Criação e Caracterização de um Corpus de Discurso Sexistas em Português

M. Luísa P. Braga, Fabiola G. Nakamura, Eduardo F. Nakamura

Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brazil

{mlpb,fabiola,nakamura}@icomp.ufam.edu.br

Resumo. *Identificar o discurso de disseminação de ódio e preconceito é um desafio contínuo para os ambientes de mídias sociais online. Uma caracterização e identificação precisa são peças-chave para tratar e mitigar a violência, assim como, educar os usuários de forma eficaz e assertiva. A disseminação online de ódio pode ser direcionada a grupos distintos de pessoas, o que gera várias classes de discurso de ódio, como por exemplo o racismo, a homofobia ou o sexismo. Esse último é um tópico cujo interesse social tem crescido a medida que a figura feminina vence as barreiras da desigualdade de gênero. Em particular, o discurso sexista propaga e incentiva o comportamento depreciativo e abusivo contra mulheres. Neste trabalho, apresentamos um corpus de discurso sexista em Português coletado a partir de portais de notícias de grande penetração popular, como G1 e UOL, por exemplo. O trabalho apresenta três contribuições principais: (1) o processo de criação do corpus e de rotulação de comentários (sexista/não sexista); (2) a caracterização e análise do corpus e do comportamento dos rotuladores anônimos; (3) uma avaliação inicial de técnicas de aprendizagem de máquina para classificação de comentários sexistas/não sexistas. Os resultados preliminares mostram que, ao utilizar support vector machine, é possível identificar comentários sexistas com uma medida F1 acima de 0,8, precisão acima de 0,9 e revocação próxima a 0,8.*

Abstract. *Identifying hate speech and prejudice is an ongoing challenge for online social media environments. Accurate characterization and identification are key for treating and mitigating violence, as well as educating users effectively and assertively. The online spread of hate can be targeted at different groups of people, which generates various classes of hate speech, such as racism, homophobia or sexism. The latter is a topic whose social interest has grown as the female figure overcomes barriers of gender inequality. In particular, sexist discourse propagates and encourages derogatory and abusive behavior against women. In this work, we present a corpus of sexist discourse in Portuguese collected from news portals of great popular penetration, such as G1 and UOL, for example. The paper presents three main contributions: (1) the process of creating the corpus and labeling comments (sexist / non-sexist); (2) the characterization and analysis of the corpus and the behavior of anonymous labelers; (3) an initial assessment of machine learning techniques for classifying sexist / non-sexist comments. Preliminary results show that, when using support vector machine, it is possible to identify sexist comments with an F1 measure above 0.8, precision above 0.9 and recall close to 0.8.*

1. Introdução

O sexismo consiste em atos e discursos que ofendem, agridem ou diminuem as pessoas de um gênero [1, 2]. Direcionar o sexismo à mulheres é muito comum na Internet e causa grande impacto social, uma vez que incentiva a prática de discriminação ou violência contra a mulher.

Apesar de ser um meio de propagar informações, a Internet também é utilizada para orquestrar e incentivar crimes. Páginas *online* têm sido utilizadas como pontos de encontros virtuais entre pessoas que compartilham comportamentos de risco direcionados a grupos sociais como negros e mulheres. Em 2018, o líder dessas páginas foi preso e condenado a 21 anos pela justiça do Paraná, sob a acusação de ter utilizado a Internet para divulgar imagens contendo pedofilia e racismo, liderar uma associação criminosa virtual, incentivar o cometimento de crimes ainda mais graves por parte de terceiros, como homicídio, feminicídio e terrorismo [3].

Outro caso do uso da Internet como meio de propagação de crimes ocorreu em maio de 2014 quando uma mulher foi espancada até a morte por conta de um boato publicado no *Facebook* [4]. No mesmo mês em 2018, foi noticiado o caso de uma jovem que cometeu suicídio por conta de mensagens de ódio que recebia em suas redes sociais [5]. Ambos os casos mostram como a disseminação do discurso de ódio pode gerar consequências letais aos seus alvos mesmo que não gere violência física.

