Submission date: 15/02/2021
Resubmission date: 25/05/2021
Camera ready submission date: 08/07/2021

1st round notification: 22/04/2021
2nd round notification: 24/06/2021
Available online: 20/08/2021
Publication date: 20/08/2021

Section: extended version
Extended versions of selected papers from the IX Brazilian Workshop on Social Network Analysis and Mining

# Criação e Caracterização de um Corpus de Discurso Sexistas em Português

# Title: Creation and Characterization of a Sexist Discourse Corpus in Portuguese

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

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

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

***Abstract.*** *Sexism is a topic whose social interest has grown as the female figure overcomes barriers of gender inequality. Sexist discourse propagates and encourages derogatory and abusive behavior against women. Accurate characterization and identification are key for treating and mitigating violence. In this work, we present a corpus of sexist discourse in Portuguese collected from news portals of great popular acceptance. 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.*

***Keywords.*** *Sexism, Hate-speach, Data Science*

***Resumo.*** *O sexismo é um tópico cujo interesse social tem crescido a medida que a figura feminina vence as barreiras da desigualdade de gênero. O discurso sexista propaga e incentiva o comportamento depreciativo e abusivo contra mulheres. Uma caracterização e identificação precisa são peças-chave para tratar e mitigar a violência. Neste trabalho, apresentamos um corpus de discurso sexista em Português coletado a partir de portais de notícias de grande aceitação popular. 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.*

***Palavras-Chave.*** *Sexismo, Discurso de ódio, Ciência de dados*

Cite as: Braga, M. L. P., Nakamura, F. G. & Nakamura, E. F. (2021). Creation and Characterization of a Sexist Discourse Corpus in Portuguese (Criação e Caracterização de um Corpus de Discurso Sexistas em Português). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 14(2), 79-95. DOI: 10.5753/isys.2021.1797

## 1. Introdução

O sexismo consiste em atos e discursos que ofendem, agridem ou diminuem as pessoas de um gênero [Von Smigay 2002, Glick and Fiske 2018]. 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 [Vianna and Hising 2018].

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* [Rossi 2014]. 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 [Oliveira 2018]. 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 [Marques and dos Santos 2018] 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 [Banks 2010] 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. [Davidson et al. 2017] e Kowk&Wang [Kwok and Wang 2013].

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. [Badjatiya et al. 2017] e Park&Fung [Park and Fung 2017] 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 [Glick and Fiske 2018] e Smigay&Ellen [Von Smigay 2002], 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. [Davidson et al. 2017].

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 [Glick and Fiske 2018] 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 [Von Smigay 2002] 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 [Kwok and Wang 2013] 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 traz o contexto necessário para que o classificador seja eficaz.

Já no trabalho de Peng et al. [Pang et al. 2002], 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

| 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.**

superiores ao uso da frequência de bigramas para o contexto do seu trabalho com 82,9% de acurácia.

Davidson et al. [Davidson et al. 2017] 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

O nosso estudo tem início na criação da base de dados, que se dá em três etapas: coleta de comentários, classificação manual dos 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. As notícias selecionadas para participar da coleta podem ser consultadas em `https://sexismo.vercel.app/news`.

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 que aparece em uma notícia referente a uma artista que se afastou da carreira para se dedicar ao feminismo. Na mesma notícia, encontramos comentários não sexistas, como "*O feminismo prega a igualdade dos gêneros. Infelizmente muitas mulheres não compreendem isso... Acham que é mimimi. Comparam a vida da mulher há 50 anos atrás com agora.*".

### 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 `https://sexismo.vercel.app`, 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. Nos casos em que o comentário responde a algum outro comentário na mesma notícia, exibimos os dois comentários, para que o rotulador tenha contexto do que está sendo tratado no texto.

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, assim evitamos que a maioria dos comentários da base tenha rótulo indefinido.

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.

