcionados a minorias sociais.

Para a orientação deste artigo focou em mensagens de intolerância associada a comunidade LGBTQ (Lesbicas, Gays, Bissexuais, Transesuxais e Queer [2]). De acordo com o relatório do Grupo Gay da Bahia[3] , no ano de 2018 foram registradas 420 mortes decorrente da discriminação. Sendo destas, 320 homicídios (76%) e 100 suicídios (24%).

Ainda ressaltando que a cada 20 horas um LGBTQ é morto ou se suicida vítima da LGBTQfobia, caracterizando o Brasil como líder mundial de crimes contra minorias sexuais, sendo responsável por mais da metade dos homicídios de transexuais no mundo. De acordo com a organização internacional Trans Respect[4], no período de janeiro de 2008 a junho de 2016 o Brasil foi responsável por 868 homicídios contra a comunidade transexual.

Com a grande popularização do Twitter e o seu grande volume de dados diários, cerca de 500 milhões de tweets, o mesmo será usado como objeto de estudo para a análise de discursos de ódio e insultos LGBTQfóbicos na Internet. E para isso será usada a área de análise de sentimentos ao qual funde diversas áreas

---

[2] Queer - Queer é um termo abrangente para minorias sexuais e de gênero que não são heterossexuais ou cisgêneros. Queer foi originalmente usado pejorativamente contra pessoas com desejos do mesmo sexo, mas, a partir do final da década de 1980, acadêmicos e ativistas queer começaram a reclamar a palavra.
https://ok2bme.ca/resources/kids-teens/what-does-lgbtq-mean/. Acessado em 20 jul. 2019.

[3] "Grupo Gay da Bahia - Quem a homofobia matou hoje - WordPress.com." https://homofobiamata.files.wordpress.com/2019/01/relatorio-2018-1.pdf. Acessado em 21 jul. 2019.

[4] "Transrespect vs Transphobia." https://transrespect.org/en/. Acessado em 21 jul. 2019.

da tecnologias tais como: inteligência artificial, recuperação de informação e mineração de dados (BECKER e TUMITAN, 2013).

O objetivo deste trabalho é classificar sentimentos no universo dos tweets publicados no Brasil que possuam em sua composição palavras ou insultos de cunho ou raízes homofóbicas como LGBTQfóbico ou não. Estando atento a hipótese de que uma mensagem LGBTQfóbica é composta por uma palavra-insulto e acompanhada por sentimento negativo.

Para testar esta hipótese, foi selecionado um conjunto de palavras-insulto e através disto foram coletados tweets que possuíam estas palavras em sua composição. Estes tweets foram salvos em um banco de dados para que posteriormente fossem classificados como negativos, positivos ou neutros de acordo com um conjunto de dados rotulados (SILVA, et al., 2019).

Após a classificação automática do algoritmo, foi realizada a validação dos resultados através de aplicação de formulários via google forms, onde o entrevistado classificou um conjunto de dezesseis tweets como LGBTQfóbicos ou não. Estes resultados serão usados como comparativo e estudo ao resultado obtido pelo algoritmo.

Este trabalho está estruturado da seguinte forma: Seção 2 descreve a fundamentação teórica com os conceitos de análise de sentimentos, aprendizagem de máquina e redes sociais; A Seção 3 apresenta a metodologia aplicada para a coleta de dados e classificação dos tweets; A Seção 4 traz análises e discussões dos resultado obtidos e na Seção 5 a conclusão.

2. Fundamentação teórica

Nesta seção serão abordados os conceitos que fundamentam este trabalho, estando subdivididos em, Redes Sociais, Análise de Sentimentos e Aprendizagem de Máquina.

2.1 Redes Sociais

Para Recuero (2009) uma rede social pode ser caracterizada como um conjunto constituído por dois elementos, os atores (pessoas, instituições ou grupos) e as conexões (interações). Partindo desse ponto, uma rede social é entendida por inúmeras interações entre pessoas ou grupos, a fim de que os atores compartilhem informações de forma independente e autônoma dentro desta rede. Como explicam Garton, Haythornthwaite e Wellman (1997:1), *"quando uma rede de computadores conecta uma rede de pessoas e organizações, é uma rede social".*

Com a popularização das redes sociais através dos smartphones, atualmente existem inúmeros e diferentes tipos de redes sociais, cada qual com sua finalidade e identidade, onde estas reúnem usuários que possuem interesses em comum; desta forma criando novas estruturas sociais, onde o fluxo de informação é mediado por um celular e gerando impactos benéficos ou não nos usuários que se conectam nesta rede.

2.2 Análise de Sentimentos

De acordo com Liu (2012) a análise de sentimentos ou mineração de sentimentos pode ser definida como a área de estudo que analisa opiniões, sentimentos, avaliações e emoções em relação a entidades, tais como, produtos, serviços e pessoas.