Ao longo de 2018, foram identificados mais de 68 mil casos de violência contra a mulher [6] e a ausência de combate à fragilização da figura feminina é um dos fatores que contribuí para o aumento dos casos. Identificar e combater o discurso ofensivo na Internet são formas de evitar a propagação de comportamentos violentos *online*, segundo Banks [7] a falta a preocupação em punir os autores de discurso de ódio agrava a propagação desse tipo de discurso na Internet.

Embora exista uma dificuldade humana de classificar o grande volume de opiniões publicadas na internet diariamente, é possível mitigar a propagação do discurso ofensivo, em particular do discurso sexista, através de ferramentas computacionais, como a classificação prévia e automática de publicações em redes sociais como sendo sexistas ou não. Para tanto, podemos utilizar técnicas de processamento de linguagem natural e aprendizagem de máquina, como é feito nos trabalhos de Davidson et al. [8] e Kowk&Wang [9].

O principal objetivo desse trabalho é a caracterização de comentários a partir da análise de uma base de dados representativa do discurso sexista, consequentemente este trabalho tem como contribuição a criação de uma base de dados de comentários sexistas ou não sexistas em portais de notícia, através da identificação das características distintivas de cada classe de comentários.

O artigo foi organizado da seguinte forma: a seção 2 apresenta conceitos relevantes para o desenvolvimento do trabalho e trabalhos relacionados à detecção de discurso de ódio. A seção 3 mostra a metodologia utilizada para o desenvolvimento do trabalho. A seção 4 exibe os resultados obtidos com a execução da metodologia. A seção 5 mostra as considerações finais sobre os resultados obtidos, seguidas das referências utilizadas.

2. Fundamentação teórica

O discurso de ódio é divido em classes como racismo, homofobia, misógina e xenofobia. Os trabalhos de Badjatya et al. [10] e Park&Fung [11] tem como objetivo a detecção automática do discurso de ódio em *tweets*, classificando o discurso como racista, sexista ou nenhum dos dois, uma vez que cada uma dessas classes de discurso tem características específicas e discriminatórias que são relevantes para qualquer tipo de classificação automática de discurso de ódio.

Os conceitos de sexismo apresentados por Glick&Fisk [2] e Smigay&Ellen [1], englobam tanto o discurso misógino como também qualquer discurso ofensivo direcionado a mulheres. Para tanto, é necessário que saibamos identificar e classificar também os discursos que são ofensivos mesmo que não apresentem ódio, como é feito por Davidson et al. [8].

A seguir serão descritos os conceitos de sexismo considerados neste artigo e trabalhos da literatura que tratam do problema de classificação automática de textos definindo discurso de ódio, discurso ofensivo ou sentimentos expressos no texto.

2.1. Sexismo

Segundo Glick&Fisk [2] existem dois tipos de sexismo, o hostil e o benevolente. Enquanto o sexismo hostil se resume aos atos e discursos misóginos, o sexismo benevolente se manifesta através de atos de proteção, idealização ou afeto dirigidos às mulheres, mesmo que não gerem ofensa. Já Smigay&Ellen [1] definem sexismo como opiniões e práticas que desprezam, desqualificam, desautorizam e violentam as mulheres, que são tomadas como seres de menor prestígio social. A partir desses conceitos, definimos sexismo como qualquer ação ou discurso com a intenção de ofender, diminuir, oprimir ou agredir pessoas do gênero feminino.

2.2. Trabalhos relacionados

Kwok&Wang [9] utilizaram um classificador de Naive Bayes com *bag of words* (BoW) de unigramas para detectar *tweets* racistas. Na montagem da sua base de dados, além da classificação entre "racista" e "não racista", os autores categorizaram os *tweets* racistas a partir do motivo pelo qual eles foram considerados ofensivos e detectou que o motivo mais comum era a presença de palavras ofensivas, por esse motivo os autores escolheram utilizar *bag of words* de unigramas como característica para o classificador. Kwok&Wang obtiveram uma acurácia de 76% e pontuaram que uma vez que palavras ofensivas não necessariamente indicam intenção racista em um *tweet*, o uso de unigrams não trás o contexto necessário para que o classificador seja eficaz.