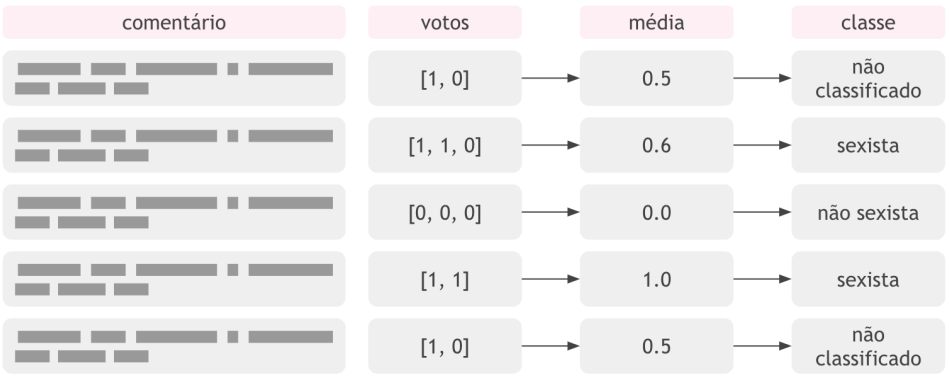


**Figura 1. Fluxo de classificação final dos comentários a partir dos votos atribuídos.**

Cada comentário recebe rótulos de mais de um usuário e o rótulo final de cada comentário é 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.

## 4. Caracterização da base de dados

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 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 $\sigma$ o vocabulário dos ngramas presentes em $D$, calculamos as frequências normalizadas $F_s$ e $F_n$ para cada ngrama $t_i \in \sigma$.

Os termos com $F_s$ maior que $F_n$ possuem maior relevância para o discurso sexista, sendo assim calculamos o valor de $F_s - F_n$ de cada ngrama em busca das palavras relevantes na diferenciação das duas classes de comentários. A tabela 2 mostra os quinze unigramas com maior valor de $F_s - F_n$ que encontramos, sendo os unigramas mais relevantes para o discurso sexista.

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*".

O termo "feia" chama atenção pois sua aparição na lista da tabela 2 indica que em contextos de notícias onde o tópico se refere a mulheres, assédio ou feminicídio, temos comentários com adjetivos pejorativos que buscam ofender mulheres.

Na tabela 3 são exibidos os unigramas mais relevantes no discurso não sexista. Os termos presentes nessa listagem não apontam referência à gênero e não possuem relação

| **Unigrama** | $F_s$ | $F_n$ | $F_s - F_n$ |
|---|---|---|---|
| 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 dos quinze unigramas com diferença positiva mais significante entre as frequências de ocorrência em cada coleção de comentários.**

semântica.

Além dos unigramas, calculamos a diferença das frequências para os bigramas mais frequentes em cada discurso, exibidos nas tabelas 4 e 5 respectivamente. A tabela 4 mostra que para a maioria dos bigramas mais frequentes no discurso sexista são relacionados à atribuição de características a um grupo social relacionado ao gênero, enquanto na tabela 5 temos bigramas sem relação semântica relevante entre si.

O bigrama "as feministas" aparece em comentários como *"as feministas são contra as cantadas porque, como não recebem cantadas, não querem que outras recebam*", "*a maioria das feministas são mulheres feias, que não consegue macho e ai ficam criticando*" ambos os exemplos tem o objetivo de diminuir mulheres feministas.

Em muitos casos, bigramas relacionados ao gênero feminino e ao gênero masculino são utilizados em conjunto para expressar opiniões que subordinam mulheres ao homem, como acontece em "*as mulheres só existem em função do homem, pois foram criadas através da costela de Adão segundo as escrituras sagradas!*" ou em "*mulher com homem mais velho e normal pois o homem serve e protege a mulher*".

O bigrama "falta de" chama atenção por não ter nada que remeta à gêneros e mesmo assim estar entre os mais frequentes do texto sexista. Este bigrama aparece em comentário ofensivos como "*a falta de beleza dela ficou em evidência*", "*a falta de homem no mercado leva as mulheres a se envolverem com esse tipo de gente*" e "*feminismo é falta de r****".

