Detecção de Mensagens Homofóbicas em Português no Twitter usando Análise de Sentimentos

Vinicius Matheus de Medeiros Silva Coutinho[1], **Yuri Malheiros**[2]

[1]Departamento de Ciências Exatas – Universidade Federal do Paraíba (UFPB)
Rio Tinto - PB - Brasil

[2]Centro de Informática - Universidade Federal do Paraíba (UFPB)
João Pessoa - PB - Brasil

vinicius.matheus@dcx.ufpb.br, yuri@ci.ufpb.br

Abstract. *Hate speech on social networks aimed at minorities brings hostility to this medium, causing suffering and harm beyond the digital world. Moderators can help control offensive messages, but with the large volume of messages posted, it is impossible to perform manual filtering. To try to combat hate speech more comprehensively, this work aims to use sentiment analysis to detect homophobic messages in Portuguese on Twitter. The results of the developed technique were compared with human interpretations. In this experiment, it was obtained accuracy of 0.6148, precision of 0.6667, recall of 0.6216, and f-measure of 0.6433.*

Resumo. *Discurso de ódio em redes sociais direcionado a minorias trazem hostilidade para este meio causando sofrimento e danos que vão além do mundo digital. Moderadores podem ajudar no controle de mensagens ofensivas, mas com o grande volume de mensagens publicadas é inviável realizar uma filtragem manual. Para tentar combater a propagação de mensagens de ódio de forma mais abrangente, este trabalho tem como objetivo utilizar análise de sentimentos para detecção de mensagens homofóbicas em português no Twitter. Os resultados da técnica desenvolvida foram comparados com as interpretações de humanos. Neste experimento, a técnica obteve 0,6148 de acurácia, 0,6667 de precisão, 0,6216 de sensibilidade e 0,6433 de f-measure.*

1. Introdução

As redes sociais mais populares atualmente interligam bilhões de pessoas em todo o planeta, permitindo o compartilhamento de informações em tempo real entre elas. Através dessas plataformas, os seus usuários conseguem acessar uma grande quantidade de outras pessoas para enviar e receber variados tipos de informações [Penni 2017].

A liberdade de criação e o vasto alcance das redes sociais trouxeram abundância e diversidade de conteúdo disponível, que muitas vezes não passa por filtros ou avaliação de qualidade. Portanto, as redes sociais podem ser facilmente distorcidas, propagando informações perigosas, como discursos de ódio, mentiras, teorias da conspiração, etc [Andrade and Pischetola 2016]. O discurso de ódio é definido como a manifestação hostil e preconceituosa de ideias que incitem discriminação de grupos específicos, geralmente por razões de gênero, etnia, religião, nacionalidade, deficiência, orientação sexual e condicionamento físico [Cohen-Almagor 2011]. Discursos de ódio que se iniciam em redes

sociais podem ir além do mundo virtual, se transformando em um problema ainda maior para a sociedade. Um dos tipos de discursos de ódio é o discurso homofóbico, que são mensagens de intolerância, muitas vezes violentas, direcionadas a pessoas homossexuais ou transexuais.

As redes sociais empregam moderadores e possuem ferramentas para que os usuários reportem conteúdo ofensivo, mas com o grande volume de mensagens publicadas é inviável filtrá-las manualmente, assim deixando uma parcela da população vulnerável aos ataques de pessoas mal intencionadas. Para tentar combater o discurso de ódio de forma mais abrangente, este trabalho tem como objetivo detectar automaticamente mensagens homofóbicas em português no Twitter, uma das maiores redes sociais da atualidade com cerca de 330 milhões de usuários ativos[1]. Para isso, foram coletadas mensagens que continham termos que popularmente são usados para fazer referência a homossexuais de uma forma ofensiva. Entretanto, dependendo do contexto, uma palavra pode ter um sentido ofensivo ou não, assim, a presença de uma palavra não determina o caráter homofóbico de uma mensagem. Por isso, usou-se um modelo de Regressão Logística treinado com um conjunto de dados de análise de sentimentos, para que apenas mensagens com sentimentos negativos e que continham termos potencialmente homofóbicos fossem consideradas homofóbicas.