Já no trabalho de Peng et al. [12], os autores utilizaram características variadas em três algoritmos de classificação diferentes afim de realizar a classificação de sentimentos em uma base de resenhas sobre filmes. As resenhas poderiam ser classificadas como "positiva" e "negativa" e os autores utilizaram *bag of words* (BoW) representativas de cada classe possível. Em suas conclusões, os autores observaram que o Support Vector Machine (SVM) foi o classificador com melhores resultados para o problema e também notaram que o uso da frequência de unigramas teve resultados superiores ao uso da frequência de bigramas para o contexto do seu trabalho com 82,9% de acurácia.

Autores	Método	Features	Resultados
Kwok&Wang	Naive Bayes	BOW de unigramas	76% de acurácia
Peng et al.	SVM	BOW de unigramas	82,9% de acurácia
Davidson et al.	Regressão Logística	TD-IDF de n-gramas de até 3 palavras	91% de precisão, 90% de revocação e 90% de F1

Tabela 1. Relação dos resultados principais apresentados pelos trabalhos relacionados.

Davidson et al. [8] realizam utilizaram Regressão Logística com regularização de L2 aplicada no TF-IDF de n-gramas para diferenciar o discurso ofensivo do discurso de ódio em tweets. Os rótulos disponíveis eram "discurso de ódio", "ofensivo sem discurso de ódio" e "nem ofensivo nem discurso de ódio" e foram atribuídos a cada *tweet* por pessoas diversas. Os resultados do autor foram 91% de precisão, 90% de revocação e 90% de F1, no entanto os autores observaram para a classe de ódio a precisão e revocação foram de 44% e 61% respectivamente, o que indica uma classificação incorreta dessa classe. Além disso, o autor afirma que é comum que as pessoas considerem discursos racistas e homofóbicos como ódio, no entanto o discurso sexista não recebe a mesma classificação.

Em todos os trabalhos o uso de unigramas mostrou resultados superiores ao uso de n-gramas com mais termos, então neste trabalho utilizaremos o TF de unigramas como *features* para os classificadores.

A tabela 1 apresenta resumidamente os resultados obtidos pelos trabalhos citados nesta seção, apresentando os métodos de aprendizagem utilizados como classificadores em cada trabalho, as *features* utilizadas e a precisão obtida em cada caso.

3. Criação da base de dados

A criação da base de dados se dá em três etapas: coleta de comentários, classificação dos manual comentários e análise dos comentários. As etapas estão descritas nesta seção.

3.1. Coleta de comentários

Na coleta de comentários, utilizamos como fonte de dados os portais de notícia G1 e UOL, neles coletamos notícias relacionadas com as palavras chave "mulher", "feminismo", "feminicídio" e "assédio", pois essas notícias têm maior probabilidade de gerar comentários e discussões sexistas, de forma que exista uma concordância entre o tema dos comentários.

Dentre as notícias que foram resultado para a busca pelas palavras chave citadas, escolhemos as que possuíam pelo menos 50 comentários na data da coleta. Selecionamos 24 notícias no total e criamos *crawlers* para coletar informações dos comentários de cada notícia, como o conteúdo dos comentários e os números de *likes* e *dislikes*.

Coletamos informações de 3.172 comentários, dentre os quais identificamos exemplos de comentários sexistas e não sexistas. O comentário "*Feministas são pessoas burras, incapazes de refletir sobre a influência do meio ambiente nas relações humanas, ao longo de sua existência.*" é um exemplo de comentário sexista.

3.2. Classificação manual dos comentários

Para que pessoas de gêneros e idades diferentes pudessem rotular a base de dados, criamos uma plataforma online, hospedada em ⟨http://sexismo.ml⟩, e divulgamos em mídias sociais.

Na plataforma, exibimos o conceito de sexismo considerado por este trabalho, o comentário que deve ser rotulado pelo usuário, o título da notícia associada ao comentário e botões de "sim" e "não" para que o usuário avalie o comentário como sexista ou não respectivamente. Não incluímos a opção "não sei" para forçar os rotuladores a sempre escolher uma das classes mesmo que considere o comentário difícil de ser rotulado.

Sempre que o usuário avalia um comentário, a plataforma exibe um novo comentário que ainda não foi rotulado pelo usuário logado. A exibição dos comentários é baseada na quantidade de votos que um comentário já recebeu, comentários com menos votos tem a maior prioridade de exibição, seguidos por comentários com empate no número de votos.