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*

| **Unigrama** | $F_s$ | $F_n$ | $F_s - F_n$ |
|---|---|---|---|
| ser | 0.005805 | 0.008533 | -0.002727 |
| não | 0.021923 | 0.024332 | -0.002409 |
| que | 0.044495 | 0.046601 | -0.002106 |
| Brasil | 0.001096 | 0.002813 | -0.001717 |
| sua | 0.001543 | 0.003188 | -0.001645 |
| presidente | 0.000365 | 0.001922 | -0.001557 |
| comentários | 0.000284 | 0.001688 | -0.001404 |
| lei | 0.000650 | 0.002016 | -0.001366 |
| em | 0.008810 | 0.010127 | -0.001317 |
| ainda | 0.001380 | 0.002672 | -0.001292 |
| está | 0.002273 | 0.003563 | -0.001290 |
| você | 0.001665 | 0.002907 | -0.001242 |
| sobre | 0.000528 | 0.001688 | -0.001160 |
| humano | 0.000244 | 0.001313 | -0.001069 |
| pessoas | 0.001259 | 0.002157 | -0.000898 |

**Tabela 3. Relação dos quinze unigramas com diferença negativa mais significante entre as frequências de ocorrência em cada coleção de comentários.**

*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.*".

O vocabulário referente à base de dados era composto por 12.200 palavras, incluindo 153 *stopwords*, 79 palavras contendo caracteres numéricos e variações das mesmas palavras.

Dentre os termos formados por letras e números, temos os exemplos "BRUX4", "FEMIB0STAN1SM0", "Saf4da" e "ordin4ria", que podem representar tentativas dos comentarista de ofender mulheres sem ser detectado por algoritmos de censura dos portais de notícia. Apesar de serem termos claramente ofensivos, os comentaristas podem variá-los de formas diversas como "0rdin4ria" e "0rd1nar14" dificultando que o algoritmo detecta comentários com esses termos como sendo ofensivos, nesse caso podemos assumir que termos compostos por números e letras provavelmente são sexistas.

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 6.

| **Bigrama** | $F_s$ | $F_n$ | $F_s - F_n$ |
|---|---|---|---|
| os homens | 0.006230 | 0.002544 | 0.003686 |
| as mulheres | 0.010619 | 0.007314 | 0.003304 |
| que as | 0.005380 | 0.002226 | 0.003154 |
| um homem | 0.003398 | 0.000954 | 0.002444 |
| uma mulher | 0.007221 | 0.004929 | 0.002291 |
| mulheres que | 0.003823 | 0.001749 | 0.002074 |
| são minoria | 0.001982 | 0.000159 | 0.001823 |
| as feministas | 0.002265 | 0.000477 | 0.001788 |
| não tem | 0.005097 | 0.003498 | 0.001599 |
| de mulheres | 0.002690 | 0.001113 | 0.001577 |
| dos homens | 0.002265 | 0.000795 | 0.001470 |
| do homem | 0.002407 | 0.000954 | 0.001453 |
| falta de | 0.002690 | 0.001272 | 0.001418 |
| com essa | 0.001841 | 0.000477 | 0.001364 |
| em um | 0.001841 | 0.000477 | 0.001364 |

**Tabela 4. Relação dos quinze bigramas com diferença positiva mais significante entre as frequências de ocorrência em cada coleção de comentários.**

Pelos dados apresentados na tabela 6, 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 [Kwok and Wang 2013] e Davidson et al. [Davidson et al. 2017].

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 2.

O gráfico da figura 2 mostra que os comentários sexistas geram mais engajamento do que os demais comentários, o que pode ser um efeito da polêmica que eles geram em discussões sobre violência contra mulheres, assédio e feminicídio, que são temas das notícias selecionadas para coleta de comentários.