Liu (2012) pontua que existem três tipos de níveis de análise de sentimentos, devido ao grau granularidade do texto. Sendo estes:

- *Documento:* Este nível de análise tem como foco classificar a opinião geral do documento com os sentimentos de positivo ou negativo.
- *Sentença:* A tarefa desse nível foca em sentenças específicas de um documento e determinando se esta sentença expressa sentimentos positivos, negativos ou neutros. Comumente usado para analisar diferentes opiniões de um documento, focando em suas sentenças.
- *Entidade e Aspecto:* Este nível de análise possui um maior grau de granularidade de classificação, que ao invés de focar nos construtores de linguagens (documentos, parágrafos, sentenças), este estará atento a opinião em si. Baseando que uma opinião é formada por um sentimento e um alvo.

Por exemplo “Eu adoro o carnaval de Olinda, apesar de não me sentir bem em multidões”, observe que nesta frase possui o sentimento geral positivo, mas percebemos que ela não é totalmente positiva. Veja que o sentimento positivo se encontra no alvo principal da frase que é “Carnaval de Olinda” (enfatizado), mas negativo em relação a “multidões” (não enfatizado).

Diante dos três níveis destacados por Liu, a análise e classificação deste presente trabalho se encaixa no primeiro nível destacado, considerando o tweet como um **documento***,* buscando classificar o sentimento geral como positivo ou negativo.

2.2.1 Técnicas de Classificação

Assim sendo, existem duas técnicas principais para realizar classificação de sentimento, sendo estas, **classificação baseada em lexicon** e **classificação utilizando aprendizagem de máquina.**

A primeira se caracteriza como o uso de palavras individuais ou expressões que possuem sentimentos bons como, maravilhoso e incrível; ou sentimentos ruins como, pobre e feio, para que seja realizada a análise de sentimento. Porém este método não é totalmente eficaz por questões da variabilidade da língua. Por exemplo, as palavras homônimas perfeitas da língua portuguesa, que possuem a mesma grafia, mesmo som, porém sentidos diferentes diante do contexto. A palavra “morro” pode ter sentidos ambíguos, por exemplo: “Os alpinistas estão escalando o morro.” (monte), sentimento positivo ou neutro, porém podemos usar a mesma palavra em outra conotação, “Eu morro de medo de altura!” (verbo morrer), sentimento negativo.

A segunda forma mais usual de classificação utiliza **aprendizagem de máquina**, que usa um conjunto de treinamento com exemplos já classificados para que padrões sejam aprendidos tornando possível a classificação de novas entradas. Esta abordagem será melhor explanada na próxima seção, onde a mesma serve como base desta pesquisa.

2.3 Aprendizagem de Máquina

Alpaydin (2010) define aprendizagem de máquina (AM) como uma otimização computadorizada de um critério usando dados como exemplos ou experiências recorrentes, desta forma, tomando decisões através das experiências acumuladas a fim de apresentar uma solução. Alpaydin afirma que o papel da aprendizagem de máquina está subdividido em duas categorias, sendo a primeira, o treinamento e execução, que faz o uso de algoritmos para resolução de problemas de otimização e o segundo, quando o modelo é compreendido após uma série de treinamentos, este deverá apresentar uma solução algorítmica eficiente.

O uso da aprendizagem de máquina é requisitado quando as estruturas e técnicas de programação clássicas não são viáveis, pois existem problemas computacionais que possuem padrões extremamente complexos, como por exemplo a classificação de imagem, sendo difícil identificar padrões. Daí, a aprendizagem de máquina, através do processo de treinamento, consegue descobrir esses padrões de forma mais eficiente do que um programador poderia fazer.

Um exemplo é classificação de e-mails em duas categorias: spam e não-spam. Para isso os algoritmos de aprendizagem de máquina irão processar milhares de exemplos reais de e-mails que são considerados spam e a partir disso a máquina começará a “aprender” e compreender os padrões de uma mensagem com perfil spam.

2.3.1 Categorização

Segundo Ayodele (2010), os algoritmos de aprendizagem de máquina são organizados de acordo com o resultado desejado do algoritmo. Desta forma, é necessário antever qual tipo de resultados será fornecido de acordo com a entrada. Pode-se classificar os sistemas da seguinte maneira (Ayodele, 2010):

- **Aprendizagem supervisionada:** Este tipo de aprendizagem se caracteriza, quando o algoritmo produz uma função que mapeia entradas às saídas

desejadas. Sendo estes exemplos ou atributos conhecido, possuindo como objetivo classificar novos exemplos entre os pré- existentes.

A Figura 1 apresenta o fluxograma de como o algoritmo de aprendizagem supervisionado funciona. Nela, pode-se observar que o treinamento ocorre no momento que o programa recebe como entrada a imagem de um gato e a resposta associada a ele (dados rotulados), evidenciando se é um gato ou não. Esse treinamento se repete múltiplas vezes com diferentes tipos de imagens. Após essas repetições o programa aprende a identificar as características de uma imagem de gato, então no momento que receber um nova entrada ele pode identificar se a imagem é ou não a imagem de um gato.