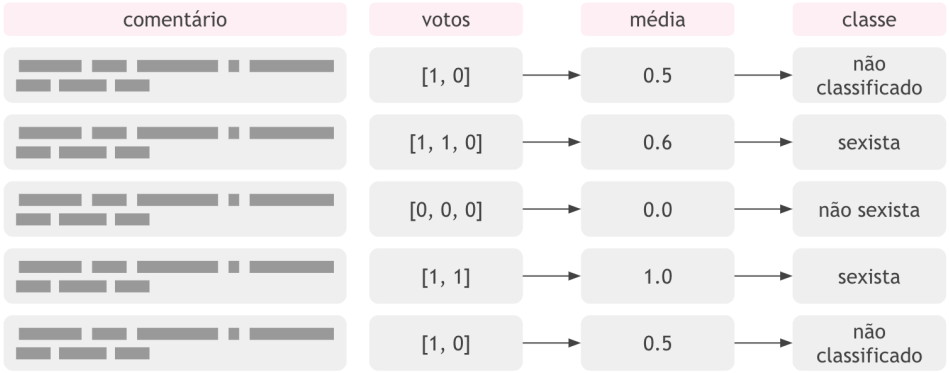

Figura 1. Fluxo de classificação final dos comentários a partir dos votos atribuídos.

A rótulo final de cada comentário foi atribuído considerando os votos "sim" com peso 1 e os votos "não" com peso 0, a média dos votos define a qual classe o comentário pertence, como mostra a figura 1. Comentários com a média de votos acima de 0,5, foram considerados como sexistas, os comentários com média abaixo de 0,5 foram considerados como não sexistas. Não atribuímos rótulo para comentários com média de votos igual a 0,5.

Esse processo gerou uma base de dados rotulada manualmente com 1.397 comentários sexistas, 1.275 comentário não sexistas e 500 comentários que não tiveram rótulo atribuído.

3.3. Análise dos comentários

Realizamos uma análise do corpus a fim de identificar características distintivas de cada classe de comentários.

Dada nossa coleção D de comentários coletados e rotulados, temos uma subcoleção D_s, composta por comentários sexistas, e uma subcoleção D_n, composta por

comentários não sexistas. Seja σ o vocabulário dos termos presentes em D, calculamos as frequências normalizadas F_s e F_n para cada termo $t_i \in \sigma$.

Calculamos o valor de $F_s - F_n$ de cada termo para encontrar palavras relevantes na diferenciação das duas classes de comentários. A tabela 2 mostra as quinze palavras com maior valor de $F_s - F_n$ que encontramos.

Palavra	F_s	F_n	$F_s - F_n$
de	0.043642	0.036990	0.006652
mulheres	0.010109	0.006892	0.003217
homens	0.006739	0.004079	0.002660
ela	0.007592	0.005063	0.002528
mulher	0.011246	0.008767	0.002479
feia	0.002639	0.000234	0.002404
assédio	0.002639	0.000891	0.001748
as	0.009459	0.007736	0.001724
na	0.008079	0.006376	0.001703
homem	0.005684	0.004079	0.001605
elas	0.002273	0.000703	0.001570
feminismo	0.002720	0.001172	0.001548
feministas	0.002030	0.000563	0.001467
uma	0.011733	0.010267	0.001465
feminista	0.001908	0.000563	0.001345

Tabela 2. Relação das quinze palavras com diferença positiva mais significante entre as frequências de ocorrência em cada coleção de comentários.

Pela tabela 2, notamos que algumas das palavras mais relevantes no discurso sexista são tradicionalmente *stopwords*, como "uma", "ela", "elas", "as' e "de". Essa última preposição foi usada em comentários da base que atribuem características ou obrigações à mulheres como em "*Mimimi é especialidade de feministas...*" e "*Toda mulher tem sim a obrigação de respeitar e atender o chefe da casa...*".

Já as *stopwords* "uma", "ela", "elas", "as" são artigos e pronomes femininos que aparecem como relevantes na diferenciação de discursos pois os comentários sexistas da base são direcionados a mulheres, como "*Cada uma sabe o risco que corre quando é negligente com seu macho e protetor*" ou "*Quando elas crescerem eu contrato elas pra fazerem uma faxina aqui em casa*".