Na figura 2, 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 2, notamos que todas as classes de comentários geram mais *likes* do que *dislikes*, mas a diferença absoluta entre a quantidade de *likes* e *dislikes*

| **Bigrama** | $F_s$ | $F_n$ | $F_s - F_n$ |
|---|---|---|---|
| ser humano | 0.000708 | 0.003657 | -0.002949 |
| que não | 0.008212 | 0.010495 | -0.002283 |
| de sua | 0.000566 | 0.002067 | -0.001501 |
| sabe que | 0.000425 | 0.001749 | -0.001324 |
| caso de | 0.000142 | 0.001431 | -0.001289 |
| as pessoas | 0.001416 | 0.002703 | -0.001287 |
| que você | 0.000991 | 0.002226 | -0.001235 |
| do Brasil | 0.000566 | 0.001749 | -0.001183 |
| todo mundo | 0.000425 | 0.001590 | -0.001165 |
| que ninguém | 0.000283 | 0.001431 | -0.001148 |
| só para | 0.000283 | 0.001431 | -0.001148 |
| maioria dos | 0.000142 | 0.001272 | -0.001130 |
| Jout Jout | 0.000142 | 0.001272 | -0.001130 |
| redes sociais | 0.000142 | 0.001272 | -0.001130 |
| de ser | 0.002124 | 0.003180 | -0.001056 |

**Tabela 5. Relação dos quinze bigramas com diferença negativa mais significante entre as frequências de ocorrência em cada coleção de comentários.**

| **Classe** | $M_c$ | $M_p$ |
|---|---|---|
| Sexista | 97 | 17 |
| Não sexista | 88 | 16 |
| Não rotulado | 78 | 14 |

**Tabela 6. Relação das medianas para quantidades de caracteres e de palavras em cada classe de comentários.**

dos comentários sexistas é menor do que nos comentários não sexistas, umas vez que comentários sexistas recebem mais engajamento negativo.

Na tentativa de aproximar os comentários não rotulados de uma das classes estudadas, calculamos a distância de Jensen-Shannon [Fuglede and Topsoe 2004] dos comentários não rotulados em relação às palavras mais relevantes para cada classe de comentários a partir do valor de $F_s - F_n$, de forma que $H_s$ é a distribuição das frequências de palavras com $F_s - F_n$ positivo e $H_n$ é a distribuição das frequências de palavras com $F_s - F_n$ negativo.

Pelo gráfico da figura 3, podemos verificar que a divergência dos comentários não rotulados em relação a $H_s$ e $H_n$ é alta, e que uma porção desses comentários tem valores mais próximos ao rótulo de "sexista" do que ao rótulo de "não sexista".

### 4.2. 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 dado mostra que a maior parte dos interessados em colaborar com uma