Para avaliar a qualidade da classificação, as mensagens foram também classificadas por pessoas. Dessa forma, foi possível analisar a percepção humana em relação às mensagens e compará-las com a classificação automática. Além disso, três pessoas foram entrevistadas para que as suas opiniões sobre as mensagens fossem analisadas mais detalhadamente.

O restante deste artigo está organizado da seguinte forma. Trabalhos relacionados são apresentados e discutidos na Seção 2. A metodologia adotada com detalhes do desenvolvimento do trabalho e experimento é descrita na Seção 3. Os resultados obtidos são apresentados na Seção 4 e discutidos na Seção 5. A Seção 6 traz a discussão das entrevistas realizadas com três pessoas sobre a classificação das mensagens. Por fim, as considerações finais e trabalhos futuros são apresentados na Seção 7.

2. Trabalhos Relacionados

O grande volume de dados disponíveis e a disseminação cada vez maior de discursos de ódio nas plataformas da Internet têm atraído a atenção e esforços da comunidade científica para criar e investigar diferentes abordagens para conseguir detectar este tipo de mensagem automaticamente.

[de Pelle and Moreira 2017] coletaram 10.336 comentários de 115 notícias das seções de política e esporte do *website* G1 (g1.globo.com). Desses dados, 1.250 foram escolhidos aleatoriamente para serem rotulados. Cada comentário foi classificado como ofensivo ou não por três juízes. Além disso, quando considerado ofensivo, o juiz categorizava o comentário como racismo, sexismo, homofobia, xenofobia, intolerância religiosa ou xingamento. Neste trabalho, dois conjuntos de dados foram criados, o primeiro tem todos os comentários e suas classes foram as escolhidas por pelo menos dois juízes. O segundo é um conjunto mais restrito, ele é composto apenas pelos comentários que os

[1]https://investor.twitterinc.com/financial-information/quarterly-results/default.aspx

três juízes escolheram a mesma classe, totalizando 1.033 comentários. A detecção dos discursos de ódio foi realizada através de dois classificadores: Naive Bayes e Support Vector Machine (SVM), que obtiveram, *f-measure* de 0,71 e 0,77 respectivamente para o primeiro conjunto de dados, e 0,79 e 0,82 respectivamente para o segundo conjunto de dados.

Continuando com detecção de discursos de ódio na língua portuguesa, [Silva and Serapião 2018] utilizaram os dados do trabalho de [de Pelle and Moreira 2017] e do trabalho de [Fortuna 2017], totalizando três conjuntos de dados, para treinar uma rede neural convolucional (CNN). O desempenho da rede para classificação foi testada em duas configurações diferentes: utilizando vetores pré-treinados com o modelo Glove e Wang2Vec (uma variação do Word2Vec) e sem a utilização de vetores pré-treinados. Utilizando validação cruzada, a rede neural foi avaliada, alcançando *f-measure* 0,89, 0,96 e 0,96 para cada um dos conjuntos de dados.

[Davidson et al. 2017] coletou *tweets* em inglês com palavras que potencialmente tornam uma mensagem ofensiva, mas, com a mesma premissa utilizada no presente trabalho, isto não é suficiente para caracterizar uma mensagem como discurso de ódio. Assim, cada um dos 25.000 *tweets* coletados foram rotulados em uma das três categorias: discurso de ódio, linguagem ofensiva e nenhuma das anteriores. Para assegurar uma maior precisão na classificação, cada mensagem foi classificada por no mínimo três pessoas. Seis algoritmos de aprendizagem de máquina foram utilizados para teste: Regressão Logística, Naive Bayes, Árvores de Decisão, Florestas Aleatórias e SVM. Apesar de ter alcançado valores máximos de precisão e sensibilidade de 0,91 e 0,90, respectivamente, quando observados apenas os discursos de ódio, esses valores caem para 0,44 e 0,61.