4. Resultados

A caracterização da base se dá não só pela classificação de cada comentário mas também pela identificação dos motivos pelo qual cada comentário está uma classe, por isso dividimos os resultados em análise dos votos atribuídos e análise dos comentários.

4.1. Análise dos votos atribuídos

Uma vez que a plataforma para classificar comentários foi divulgada em diversas redes sociais, 247 pessoas mostraram interesse em se cadastrar no site rotular comentários, das quais aproximadamente 57% são do gênero feminino e os 43% restantes são do gênero

masculino. Esse fato mostra que a maior parte dos interessados em colaborar com uma pesquisa sobre “sexismo” são mulheres.

Os usuários da plataforma online poderiam votar em quantos comentários quisessem, e no final coletamos 7.089 votos. Todos os comentários da base receberam pelo menos um voto, apenas um comentário recebeu quatro votos, 1.168 comentários receberam três votos, 1.581 comentários receberam dois votos e 419 comentários ficaram com apenas um voto.

Dos votos atribuídos a cada comentário, 67,2% foram de pessoas do gênero feminino, enquanto os 32,8% restantes, foram atribuídos por pessoas do gênero masculino. Esse fato reforça a hipótese de que as mulheres se sentem mais motivadas do que os homens ao colaborar com esse tipo de pesquisa.

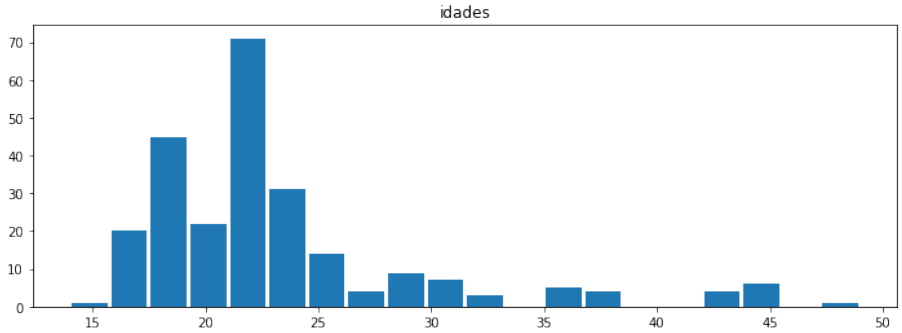

Figura 2. Gráfico de barras da distribuição de idades dos usuários cadastrados plataforma de votação.

A média de votos atribuídos por usuário é de 61 votos, e dos 247 usuários, 11 votaram apenas em um comentário, mas nenhum dos usuários teve zero votos. Pelo gráfico da figura 2, podemos ver a distribuição de idades dos rotuladores, com o gráfico vemos que a faixa etária de 20 à 25 anos tem a maior quantidade de rotuladores.

As figuras 3 e 4 mostram o volume de votos distribuídos por idade e gênero. Podemos visualizar que as mulheres tiverem um maior volume de votos na faixa etária de 17 à 25 anos, enquanto os votos masculinos ficaram mais distribuídos entre as faixas etárias.

Observamos no gráfico da figura 3 que as mulheres têm um volume maior de votos corretos para a classe sexista em relação aos homens, mas na figura 4 podemos ver que elas também tem uma porcentagem maior de votos em falsos positivos. Isso pode ser explicado pelo fato de que as mulheres são mais sensíveis à conteúdos sexistas do que os homens, dada a percepção mais recente que a mulheres têm da necessidade de combate ao machismo.