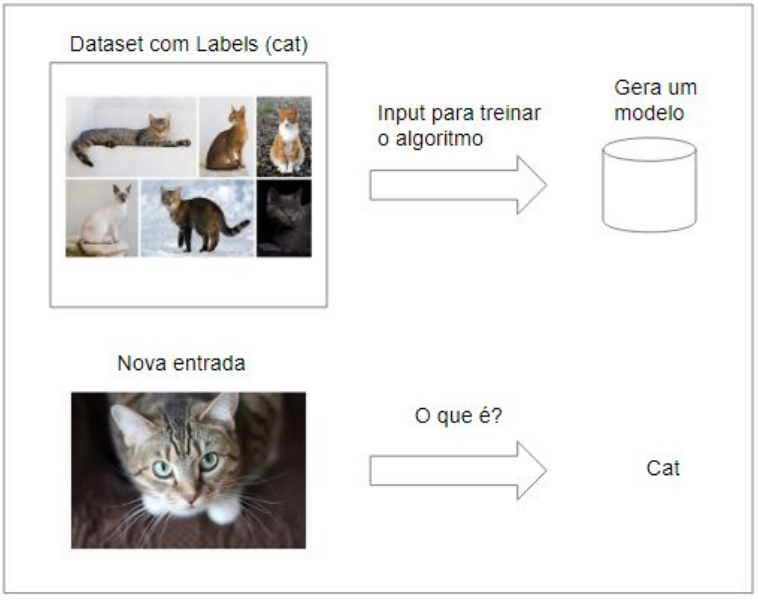

Figura 1. Fluxograma de aprendizagem Supervisionada.

- **Aprendizagem não-supervisionada**: É o algoritmo que modela um conjunto de entradas que não possuem valores de saída desejada (não rotulados), e desta forma a classificação é realizada de acordo com as similaridades entre os exemplos.

A Figura 2 apresenta um agrupamento de imagens. O modelo de aprendizado entendeu como classificar imagens entre dois grupos distintos, sendo o primeiro dele rostos e o segundo casas. Ao receber uma nova

entrada (imagem), o modelo observa os atributos e identifica a qual grupo este pertence.

Figura 2. Exemplo de aprendizagem não-supervisionada.

Tendo em vista as duas categorias apresentadas por Ayodele, a pesquisa irá usar a **aprendizagem de máquina supervisionada**, na qual é utilizado um conjunto de dados mapeados (SILVA et al., 2019), que fornece um resultado conhecido ou previsto.

3. Metodologia

Para obtenção dos resultados dessa pesquisa foram realizados os seguintes passos: (i) seleção de palavras de cunho LGBTQfóbica, (ii) coleta de tweets através das palavras-chave previamente selecionadas, (iii) uso de classificadores de sentimentos, (iv) validação, (v) análise dos resultados. A Figura 3 ilustra os passos desenvolvidos nesta pesquisa.

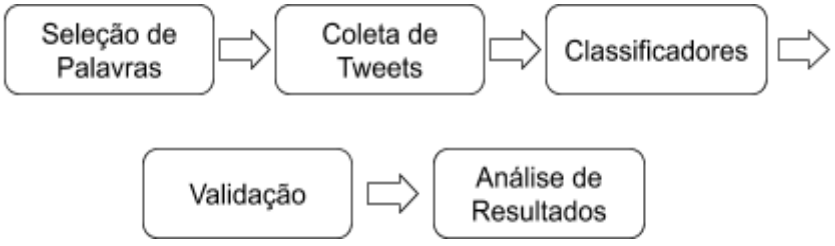

Figura 3. Fluxograma dos passos da pesquisa

3.1 Coleta de Dados

O primeiro passo para obtenção do objetivo desta pesquisa foi a seleção de palavras com raízes LGBTQfóbicas, foram selecionadas 4 palavras com conotação ofensiva a comunidade LGBTQ, cada palavra-insulto escolhida tem como alvo uma parte da comunidade LGBTQ. O insulto “Gay”, foi selecionado com o objetivo de coletar tweets de forma mais abrangente, incluindo todos da comunidade; Já o insulto “Sapatão” foi selecionado pois foi considerado a palavra-insulto mais comum para prática de lesbofobia; “Traveco” insulto direcionado a comunidade transgênero e por fim “Viado” que é a palavra homofóbica mais usada e conhecida para a prática da homofobia.

Palavra-Insulto	Quantidade de Tweets Coletados
Gay	1.166
Sapatão	1.957
Traveco	1.267
Viado	3.959
TOTAL	8.349

Tabela 1. Quantidade de Tweets por Palavra-Insulto

Estas palavras serviram para a coleta de tweets, realizada através da API do Twitter, que permite a busca de mensagem por palavras-chave. Por fim, estes tweets foram salvos numa base de dados local, que serviu para os estudos deste trabalho. A coleta foi realizada durante cinco dias no período de 27 de julho de 2019 a 31 de julho de 2019, coletando 8.349 tweets. A Tabela 1 apresentada as quantidades de tweets coletados por palavras-chave e na Tabela 2 tem-se um exemplo coletado para cada palavra-chave.

Tweet Coletado	Palavra-Chave
“To querendo virar sapatao, porém e se a mina for paranóica igual eu?”	Sapatão
“Gente pq usam o termo traveco pra ofender alguém como "pessoa feia"Vcs já viram um traveco? É mais bonito que eu kkkkkk”	Traveco
“ Messi ta virando viado, que papo de vacilão ”	Viado
“o caso do gay que acha que por ser gay não tem como ser machista”	Gay

Tabela 2. Exemplos de Tweets Coletados.