Ainda para mensagens em língua inglesa no Twitter, [Burnap and Williams 2016] utilizaram modelos de aprendizagem de máquina para classificação de discursos de ódio em diferentes características, como: raça, deficiência e orientação sexual. Foram coletados dados sobre eventos que estavam relacionados às três características citadas. Para raça, foram coletados *tweets* sobre a reeileção de Barack Obama, para orientação sexual, o anúncio público de Jason Collins, o primeiro atleta olímpico americano que se assumiu homossexual, e para deficiência, a cerimônia de abertura dos jogos paraolímpicos de Londres. Para cada evento, 2.000 mensagens foram rotuladas por humanos como ofensivas em relação a raça, orientação sexual e deficiência. SVM e Florestas Aleatórias foram treinados e testados através de validação cruzada, tendo o primeiro obtido os melhores resultados. O *f-measure* médio para as classificações foi 0,68, sendo a precisão média 0,79 e a sensibilidade média 0,59.

O trabalho de [Gitari et al. 2015] teve como objetivo a criação de um classificador para detectar discurso de ódio em plataformas da *web* como blogs e fóruns. Para isso, foi desenvolvido uma abordagem em três etapas. A primeira envolveu a detecção de sentenças subjetivas, pois discursos de ódio costumam carregar expressões deste tipo. Na segunda etapa, um *lexicon* com palavras relacionadas a discursos de ódio foi construído através de um método baseado em regras. Por fim, na terceira etapa, um classificador baseado em regras foi criado usando o *lexicon*, para classificar sentenças de um documento como "ódio forte", "ódio fraco" e "sem ódio". O classificador foi avaliado usando dois conjuntos de documentos distintos. Para o primeiro, os valores máximos de precisão, sensibilidade e f-measure foram respectivamente: 0,7342, 0,6842 e 0,7083. Para o segundo

conjunto os valores máximos de precisão, sensibilidade e f-measure foram respectivamente: 0,7155, 0,6824 e 0,6985.

Pode-se perceber que a maioria dos trabalhos passa pela fase de criação e rotulação de um conjunto de dados, algo que evitamos usando um conjunto de dados previamente disponível. [Silva and Serapião 2018] é o único dos trabalhos apresentados que partiu de conjuntos já criados. Como trabalhos em análise de sentimentos são mais abrangentes que trabalhos específicos sobre discurso de ódio ou homofobia, interligar os dois problemas pode nos levar a avanços nessa área. Em relação aos algoritmos utilizados, três dos quatro trabalhos utilizaram SVM com sucesso. O presente trabalho utilizou Regressão Logística para classificação de sentimentos, classificador também utilizado por [Davidson et al. 2017]. Com exceção de [de Pelle and Moreira 2017] e [Silva and Serapião 2018], os trabalhos são difíceis de comparar, pois usam dados diferentes. Entretanto, os mais próximos ao nosso trabalho são os trabalhos de [Davidson et al. 2017] e [Burnap and Williams 2016] que também classificaram mensagens coletadas no Twitter.

3. Metodologia

Para alcançar o objetivo deste trabalho, inicialmente foi construído um conjunto de dados com mensagens do Twitter que continham palavras potencialmente homofóbicas. Em seguida, foi aplicada análise de sentimentos para diferenciar as mensagens negativas das positivas ou neutras. Por fim, 160 mensagens foram classificadas por seres humanos para validação e análise dos resultados alcançados. As subseções a seguir explicam em detalhes cada uma dessas etapas.

3.1. Coleta dos dados

Nesta etapa, inicialmente foram selecionadas palavras potencialmente homofóbicas para guiar a busca por mensagens no Twitter. Foram escolhidas quatro palavras, cada uma representando um grupo de pessoas específico. As palavras selecionadas foram:

- Sapatão: palavra usada como referência a pessoas homossexuais do sexo biológico feminino;
- Viado: palavra usada como referência a pessoas homossexuais do sexo biológico masculino;
- Traveco: palavra usada como referência a transexuais;
- Gay: palavra usada de forma genérica como referência a pessoas homossexuais.

Através da API do Twitter, essas palavras foram usadas para buscar mensagens compartilhadas na rede social. A coleta foi realizada durante cinco dias no período de 27 de julho de 2019 e 31 de julho de 2019, coletando 8.349 *tweets*. Na Tabela 1 são apresentadas as quantidades de *tweets* coletados por palavras-chave e na Tabela 2 tem-se um exemplo coletado para cada palavra-chave.