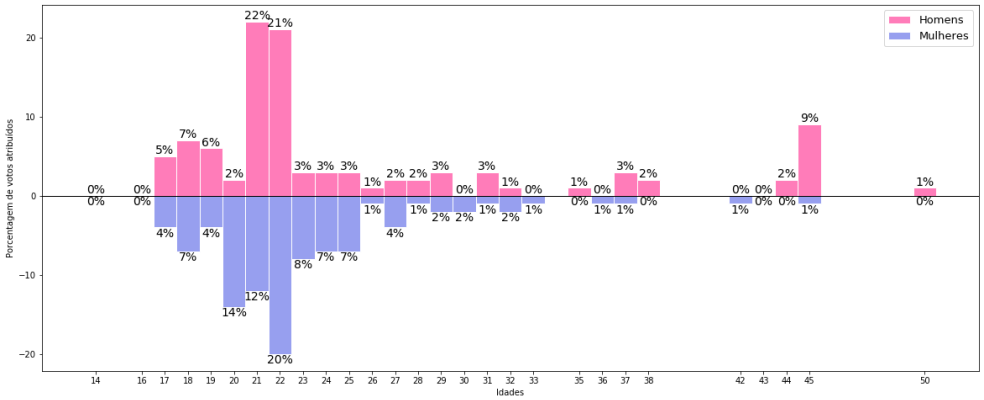

Figura 3. Gráfico de barras relacionando as porcentagens dos votos corretos para a classe "sexista" separado por gênero e distribuído por idades.

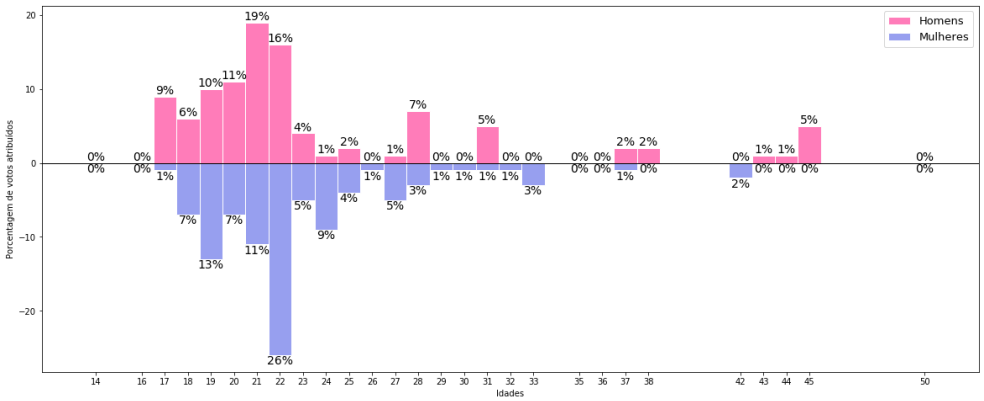

Figura 4. Gráfico de barras relacionando as porcentagens dos votos incorretos para a classe "sexista" separado por gênero e distribuído por idades.

Nesta seção, identificamos que mulheres classificam mais comentários como sexistas do que os homens. Uma hipótese sobre esse fato é que a tolerância de homens ao discurso sexista é maior do que a das mulheres. O comentário "*Ela é muito linda. Merece ser paquerada e amada. Em todos os sentidos*" cai no conceito de sexismos benevolente e homens podem ter dificuldade de identificar o desconforto que comentários como esse causam em algumas mulheres.

4.2. Análise dos comentários

Dos 3.172 comentários coletados, 1.397 foram classificados como sexistas, 1.275 como não sexistas e 500 não receberam rótulo.

Um exemplo de comentário rotulado como sexista é "*Mimimi é especialidade de feministas, sempre irão problematizar alguma coisa.*", e um exemplo de comentário rotulado como não sexista é "*Ninguém é de ninguém! Por favor não venham com essas histórias de que a mulher é do homem, nunca foi é nunca será!!!*".

Dentre os comentários que não receberam rótulo, estão os comentários "*É difícil julgar uma pessoa, acho que teria que fazer um estudo mais aprofundado para saber*

o porque desses assassinatos brutais, não é apenas um 'não da mulher' que motiva ele a mata-la, tem algo mais nisso." e "*Feminismo, hoje em dia, é tratado como piada no mundo inteiro. Feminismo nunca produziu nada, e apenas adere ao que outros produzem.*".

Para coletar características sobre as classes de comentários, calculamos a mediana da quantidade de caracteres M_c e a mediana da quantidade de palavras M_p para cada uma das classes de comentário possíveis e também para os comentários que não foram classificados e registramos os resultados na tabela 3.

Classe	M_c	M_p
Sexista	97	17
Não sexista	88	16
Não rotulado	78	14

Tabela 3. Relação das medianas para quantidades de caracteres e de palavras em cada classe de comentários.