3.2 Classificação de Dados

Diante a hipótese apresentada por esse trabalho, ao qual uma mensagem que tenha o caráter LGBTQfóbica (tweet) seja composta por uma palavra-insulto e que possua o sentimento negativo são de fato mensagens homofóbicas, foi realizada classificação das mensagens com o auxílio da aprendizagem de máquina.

A classificação dos tweets foi realizada fazendo-se o uso da linguagem de programação Python que possui uma grande variedade de bibliotecas científicas para o desenvolvimento de algoritmos e análise de dados. Para esta pesquisa foi utilizada a biblioteca Scikit Learn, onde esta integra diferentes algoritmos de aprendizado de máquina para problemas supervisionados e não supervisionados (PEDREGOSA et al., 2011).

Utilizando um conjunto de 2787 mensagens do Twitter (tweets) previamente classificados por pessoas quanto aos seus sentimentos (positivos, negativos ou neutros) (SILVA et al., 2019), foi possível classificar os sentimentos (positivo ou negativo) de um novo conjunto de mensagens não rotulados coletados por esta pesquisa. Para tal, este trabalho considerou as classificações positivas e neutras como pertencentes ao conjunto dos positivos.

Após realizada a classificação do novo conjunto pelo algoritmo foi necessária a validação destes resultados com pessoas reais, tendo em vista perceber em qual

contexto a palavra-insulto está incluída, pois pode ocorrer situações em que o tweet composto com a palavra-insulto esteja classificado como negativo, mas que não está incluso em um contexto pejorativo, ou seja normalizado e diante aos resultados dos entrevistados poder identificar a taxa de acerto do algoritmo. A Tabela 3 a seguir exibe alguns resultados de classificação obtidos pelo algoritmo.

Tweet Coletado	Sentimento	Palavra-Chave
"A mulher parece sapatão de qualquer forma, não adianta mudar o personagem."	Negativo (-1)	Sapatão
"@ Coisa horrível... sem conteúdo... traveco"	Negativo (-1)	Traveco
"brega e xote melhores tipo de música quem discorda eh viado"	Positivo (1)	Viado
"putz que bonito esse menino barbado e de óculos quadrado, será se é gay?"	Neutro (0)	Gay

Tabela 3. Exemplos de Classificações do Algoritmo

3.3 Validação dos Resultados

A validação dos resultados foi realizada por formulários via *google forms,* nos quais os entrevistados deveriam responder sobre sua orientação sexual e identidade de gênero e posteriormente responder de acordo com sua interpretação se um tweet tinha sentido homofóbico ou não. No total, cada participante analisou 16 tweets.

Este conjunto de dezesseis tweets é composto por tweets que contêm as quatros palavras-chaves selecionadas para esta pesquisa (gay, sapatão, traveco e viado). Neles, cada palavra-insulto possui um conjunto de quatro tweets que estão dispostos da seguinte forma: dois tweets negativos considerados como homofóbicos e um positivo e um neutro como não homofóbicos. Assim sendo dois homofóbicos e dois não homofóbicos.

Para melhores resultados foram produzidos dez formulários, cada um com um conjunto de tweets diferentes, totalizando assim 160 perguntas. A aplicação destes formulários foi realizada através de um link, que redirecionava o entrevistado aleatoriamente para um dos dez formulários. A pesquisa foi realizada com os alunos e professores dos cursos de Sistemas de Informação e Licenciatura em Ciência da Computação da Universidade Federal da Paraíba - Campus IV via grupo de emails dos alunos e professores dos cursos, no período dos dias 23 a 28 de agosto de 2019.

4. Resultados e Discussões

Após a aplicação dos formulários, foram obtidas respostas de 57 alunos e professores dos cursos de Sistemas de Informação e Licenciatura em Ciência da Computação. A seguir será apresentada a análise e discussão desses resultados .

4.1 Informações Pessoais

Os primeiros dados a serem analisados, são sobre as informações pessoais dos entrevistados. Estas informações mostram o panorama entre as identidades de gêneros e orientação sexual dos alunos e professores dos curso de SI e LCC.

A Tabela 3 exibe que de acordo com a identidade gênero 46 (80,7%) dos entrevistados são do sexo masculino, 10 (17,5%) são do sexo feminino e 1 (1,8%) se auto declarou como não binário. Estes dados evidenciam que os cursos de tecnologia do Campus IV são compostos majoritariamente pelo sexo masculino.

Em relação à orientação sexual temos que os pesquisados em sua maioria, 44 (77,2%) são heterossexuais, 5 homossexuais (8,8%), 6 bissexuais (10,5%) e 2 assexuais (3,5%).

Gênero/Orientação	Heterossexual	Homossexual	Bissexual	Assexual	Total
Masculino	36	4	5	1	46
Feminino	8	1	1	0	10

Não-Binário	0	0	0	1	1
TOTAL	44	5	6	2	57

Tabela 4. Informações Pessoais

4.2 Taxa de Acerto Geral do Algoritmo

Para calcular a taxa de acerto do algoritmo produzido foi necessário comparar a classificação do algoritmo com os resultados obtidos através do formulários, desta forma considerando o acerto quando a maioria dos participantes votaram na mesma classificação do algoritmo.