3.2. Análise de sentimentos

Para etapa de análise de sentimentos das mensagens, inicialmente os *tweets* coletados foram pré-processados através de quatro passos: remoção de links, remoção de *stopwords* em português, aplicação de *stemming* e remoção de acentuação. Após realizadas essas

Palavra-chave	Quantidade de tweets
Gay	1.166
Sapatão	1.957
Traveco	1.267
Viado	3.959
Total	8.349

Tabela 1. Quantidade de tweets coletados por palavras-chave

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

Tabela 2. Exemplos de tweets coletados

etapas, o texto foi vetorizado utilizado o *TFIDF Vectorizer* da biblioteca *Scikit-learn* [Pedregosa et al. 2011].

Com os *tweets* pré-processados, foi treinado um modelo de Regressão Logística utilizando o conjunto de dados TASH-PT para análise de sentimentos, que possui 2.787 *tweets*, sendo 888 positivos, 881 negativos e 1.018 neutros [Silva et al. 2019]. A Tabela 3 exibe quatro resultados de classificação obtidos pelo modelo. Na abordagem deste trabalho, para determinar se uma mensagem é homofóbica, ela precisa ter uma das palavras-chave, que caracterizam uma mensagem como potencialmente homofóbica, e ter sentimento negativo. Assim, dos exemplos da Tabela 3, os dois primeiros não são considerados homofóbicos e os dois últimos são.

3.3. Validação

Para validação foram aplicados questionários com 16 *tweets* coletados e classificados pelo modelo. Para cada *tweet*, o entrevistado marcava se ele interpretava a mensagem como sendo homofóbica ou não. O conjunto de 16 *tweets* foi composto por *tweets* que contêm as quatro palavras-chave selecionadas para esta pesquisa (gay, sapatão, traveco e viado). Neles, cada palavra aparece em quatro *tweets* que estão dispostos da seguinte forma: dois *tweets* negativos considerados como homofóbicos e um positivo e um neutro considerados como não homofóbicos. Dessa forma temos uma distribuição equivalente de *tweets* homofóbicos e não homofóbicos para cada uma das palavras-chave. Além das classificações, o questionário também perguntava sobre a orientação sexual e a identidade de gênero do entrevistado.

Dez questionários distintos foram elaborados, para que os resultados abrangessem

Tweet	Sentimento
“brega e xote melhores tipos de música quem discorda eh viado”	Positivo
“putz que bonito esse menino barbado e de óculos quadrado, será se é gay?”	Neutro
“Coisa horrível... sem conteúdo... traveco”	Negativo
“A mulher parece sapatão de qualquer forma, não adianta mudar o personagem.”	Negativo

Tabela 3. Exemplos de classificação de sentimento

	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 dos entrevistados

uma quantidade maior de *tweets*. Com isso, foram analisados por seres humanos 160 *tweets* classificados como homofóbicos ou não. A seguir, usa-se QN para representar o questionário N, por exemplo, $Q1$ identifica o questionário de número 1, $Q2$ o de número 2 e assim sucessivamente. A aplicação destes questionários 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 e-mails dos discentes dos cursos. No total, 57 pessoas responderam os questionários completamente.

4. Resultados

Os primeiros dados apresentados são as informações pessoais dos entrevistados. Com eles, tem-se uma visão geral das identidades de gênero e orientações sexuais de todos os participantes. A Tabela 4 exibe que de acordo com a identidade de gênero, 46 (80,7%) dos entrevistados são do sexo masculino, 10 (17,5%) são do sexo feminino e 1 (1,8%) se auto declarou não-binário. Em relação à orientação sexual, temos que 44 (77,2%) são heterossexuais, 5 (8,8%) são homossexuais, 6 (10,5%) bissexuais e 2 (3,5%) assexuais.