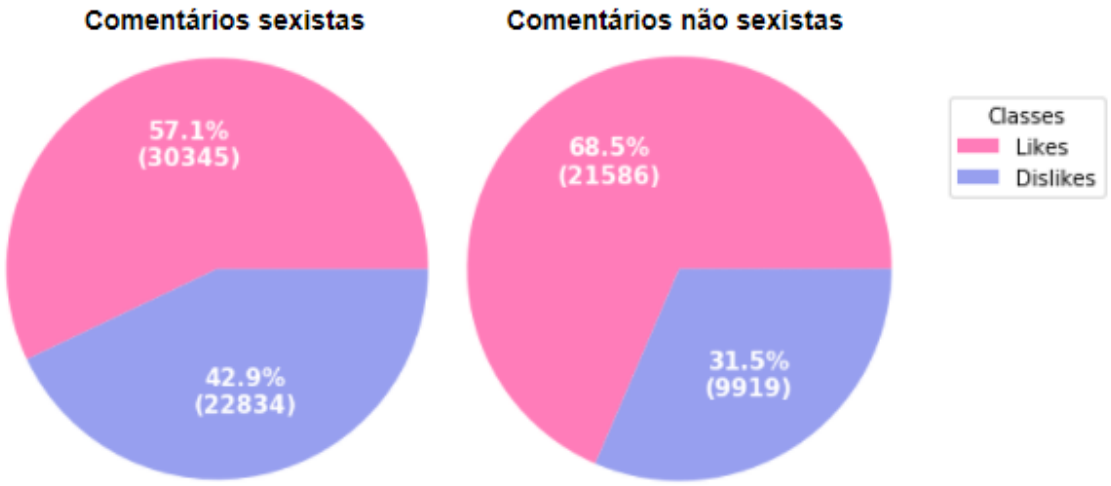


**Figura 2. Gráficos de pizza representando as quantidades de *likes* e *dislikes* para cada classe de comentários.**

pesquisa sobre "sexismo" foram 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 sentiram mais motivadas do que os homens ao colaborar a pesquisa.

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 4, podemos ver a distribuição de idades dos rotuladores, nele vemos que a faixa etária de 18 à 24 anos tem a maior quantidade de rotuladores. Podemos notar também que a quantidade de mulheres é superior a quantidade de homens na maioria das faixas etárias.

As figuras 5, 6, 7 e 8 mostram o volume de votos distribuídos por idade e gênero dos rotuladores. Na análise dessas imagens consideramos que votos "corretos" são os votos atribuídos que correspondem ao rótulo final de um comentário, enquanto os votos "incorretos" correspondem aos votos atribuídos para a classe contrária ao rótulo final de um comentário.

Na figura 5 observamos os votos "corretos" para a classe sexista. Podemos observar pelo gráfico que os "acertos" em comentários sexistas estão bem equilibrados entre homens e mulheres, embora as mulheres acertem um pouco mais que os homens. Nas faixas etárias de 28 anos para cima, observamos que os homens acertam mais que as mu-

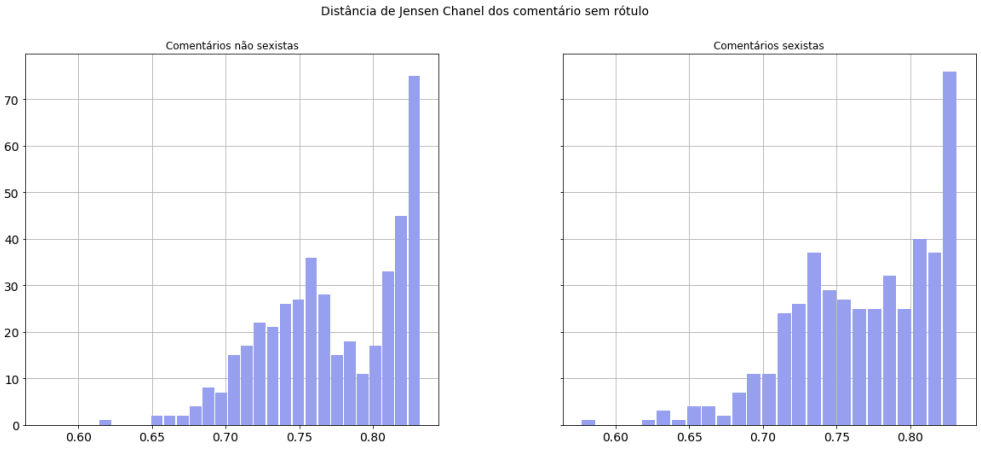


**Figura 3. Gráfico de barras relacionando as quantidades de comentários em cada faixa de valores para JSD quando comparados com $H_s$ e $H_n$.**

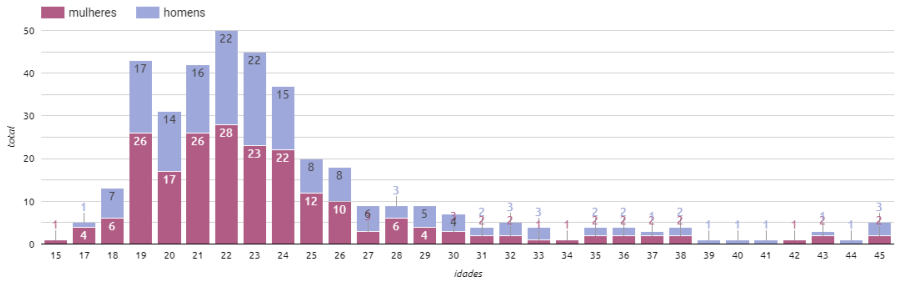


**Figura 4. Gráfico de barras da distribuição de idades dos usuários cadastrados plataforma de votação separados por gênero.**

lheres com mais frequência, o que pode indicar uma diferença na tolerância ao sexismo entre mulheres mais novas e mais velhas.