Ao analisar os resultados foi percebido que ocorreu empate em 16% das perguntas (26) das 160 perguntas selecionadas, ou seja, os entrevistados responderam igualmente como homofóbico e não homofóbico a uma mesma pergunta. Diante deste cenário foi necessário não contabilizar esses empates e reduzir os números das perguntas para 134.

O cálculo da taxa de acerto geral foi realizado da seguinte forma, somatório das questões em que a maioria dos participantes votaram na mesma classificação do algoritmo, dividido pelo total de perguntas menos os empates. E em comparação aos resultados dos entrevistados em relação aos do algoritmo, ocorreu uma taxa de acerto geral de 64,92% (87 acertos) e taxa de erro 35,1% (47 erros).

4.2.1. Taxa de Acerto em Tweets Negativos

Tendo em vista que um tweet possivelmente LGBTQfóbico é composto por uma palavra-insulto e acompanhado pelo sentimento negativo, foram selecionados oito tweets classificados como negativos para cada formulário, totalizando oitenta tweets negativos. Esta análise busca identificar a taxa de acerto do algoritmo em relação às respostas dos entrevistados em relação aos tweets negativos e possivelmente de caráter LGBTQfóbico.

Observando os resultado foi identificado que ocorreu empate em 13,7% das perguntas (11) das 80 mensagens classificadas como negativas e diante deste cenário foi necessário não contabilizar os empates, assim então, reduzindo o

número de tweets negativos (69). Para este a cálculo foi utilizada a mesma lógica matemática da Taxa de Acerto Geral do Algoritmo e o resultou que algoritmo concordou em relação às respostas dos usuários em 66,67% ou seja, 46 mensagens das 69 estudadas possuem caráter LGBTQfóbico em relação ao entendimento dos entrevistados.

4.3 Análise Perceptiva

Compreendendo que a partir do momento que os tweets foram classificados por pessoas, foi perceptível a existência de inúmeras interpretações diante aos exemplos apresentados, revelando a subjetividade dos indivíduos em relação ao problema de classificação proposto.

Para algumas mensagens foi observado que ocorreram 26 empates, ou seja, os participantes da pesquisa não chegaram a uma conclusão majoritária e em outras situações foi observado que a diferença entre os resultados foi de apenas uma resposta, expressando pouco mais de 60% em concordância.

Nesse sentido, esta análise observou alguns exemplos desses resultados, a fim de analisar a percepção dos participantes em relação ao que eles consideram como homofóbico ou não. Além disso, foram entrevistadas três pessoas, cada qual representando um alvo dos insultos presentes nos tweets-exemplos selecionados, e foi apresentada uma das mensagens para que classificassem e justificassem sua resposta.

Estas entrevistas foram realizadas com o intuito de que as pessoas que pertencem a estas orientações sexuais opinassem sobre a conotação da mensagem apresentada, tendo em vista que a mesmas estão em uma posição mais adequada para o tal, pois se encontram em seus locais de fala e além de possuírem vivências próprias perante a discriminação sexual.

A Tabela 5 mostra os *tweets* selecionados, assim como o resultado do algoritmo e as repostas dos entrevistados que responderam ao formulário que foi aplicado entre os discentes e docentes dos cursos de LCC e SI.

Tweet	Resultado Algoritmo	Resultado das Respostas do Formulário
“to com voz de traveco q m*rda, cadê minha voz irmão”	Negativo	Empate (2/2)
“Homem eh ridículo d+ né vc fala uma coisinha e eles AIN NÃO GENERALIZA meu se não te diz respeito fica quieto vc eh viado meu filho"”	Neutro	Negativo (3/2)
"A mulher parece sapatão de qualquer forma, não adianta mudar o personagem."	Negativo	Positivo (3/2)

Tabela 5. Análise da percepção dos Tweets

4.3.1 Discussão das Respostas dos Entrevistados

Em relação ao primeiro exemplo, com comunidade alvo Transsexuais e Travesti, a entrevistada apresentou a seguinte justificativa, “Primeiro que Traveco já é um termo ofensivo, e a ideia de "voz de traveco" é uma voz aguda porém fina o que não é regra, então esse tweet tem sim lgbtfobia nele mesmo que implícita ou explícita, porque vai depender de alguns fatores externos também” (Estudante e Professora de Inglês, Transsexual). Como observado a entrevistada considera o tweet como LGBTfóbico, concordando assim com o resultado do algoritmo, pois considera o termo “Traveco” ofensivo a comunidade Transexual, mas debate que pode existir outras interpretações para “voz de traveco” e que a transfobia presente pode ser explícita ou não dependendo do contexto.

Diante ao segundo exemplo apresentado, o entrevistado apresentou a seguinte resposta, “Pela forma como ela fala como se eu não tivesse direito de ter uma opinião contrária apenas por ser homem e “viado”, independente da minha opinião está certa ou errado, dentro ou fora de contexto, liberdade de expressão tá

aí.” (Bacharel em Ciência Biológicas, Homossexual). Partindo da análise do entrevistado, ele apresenta que o tweet possui caráter homofóbico, concordando com o resultado majoritário dos entrevistados , pois o sentido do tweet mostra que pelo fato do indivíduo ser homem e homossexual ele não tem direito a opinião, independente se está correto ou não.