Pelos dados apresentados na tabela 3, podemos assumir que os comentários que geram divergência de votos entre os rotuladores são mais curtos que os comentários que receberam classificação, o que indica que os comentários com menos palavras e caracteres tendem a ter menos contexto e podem gerar ambiguidades na interpretação do leitor.

Um exemplo de comentário não rotulado que possui poucas palavras é "*Já estou vendo a Fátima Bernardes convidando ela para falar no programinha matinal dela*", a frase não possuí nenhuma palavra ofensiva, mas dá a entender que o autor está menosprezando a mulher a qual se refere a notícia, problema que também foi detectado nos trabalhos de Kwok&Wang [9] e Davidson et al. [8].

Para obter informações sobre o tipo de comentário que gera mais engajamento, montamos um gráfico com a soma de *likes* e *dislikes* para cada classe de comentários, que pode ser visualizado na figura 5.

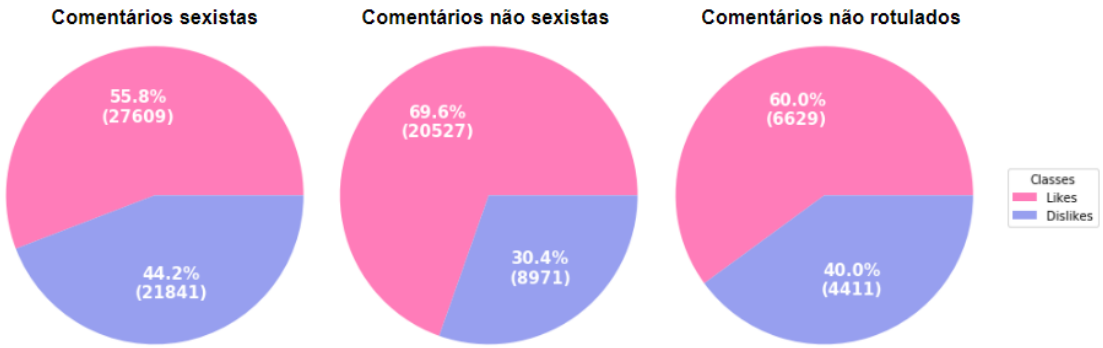

Figura 5. Gráficos de pizza representando as quantidades de *likes* e *dislikes* para cada classe de comentários.

O gráfico da figura 5 mostra que os comentários sexistas geram mais engajamento do que os demais comentários, o que pode ser um efeito do impacto que eles geram

Classe	Modelo	*P*	*R*	*F1*
Sexista	SVM	0,97	0,87	0,92
Não sexista	SVM	0,88	0,87	0,92
Sexista	KNN	0,88	0,98	0,93
Não sexista	KNN	0,98	0,88	0,93

Tabela 4. Relação das quinze palavras com diferença positiva mais significante entre as frequências de ocorrência em cada coleção de comentários.

em discussões sobre violência contra mulheres, assédio e feminicídio, que são temas das notícias selecionadas para coleta de comentários.

Na figura 5, podemos ver também que os comentários não rotulados possuem menos engajamento do que os demais, o que pode indicar que os leitores não compreendem a opinião expressa nesses comentários ou que sua relevância no contexto é baixa. Ainda assim, os comentários não rotulados possuem uma distribuição de *likes* e *dislikes* similar a mesma distribuição para os comentários sexistas.

Ainda no gráfico da figura 5, notamos que todas as classes de comentários geram mais *likes* do que *dislikes*, mas a diferença absoluta entre a quantidade de *likes* e *dislikes* dos comentários sexistas é menor do que nos comentários não sexistas, umas vez que comentários sexistas recebem mais engamento negativo.

Utilizamos no nosso *dataset* os métodos de aprendizagem Support Vector Machine (SVM) e K-Nearest Neighbors (KNN), a fim de validar a possibilidade da classificação automática dos comentários.

Os parâmetros de cada modelo foram escolhidos utilizando Grid Search, de forma que as características eram compostas pelo *Term Frequency (TF)* dos 100 termos com maior valor de $F_s - F_n$. Testamos cada modelo usando a validação cruzada de 5 vezes, mantendo 20% da amostra para teste, os resultados obtidos estão listados na tabela 4 com as métricas de precisão (P), revocação (R) e F1 para cada modelo. Toda modelagem foi realizada usando o scikit-learn [13].