As respostas do classificador foram comparadas com as interpretações das pessoas que responderam os questionários. A Tabela 5 apresenta quantas pessoas responderam cada questionário. Como cada mensagem foi avaliada por mais de uma pessoa, o valor final da interpretação é dado de acordo com a maioria. Por exemplo, se, de cinco pessoas,

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Número de entrevistados	5	4	9	7	10	2	6	4	6	4

Tabela 5. Quantidade de entrevistados por questionário

	Acurácia	Precisão	Sensibilidade	F-measure
Q1	0,4375	0,6300	0,4545	0,6369
Q2	0,7000	0,4000	1,0000	0,5714
Q3	0,6250	0,7500	0,6000	0,6667
Q4	0,6875	0,8750	0,6363	0,7368
Q5	0,6428	0,6300	0,7142	0,6667
Q6	0,5000	0,2500	1,0000	0,4000
Q7	0,5000	0,6700	0,5000	0,5714
Q8	0,7500	0,8571	0,7500	0,8000
Q9	0,7500	0,6700	0,7500	0,7059
Q10	0,5333	0,5714	0,5000	0,5333
Total	0,6148	0,6667	0,6216	0,6433

Tabela 6. Acurácia, precisão, sensibilidade e f-measure do classificador

duas interpretaram uma mensagem como homofóbica e três como não homofóbica, então considera-se a interpretação como não homofóbica.

Em 26 das 160 perguntas aconteceram empates, ou seja, os entrevistados interpretaram na mesma proporção a mensagem como homofóbica e não homofóbica. Sendo assim, estas mensagens foram desconsideradas na nossa análise, pois a interpretação pelas pessoas é inconclusiva. Com isso, o número total de mensagens avaliadas foi de 134.

A Tabela 6 apresenta os valores da acurácia, precisão, sensibilidade e *f-measure* obtidos pelo classificador em relação às interpretações das pessoas que responderam os questionários. Nela, tem-se os valores das métricas para cada questionário e, na última linha, os cálculos foram feitos considerando todas as perguntas.

A Tabela 7 traz o cálculo do *f-measure* separado por grupos de entrevistados. Pode-se observar o valor da métrica para pessoas com identidade de gênero masculino, feminino e não-binário, além das pessoas homossexuais, heterossexuais, bissexuais e assexuais. Quando os grupos foram separados, em alguns casos, também existiram empates. Essa situação foi tratada com a mesma abordagem feita para o cálculo das métricas de forma geral (Tabela 6), ou seja, mensagens com empate de interpretações foram removidas do cálculo. Os valores faltantes são os casos que o questionário não foi respondido por ninguém de um determinado grupo. Por exemplo, os questionários 2, 6 e 10 não foram respondidos por nenhuma pessoa do sexo feminino.

Para avaliar a concordância entre os entrevistados foi calculado o coeficiente Kappa de Fleiss [Fleiss 1971]. A Tabela 8 apresenta os valores dos coeficientes para cada questionário. Na primeira coluna tem-se o valor considerando todas as pessoas e nas colunas seguintes têm-se os valores separados por grupos. Nesta tabela pode-se observar mais valores faltantes que na Tabela 7, isto acontece, pois para medir a concordância, um

	Masculino	Feminino	Não binário	Heteros-sexual	Homos-sexual	Bissexual	Assexual
Q1	0,9615	0,5000	-	0,9615	0,5000	-	-
Q2	0,5714	-	-	0,5714	-	-	-
Q3	0,6250	0,5000	-	0,5882	0,5455	-	-
Q4	0,7368	0,6667	-	0,7500	-	0,7000	-
Q5	0,6250	0,6667	-	0,6667	0,5333	0,6154	-
Q6	0,4000	-	-	0,4000	-	-	-
Q7	0,5000	0,6316	-	0,6667	-	-	0,5000
Q8	0,7273	0,7059	-	0,6667	0,7368	-	0,6667
Q9	0,7059	0,7273	-	0,7500	0,7500	0,7059	-
Q10	0,5000	-	0,6667	0,5000	-	-	-
Total	0,6277	0,6296	0,6667	0,6479	0,6234	0,6800	0,5882

Tabela 7. F-measure do classificador dividido por grupos de entrevistados

	Todos	Masculino	Feminino	Heterossexual	Homossexual	Bissexual
Q1	0,0805	0,1259	-	0,1259	-	-
Q2	0,2954	0,2954	-	0,2954	-	-
Q3	0,2151	0,2260	-	0,1599	0,4920	-
Q4	0,2831	0,3952	0,4514	0,2029	-	0,6467
Q5	0,2295	0,2373	-	0,1624	-	0,4920
Q6	-0,3333	-0,3333	-	-0,3333	-	-
Q7	0,3614	0,3182	-	0,3553	-	-
Q8	0,3778	0,5000	0,3333	0,2889	-	-
Q9	0,5595	0,4987	0,6113	0,5997	-	-
Q10	0,3535	0,2487	-	0,2487	-	-
Média	0,2422	0,2512	0,4654	0,2106	0,4920	0,5697