E por fim, a justificativa do entrevistado em relação ao terceiro exemplo foi “Sim, na minha visão é (LGBTQfóbico). Até porque essa frase, já rotula uma estética atribuída às pessoas lésbicas.” (Desenvolvedora de Software, Lésbica). Discordando dos resultados apresentados pelos entrevistados e concordando com o resultado do algoritmo, a entrevistada classifica a mensagem como LGBTQfóbica, partindo da análise que ocorre uma rotulação da estética e comportamento da comunidade lésbica, atribuindo a concepção masculina ao indivíduo desta orientação.

Através destas análises obtidas pelos entrevistados é perceptível a pluralidade de interpretações diante ao assunto apresentado, onde todos participantes deste estudo consideraram os tweets LGBTQfóbico, já os entrevistados que responderam ao formulários previamente obtiveram uma heterogeneidade de compreensões em suas respostas.

4.4 Análise Comparativa de Respostas

Para melhor compreensão das respostas dos formulários e a fim de perceber a existência de concordâncias entre os entrevistados foram feitas as seguintes análises, primeiramente observar a taxa de concordância geral entre as perguntas dos formulários, assim como a análise detalhada dos usuários baseando-se em suas identidades de gênero (masculino, feminino e não-binário) e orientações sexuais (heterossexual, homossexual e assexual).

As subseções a seguir irão abordar a taxa de **concordância geral**, o comparativo das respostas entre as **mesmas identidades de gênero** (independente da orientação sexual), assim como o comparativo entre as mesmas **orientações sexuais** (independente da identidade de gênero), além do comparativo entre **orientações sexuais divergentes** e por fim, a análise da concordância entre a **comunidade LGBTQ**.

Para obtenção dos resultados baseados em identidade de gênero e orientação sexual, os formulários em suas análises individuais deveriam seguir a condição de conter no mínimo duas respostas para um mesmo gênero ou orientação sexual e para a análise de concordância entre orientações sexuais diferentes, é necessário obter no mínimo duas respostas de orientações diferentes.

4.4.1 Concordância Média Geral

Esta análise tem como propósito identificar a homogeneidade das respostas obtidas nos dez formulários aplicados e não foi levado em consideração as orientações sexuais e identidades de gênero dos participantes.

Para a realização do cálculo da Concordância Geral, foi feita primeiramente a média percentual das concordâncias de cada formulário, para este cálculo foi utilizado o somatório das porcentagens médias de concordância de cada pergunta dividido pelo total de perguntas (16). E após obtenção dos resultados de cada formulário, foi realizado o somatório dos resultados e dividido pelo total de formulários (10), e assim resultando a taxa de Concordância Geral das respostas dos participantes que concluiu-se em 77,7%.

A Tabela 6 exibe as porcentagens médias de cada formulário, assim como a taxa de Concordância Geral das respostas.

Formulário	Porcentagem Média de Concordância por Formulário
Formulário 1	75%
Formulário 2	73,4%
Formulário 3	74,3%
Formulário 4	81,2%
Formulário 5	73,7%
Formulário 6	68,7%
Formulário 7	80,2%

Formulário 8	80,7%
Formulário 9	86,7%
Formulário 10	82,8%
CONCORDÂNCIA GERAL	**77,7%**

Tabela 6.Concordância Média Geral

4.4.2 Concordância por Grupos

Esta análise tem como propósito dividir os entrevistados por grupos a fim de identificar a homogeneidade das respostas obtidas nos dez formulários aplicados.

O primeiro grupo a ser analisado será o da Identidade de Gênero, que foi subdividido em dois subgrupos, o gênero feminino e masculino, onde foi observado a concordância média desses gêneros.

Após a análise deste grupo, o próximo a ser estudado será o da Orientação Sexual, subdividindo-se entre os subgrupos dos heterossexuais, homossexuais e bissexuais, assim como a cordâncias entre as orientações sexuais diferentes; e por fim uma análise da concordância entre a comunidade LGBTQ.

Para a realização do cálculo da Concordância por Grupos, foi seguida a mesma lógica matemática da Concordância Geral, porém estando atento a condição aplicada aos formulários definida previamente.

4.4.2.1 Identidade de Gênero

Ao aplicar a condição de comparação aos formulários foi percebido que para identidade de gênero, apenas alguns gêneros validaram corretamente para a análise entre identidades, que foram o masculino e feminino. O gênero não-binário não se encaixou na condição pois foi obtida apenas uma resposta. A Tabela 7 exibe a taxa média de concordância dos gêneros masculino e feminino.

Identidade de Gênero	Porcentagem Média das Concordâncias
Feminino	89,5%

Masculino	77,7%

Tabela 7. Porcentagem Média da Concordâncias por Identidade de Gênero.