A figura 6 mostra os votos "incorretos" para a classe sexista. Na imagem observamos que as mulheres possuem um maior número de falsos positivos, esse dado mostra que mulheres adotam um conceito de sexismo diferente do conceito adotado pelos homens, uma vez que o sexismo não se resume apenas a ofensas a figura feminina, mas também inclui o sexismo benevolente [Glick and Fiske 2018], que raramente é percebido por pessoas do gênero masculino como discurso sexista.

Os gráficos das figuras 7 e 8 exibem a quantidade de votos "corretos" e "incorretos" para a classe não sexista respectivamente. Observamos que, similar ao fenômeno que ocorre com as mulheres, os homens mostram uma taxa alta de falsos positivos para a classe não sexista. Podemos destacar as faixas etárias de 21 e 45 anos, onde os rotuladores erraram com alta frequência ao afirmar que os comentários eram não sexistas.

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, por conta do conceito que cada gênero adota de sexismo. O comentário "*Ela é muito linda. Merece ser paquerada e amada.*

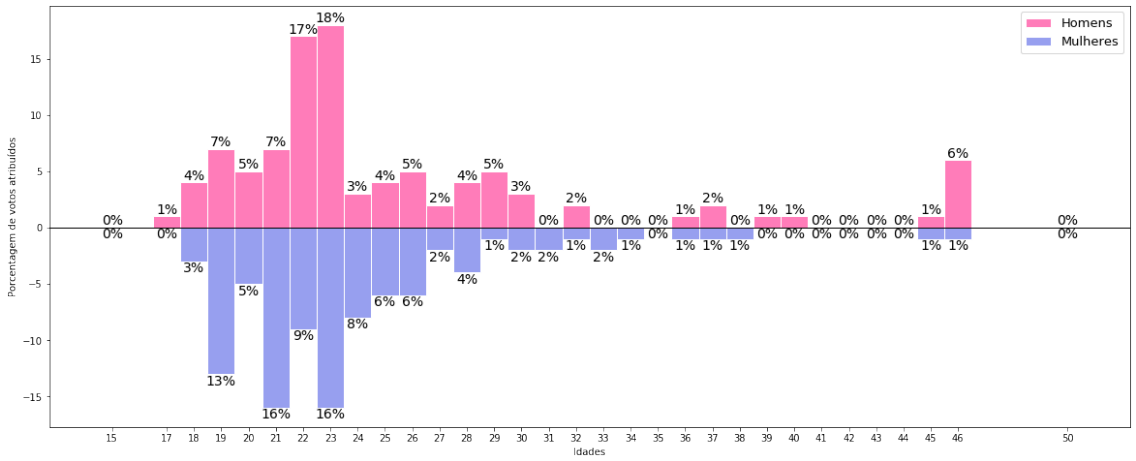


**Figura 5. Gráfico de barras relacionando as porcentagens dos votos corretos para a classe "sexista" separado por gênero e distribuído por idades.**