Tabela 8. Coeficientes Kappa de Fleiss para verificar concordância entre os entrevistados

questionário precisa ser respondido por pelo menos duas pessoas de um mesmo grupo.

Por fim, a Tabela 9 traz a proporção de classificações de mensagens como homofóbicas. A primeira coluna apresenta o valor da proporção considerando todas as pessoas que responderam os questionários e as colunas seguintes mostram os valores separados por grupos.

5. Discussão

A acurácia média dos questionários foi de 0,6148, sendo o maior valor 0,75 para os questionários 8 e 9 e o menor valor 0,4375 para o questionário 1. Calculando para todos os questionários, os valores de precisão e sensibilidade são próximos, respectivamente 0,6667 e 0,6216. Isto indica que os números de falsos positivos e de falsos negativos são parecidos. Ou seja, a identificação não está tendendo a errar mais uma classificação que outra. Estes resultados diferem dos valores de precisão e sensibilidade encontrados em [Burnap and Williams 2016], que obteve precisão de 0,79 e sensibilidade de 0,59. Ou

Todos	Masculino	Feminino	Não binário	Heteros-sexual	Homos-sexual	Bissexual	Assexual
0,5407	0,5338	0,5588	0,6250	0,5407	0,5131	0,6364	0,5625

Tabela 9. Proporção de respostas "homofóbico" nos questionários

seja, nosso filtro é mais rigoroso, fazendo com que menos mensagens homofóbicas passem sem ser detectadas, por outro lado, nossa abordagem é menos precisa.

Analisando os questionários individualmente, percebe-se que existe também uma regularidade entre os valores de precisão e sensibilidade, com exceção do questionário 6. Este questionário foi respondido por apenas duas pessoas e 10 das 16 perguntas foram empates, isto pode ter influenciado na variação dos seus resultados. O questionário 2 apresenta a segunda maior disparidade entre precisão e sensibilidade, mas, assim como o questionário 6, ele também foi um dos que obteve menos respostas, quatro no total e seis perguntas tiveram interpretações empatadas. Por outro lado, os questionários 8 e 10 também foram respondidos por quatro pessoas, com quatro e um empate respectivamente, e não possuem uma variação acentuada. O *f-measure* apresentado na Tabela 6 sumariza as métricas de precisão e sensibilidade. Nela observa-se que o maior valor de *f-measure* foi do questionário 8 e o menor do questionário 6. Calculando para todos os questionários o valor da métrica foi 0,6433.

Na Tabela 7 pode-se perceber que os valores de *f-measure* dos grupos são semelhantes a 0,6433, valor obtido no cálculo desta métrica para todas as pessoas. O menor valor 0,5882 foi obtido para o grupo das pessoas assexuais e o maior 0,68 para as pessoas bissexuais. Para o questionário 1 o grupo masculino e o grupo heterossexual obtiveram o valor 0,9675, o maior de todas as análises. Entretanto, isto não aconteceu para os grupos feminino e homossexual neste mesmo questionário, para eles, o resultado foi de 0,5.

Calculando o coeficiente Kappa de Fleiss, pode-se observar na Tabela 8 que a média dos valores dos coeficientes foi de 0,2422, que é considerada uma concordância razoável. Entre os formulários, o maior valor do coeficiente foi 0,5595, que indica uma concordância moderada, no questionário 9, e o menor foi -0,3333 no questionário 6, indicando concordância insignificante. Entre os grupos, o maior valor do Kappa de Fleiss foi para os bissexuais e o menor para as pessoas heterossexuais, entretanto todos podem ser considerados valores de concordância razoável. Com este resultado percebe-se que o problema de classificação de uma mensagem como homofóbica não é claro para as pessoas, existe um certo nível de concordância, mas ele poderia ser maior. Isto se reflete no resultado de abordagens usando aprendizagem de máquina para classificação. Se não é claro para pessoas se uma mensagem tem sentimento positivo ou negativo, os classificadores também terão dificuldade em diferenciá-la. Assim, usando os valores dos coeficientes de Kappa de Fleiss, tem-se um forte indício da complexidade inerente de um problema em muitos casos subjetivo como o de classificar mensagens como homofóbicas.