Para concordância de respostas entre o gênero Feminino, os formulários que obtiveram duas ou mais respostas do mesmo gênero, foram os 4, 8 e 9, sendo respondidas 48 perguntas por 6 entrevistadas do gênero feminino. E através da média da soma percentual dos formulários que satisfazem a condição, foi observada uma taxa de concordância de 89,5% entre o gênero feminino. Utilizando a mesma condição para o gênero Masculino, que abrangeu todos os formulários, foram respondidas 160 perguntas por 47 entrevistados do gênero Masculino, com uma taxa de concordância de 77,7%.

4.4.2.2 Orientação Sexual

Ao aplicar a condição de comparação aos formulários foi percebido que para orientação sexual, apenas algumas orientações validaram corretamente para a análise entre orientações, que foram heterossexual, homossexual e bissexual. A orientação assexual não se encaixou na condição pois foi obtida apenas uma resposta. Entretanto, para a análise entre orientações diferentes todos os formulários se aplicaram corretamente. A Tabela 8 exibe a taxa média de concordância das orientações sexuais.

Orientação Sexual	Porcentagem Média das Concordâncias
Heterossexuais	77,8%
Homossexuais	87,5%
Bissexuais	87,5%
Entre Todas as Orientações	79,8%

Tabela 8. Porcentagem Média das Concordâncias por Orientação Sexual

Analisando a concordância entre mesma orientação sexual, os heterossexuais, obtiveram duas ou mais respostas em todos os formulários, sendo

respondidas 160 perguntas por 7 entrevistadas do gênero feminino, 39 do gênero masculino, com a uma taxa de concordância de 77,8%. Já para os bissexuais a condição foi aceita nos formulários 4 e 5, sendo respondidas 32 perguntas por 1 entrevistada do gênero feminino, 4 do gênero masculino, com uma ocorrência de 87,5% de concordância. E para os homossexuais a condição foi aceita em apenas um, formulário 3, sendo respondidas 16 perguntas por 4 entrevistados do gênero masculino, com a taxa de concordância de 87,5%.

Observando a concordância entre orientações sexuais, a condição foi aceita em 8 formulários; os formulários 2 e 6 não se aplicaram pois apresentaram em seus resultados apenas a orientação heterossexual. Foram respondidas 128 perguntas por 10 entrevistados do gênero feminino, 1 não-binário e 42 do gênero masculino, com uma taxa de concordância 79,3% entre a comunidade LGBTQ e Heterossexuais em suas respostas.

4.4.2.3 Concordância entre a comunidade LGBTQ

Esta análise comparativa tem como objetivo observar a concordância das respostas entre toda comunidade LGBTQ que participou da pesquisa. Dentre os dez formulários aplicados, condição comunidade LGBTQ foi aceita em quatro formulários (obtiveram duas ou mais respostas de orientações diferentes do heterossexual); formulários 3, 4, 5 e 9. Foram respondidas 64 perguntas por 2 entrevistadas do gênero feminino, 1 não-binário e 18 do gênero masculino. A soma das porcentagens média dos resultados obtidos entre os formulários que satisfizeram a condição, observou a ocorrência de 87,8% de concordância entre a comunidade LGBTQ em suas respostas.

5.Conclusão

A presente pesquisa tem como objetivo a classificação automática de tweets como LGBTQfóbico ou não. Para isso foi desenvolvido um algoritmo utilizando os conceitos de aprendizagem de máquina supervisionada com o auxílio da biblioteca Python Scikit-Learn, que através de um conjunto de dados previamente rotulados foi

possível classificar um novo conjunto de dados, que neste presente trabalho se caracteriza como o conjunto de tweets coletados por palavras-chave através da API do Twitter.

Para validar os resultados do algoritmo desenvolvido, foi feita a aplicação de dez formulários com um conjunto total de 160 tweets, tendo em vista medir a taxa de acerto entre os resultados dos entrevistados em relação aos apresentados pelo algoritmo.

A partir dos resultados obtidos através da análise das respostas coletadas dos formulários, foi possível perceber uma taxa de acerto geral de 64,92% e que em relação aos tweets possivelmente LGBTQfóbicos ocorreu uma taxa de acerto em relação aos tweets negativos de 66,67%.

Diante ao resultado foi analisado a Concordância Média Geral dos entrevistados, que obteve uma taxa de 77,7% de concordância entre suas respostas, mostrando que o grupo entrevistado possui uma homogeneidade entre suas opiniões individuais e expressando uma diferença percentual baixa em relação a taxa de acerto do algoritmo.

Para melhor entendimento do resultado da Concordância Geral foram estudadas as concordâncias das respostas dos entrevistados a partir da óptica da subdivisão de grupos. O primeiro grupo a ser analisado foi o da Identidade de Gênero, ao qual gênero feminino obteve uma taxa de concordância de 89,5%, e o gênero masculino ocorreu uma taxa de 77,7%, é perceptível que os participantes possuem opiniões homogêneas, concordando com o resultado da Cordância Média Geral.

O segundo grupo a ser estudado foi o da Orientação sexual, onde o subgrupo heterossexual obteve uma taxa de 77,8% de concordância entre suas respostas, já os subgrupos homossexuais e bissexuais tiveram a mesma taxa de concordância um pouco maior, 87,5%. Mostrando que ambos estão dentro do espectro do resultado da Concordância Geral, porém observando que os subgrupos que se encaixam na comunidade LGBTQ obtiveram um resultado acima da média.