Os resultados exibidos na tabela 4 mostram que a as palavras escolhidas como relevantes para o discurso sexista a partir do valor de $F_s - F_n$ são representativas na distinção das classes sexista e não sexista.

5. Conclusão

Neste artigo, apresentamos a caracterização de uma base de dados composta por comentários sexistas e não sexistas em portais de notícias. Com os comentário classificados manualmente por pessoas de idades e gêneros distintos, estudamos as características dos discursos para determinar quais delas podem ser utilizadas na diferenciação entre as classes de comentários aqui consideradas. Também realizamos o estudo dos votos atribuídos por faixa etária e gênero, a fim de construir uma análise da influência que essas características tem no voto final de cada rotulador.

As análises realizadas evidenciaram que a tolerância dos homens ao sexismo é menor que a das mulheres ao rotular manualmente um comentário, uma vez que o alvo do discurso sexista são as mulheres, o que se mostra verdade pela presença de artigos e

pronomes femininos como palavras mais relevantes do discurso sexista. Utilizamos a base de dados construída com 2.672 comentários rotulados e vimos que a detecção automática desse tipo de discurso é viável considerando o TF das palavras mais frequentes no discurso sexista. Utilizando o SVM, obtivemos precisão de aproximadamente 0,92, e com o KNN esse valor foi de 0,93.

Como trabalhos futuros, pretendemos seguir o estudo da base de dados a fim de encontrar novas características que diferenciam as duas classes de discurso. Pretendemos avaliar como características os engajamento recebido nos comentários e também a quantidade de palavras em cada comentário. Por fim, pretendemos realizar um estudo comparativo entre os resultados de classificadores automáticos quando utilizados nesta base de dados, a fim de ter uma identificação correta do discurso sexista em novos comentários a serem avaliados.

Referências

1 SMIGAY, K. E. V. Sexismo, homofobia e outras expressões correlatas de violência: desafios para a psicologia política. *Psicologia em revista*, v. 8, n. 11, p. 32–46, 2002. 2, 3

2 GLICK, P.; FISKE, S. T. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In: *Social Cognition*. [S.l.]: Routledge, 2018. p. 116–160. 2, 3

3 VIANNA, J.; HISING, E. *Homem é condenado a 41 anos de prisão por crimes como racismo, terrorismo e divulgação de pedofilia na internet*. G1, 2018. Disponível em: ⟨https://glo.bo/2sjxAJ5⟩. Acesso em: 30 de abril de 2020. 2

4 ROSSI, M. *Mulher espancada após boatos em rede social morre em Guarujá, SP*. G1, 2014. Disponível em: ⟨https://glo.bo/37GfDVv⟩. Acesso em: 30 de abril de 2020. 2

5 OLIVEIRA, S. *Adolescente vítima de bullying se suicida por 'não aguentar mais'*. Redação Amazonas1, 2018. Disponível em: ⟨http://bit.ly/2st6sHX⟩. Acesso em: 30 de abril de 2020. 2

6 MARQUES, J. J.; SANTOS, J. L. dos. *Mapa da Violência Contra a Mulher*. [S.l.]: CMULHER, 2018. 2

7 BANKS, J. Regulating hate speech online. *International Review of Law, Computers Technology*, p. 233–239, 2010. 2

8 DAVIDSON, T.; WARMSLEY, D.; MACY, M. Automated hate speech detection and the problem of offensive language. *Eleventh International AAAI Conference on Web and Social Media*, 2017. 2, 3, 4, 9

9 KWOK, I.; WANG, Y. Locate the hate: Detecting tweets against blacks. In: *Twenty-seventh AAAI conference on artificial intelligence*. [S.l.: s.n.], 2013. 2, 3, 9

10 BADJATIYA, P.; GUPTA, S.; GUPTA, M. Deep learning for hate speech detection in tweets. p. 759–760, 2017. 3

11 PARK, J. H.; FUNG, P. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*, 2017. 3

12 PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. [S.l.], 2002. p. 79–86. 3

13 PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011. 10