Nos questionários, a metade das mensagens foram classificadas como homofóbicas e a outra metade como não homofóbicas. Assim, idealmente, os entrevistados deveriam responder metade das questões com uma classificação e metade com a outra. A Tabela 9 traz a proporção de classificações homofóbicas dadas pelas pessoas para as mensagens do questionário. Pode-se observar que, para todas as pessoas, tem-se uma

Mensagem	Resultado do algoritmo	Resultado do questionário
“To com voz de traveco merda, cadê minha voz irmão”	Homofóbico	Empate
“Homem é 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”	Não homofóbico	Homofóbico
“A mulher’ parece sapatão de qualquer forma, não adianta mudar o personagem’	Homofóbico	Não Homofóbico

Tabela 10. Mensagens analisadas pelos entrevistados

proporção de 0,5407. Isto indica que na interpretação dos entrevistados, 54,07% das mensagens eram homofóbicas, um pouco mais que os 50% dos questionários. Assim, os entrevistados têm uma tendência maior que o algoritmo de perceber as mensagens como homofóbicas.

As maiores proporções de percepção de mensagens homofóbicas foram dos grupos bissexual e não-binário, respectivamente. O primeiro classificou 65,50% das mensagens como homofóbicas e o segundo 64,54%. Por outro lado, mesmo sendo um dos grupos afetados diretamente pelas mensagens homofóbicas, os homossexuais foram os entrevistados que menos interpretaram as mensagens como sendo homofóbicas, a proporção para essa classificação foi de 0,5131.

6. Entrevistas

Para compreender melhor os resultados e para conseguir saber de forma mais próxima a opinião das pessoas sobre suas interpretações das mensagens, três pessoas foram convidadas a comentar sobre as classificações da técnica proposta. Estas pessoas não participaram do experimento de classificação através dos questionários. O primeiro entrevistado foi uma professora de inglês transsexual, o segundo um bacharel em ciências biológicas homossexual e o terceiro uma desenvolvedora de software homossexual. A seguir apresentamos os comentários deles sobre as mensagens da Tabela 10, nela, na primeira coluna tem-se o texto das mensagens, na segunda o resultado obtido pela abordagem deste trabalho e na terceira o resultado das respostas dos questionários.

A primeira participante comentou sobre o primeiro *tweet* da Tabela 10, ela disse: “primeiro que traveco já é um termo ofensivo, e a ideia de voz de traveco, que é uma voz aguda, que não é regra. Este *tweet* tem sim homofobia nele, mesmo que implícita ou explícita, por que vai depender de fatores externos também.” A interpretação da entrevistada é a mesma do algoritmo, ela considera o termo traveco e a expressão voz de traveco como ofensivas, entretanto ela ressalta que podem existir outras interpretações e que a homofobia pode depender do contexto.

Para a segunda mensagem, o segundo entrevistado comentou: “pela forma como ela fala é como se eu não tivesse direito de ter uma opinião contrária apenas por ser homem e viado, independente da minha opinião está certa ou errada, dentro ou fora de contexto. Liberdade de expressão ta aí”. O entrevistado interpretou o *tweet* como homofóbico, resposta contrária a do algoritmo, mas concordando com a maioria das pessoas que responderam o questionário.

Por fim, a terceira entrevistada comentou sobre a terceira mensagem: “sim, na minha visão é homofóbico. Até por que essa frase já rotula uma estética atribuída às pessoas lésbicas”. Discordando dos resultados apresentados pelos entrevistados e concordando com o algoritmo, a entrevistada classifica a mensagem como homofóbica, partindo da análise que ocorre rotulação estética e comportamental, que atribui a concepção masculina às homossexuais do gênero feminino.