E por fim o grupo da Comunidade LGBTQ expressou uma concordância de 87,8% das respostas, sendo um valor aproximado aos dos subgrupos dos homossexuais e bissexuais, mostrando uma conformidade entre os resultados.

Apesar de uma boa taxa de concordância entre os subgrupos apresentados em relação ao resultado geral, ocorreu uma taxa de empate de 16% e além de que alguns resultados obtiveram pouco mais do 60% de concordância entre as respostas coletadas e diante a este panorama foi necessária da interpretação de terceiros em alguns exemplos selecionados.

Os entrevistados foram selecionados com o requisito de fazerem parte da comunidade LGBTQ e de serem alvo dos insultos que eles analisaram. E através disso foi percebido que todos os entrevistados consideraram os tweets-exemplos como LGBTQfóbicos.

Diante os estudos e análises realizadas, esta pesquisa apresentou uma boa taxa de concordância média geral apresentando valores próximos a 80% e em relação a taxa de acerto do algoritmo expressou um resultado aceitável de quase 65%. Porém o assunto em questão desperta uma pluralidade de entendimentos, sendo assim difícil convergir as subjetividades individuais dos entrevistados em um resultado homogêneo, e que diante a este panorama os resultados da taxa de acerto podem ter sidos induzidos de certa maneira.

Referências

ANDRADE, Marcelo., PISCHETOLA, Magda., **O Discurso De Ódio Nas Mídias Sociais**: A Diferença Como Letramento Midiático E Informacional Na Aprendizagem. Revista e-Curriculum, 2016.

ALPAYDIN, E. (2010) **Introduction to Machine Learning**, The MIT Press, Cambridge, Massachusetts, EUA, 537 páginas.

AYODELE, T. **Types of Machine Learning Algorithms, New Advances in Machine Learning**, Yagang Zhang, 2010.

BECKER, K.; TUMITAN, D. **Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios.** Simpósio Brasileiro de Banco de Dados, 2013.

BRANDTZAEG, Petter Bae & HEIM, Jan; (2007). **Initial context, user and social requirements for the Citizen Media applications: Participation and motivations in off- and online communities.** Citizen Media Project.

GARTON, L.; HAYTHORNTHWAITE, C. e WELLMAN, B. **Studying Online Social Networks**. Journal of Computer Mediated Communication, n. 3, vol 1, 1997.

COHEN-ALMAGOR, Raphael, **Liberal Democracy and the Limits of Tolerance**. Ann Arbor: The University of Michigan Press, 2000.

GARTON, L.; HAYTHORNTHWAITE, C. e WELLMAN, B. **Studying Online Social Networks**. Journal of Computer Mediated Communication, n. 3, vol 1, 1997.

LIU, B. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers, 2012.

MARTINS MARQUES, Amanda Ravena; MUNIZ, Arnaldo Brasil: BRANDÃO, Maureen da Silva. **A liberdade de expressão e suas ameaças: reflexões a partir do caso Ellwanger.** [2013]. (Série Monografias do CEJ, v. 16).

MEYER-PFLUG, Samantha Ribeiro. **Liberdade de expressão e discurso do ódio. São Paulo: Editora Revista dos Tribunais**, 2009. 271 p.

NASCIMENTO, P.; OSIEK, B. A.; XEXÉO, G. **Análise de Sentimento de Tweets com Foco em Notícias**. Revista Eletrônica de Sistemas de Informação, v. 14, n. 2, p. 1-14, 2015.

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.

PETTENATI, Maria Chiara & RANIERI, Maria; (2006). **Informal learning theories and tools to support knowledge management in distributed CoPs.** IN Innovative Approaches for Learning and Knowledge Sharing, EC-TEL. Workshop Proceeding.

PRATI, Ronaldo Cristiano.**Novas Abordagens em Aprendizado de Máquina para a Geração de Regras, Classes Desbalanceadas e Ordenação de Caos**. 2006. 191 p. Tese de Doutorado - Universidade de São Paulo, São Paulo, 2006.

RECUERO, Raquel da Cunha. **Comunidades em Redes Sociais na Internet: Proposta de Tipologia baseada no Fotolog.com.** 2006. 334 p. Tese de Doutorado - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

SCHÄFER, Gilberto; LEIVAS, Paulo Gilberto Cogo; SANTOS, Rodrigo Hamilton dos. **Discurso de ódio:** Da abordagem conceitual ao discurso parlamentar. *Revista de informação legislativa*: RIL, v. 52, n. 207, p. 143-158, jul./set. 2015. Disponível em: http://www12.senado.leg.br/ril/edicoes/52/207/ril_v52_n207_p143.

SILVA, E. P. ; MALHEIROS, Y ; NUNES, R. T. A. ; ANTUNES, I. L. ; REGO, T. G. . **Um Conjunto de Dados Extraído do Twitter para Análise de Sentimentos na Língua Portuguesa.** In: Symposium in Information and Human Language Technology, 2019, Salvador. Symposium in Information and Human Language Technology, 2019.