Através destas análises obtidas pelos entrevistados é perceptível a pluralidade de interpretações diante ao assunto apresentado. Muitos são os fatores, além do próprio conteúdo da mensagem, que podem influenciar na identificação de um texto como homofóbico ou não. Uma ferramenta de detecção automática é um passo importante, mas a visão das pessoas ainda é crucial para um entendimento aprofundado do significado das mensagens.

7. Conclusão

Este trabalho apresentou um método de detecção de mensagens homofóbicas em português no Twitter utilizando análise de sentimentos. Para isso, um classificador de aprendizagem de máquina foi treinado utilizando um conjunto de dados previamente disponível e aplicado a *tweets* que continham palavras potencialmente homofóbicas.

Entrevistados deram suas interpretações às 160 mensagens classificadas pela técnica proposta, que obteve acurácia de 0,6148, precisão de 0,6667, sensibilidade de 0,6216 e *f-measure* de 0,6433. A concordância entre os entrevistados foi medida através do coeficiente Kappa de Fleiss, que teve um valor médio de 0,2422, que representa uma concordância razoável. Isto mostra que mesmo as pessoas não têm uma interpretação homogênea do que são mensagens homofóbicas. Confirmando esse resultado, três pessoas foram convidadas para opinar mais detalhadamente sobre as mensagens, duas delas mencionaram de forma explícita que a interpretação depende do contexto e de fatores externos. Com isso, tem-se mais um indício da subjetividade das interpretações que são possíveis para mensagens compartilhadas em redes sociais.

Os resultados obtidos são similares a outros trabalhos que propõem técnicas de classificação de *tweets* que contém algum tipo de discurso de ódio. Assim, mostramos a viabilidade de utilizar técnicas de análise de sentimentos para essa finalidade. Como a área de análise de sentimentos é amplamente estudada, pode-se usar os seus avanços para aplicações voltadas a detecção e filtragem de mensagens homofóbicas.

Como trabalhos futuros, pretende-se investigar o uso de mais classificadores de análise de sentimentos para detecção de mensagens homofóbicas ou para outros tipos de discursos de ódio. Além disso, pode-se melhorar classificadores já existentes, com novos e melhores conjuntos de dados, e aplicando outras técnicas de aprendizagem de máquina supervisionada. Outra linha de investigação é a realização de testes mais abrangentes, com mais pessoas e mais mensagens, para confirmar os resultados obtidos neste trabalho.

Utilizar um classificador de sentimentos previamente construído é uma vantagem, pois reutiliza um artefato já desenvolvido e testado, entretanto, o problema de classificação de sentimentos é subjetivo, complexo e pode ser ambíguo em alguns casos. Assim, usar um classificador desse tipo, faz o classificador de mensagens homofóbicas depender de bons resultados da análise de sentimentos, que nem sempre são possíveis. Outra limitação importante é que mensagens criticando homofobia podem possuir sentimento negativo e citar termos potencialmente homofóbicos. Estas mensagens não são homofóbicas, mas poderiam ser classificadas dessa forma na nossa abordagem.

Referências

Andrade, M. and Pischetola, M. (2016). O discurso de ódio nas mídias sociais: a diferença como letramento midiático e informacional na aprendizagem. *Revista e-Curriculum*, 14(4):1377–1394.

Burnap, P. and Williams, M. L. (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5(1):11.

Cohen-Almagor, R. (2011). Fighting hate and bigotry on the internet. *Policy & Internet*, 3(3):1–26.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fortuna, P. C. T. (2017). Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

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

Penni, J. (2017). The future of online social networks (osn): A measurement analysis using social media tools and application. *Telematics and Informatics*, 34(5):498–517.

Silva, E. P., Malheiros, Y., Nunes, R. T. A., Antunes, I. L., and Rêgo, T. G. (2019). Um conjunto de dados extraído do twitter para análise de sentimentos na língua portuguesa. In *Proceedings of XII Symposium in Information and Human Language Technology*, pages 53–60.

Silva, S. and Serapião, A. (2018). Detecção de discurso de ódio em português usando cnn combinada a vetores de palavras. In *Proceedings of KDMILE 2018, Symposium on Knowledge Discovery, Mining and Learning, São Paulo, SP, Brazil*.