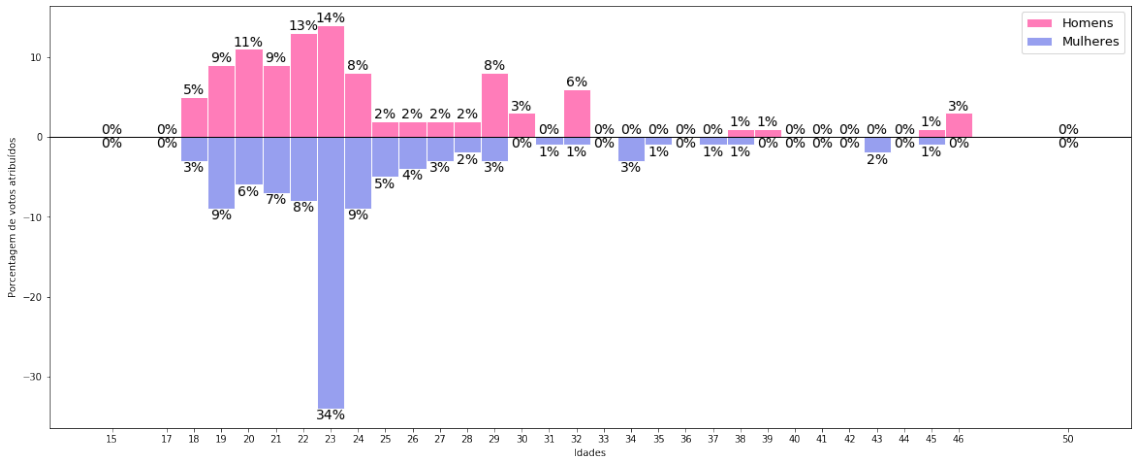


**Figura 6. Gráfico de barras relacionando as porcentagens dos votos incorretos para a classe "sexista" separado por gênero e distribuído por idades.**

*Em todos os sentidos*" cai no conceito de sexismo benevolente e homens podem ter dificuldade de identificar o desconforto que comentários como esse causam em algumas mulheres.

## 5. Resultados

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

Os parâmetros de cada modelo foram escolhidos utilizando Grid Search, usando combinações distintas das características listadas abaixo:

- *Term Frequency (TF)* dos 100 unigramas com maior valor de $F_s - F_n$ ($TF_{us}$)
- *Term Frequency (TF)* dos 100 unigramas com menor valor de $F_s - F_n$ ($TF_{un}$)
- *Term Frequency (TF)* dos 100 bigramas com maior valor de $F_s - F_n$ ($TF_{bs}$)

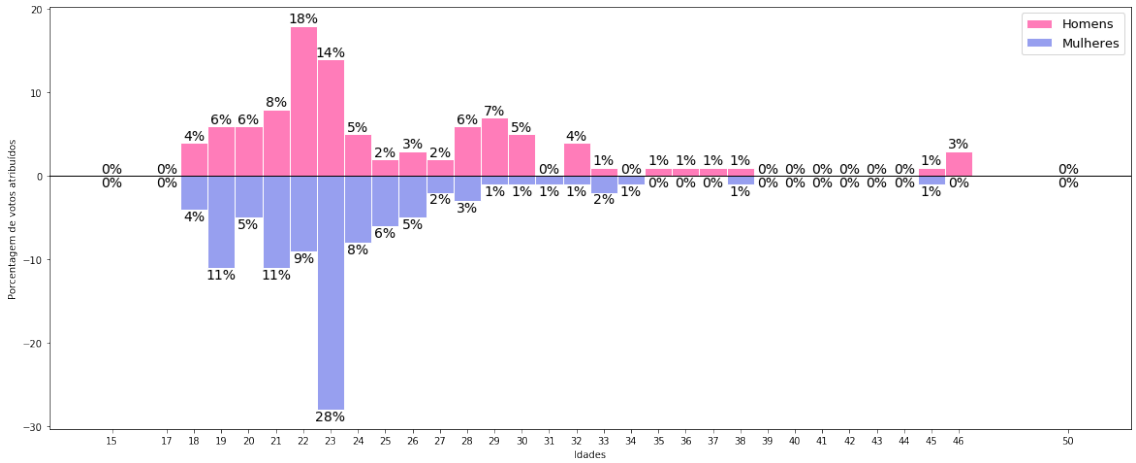


**Figura 7. Gráfico de barras relacionando as porcentagens dos votos corretos para a classe "não sexista" separado por gênero e distribuído por idades.**

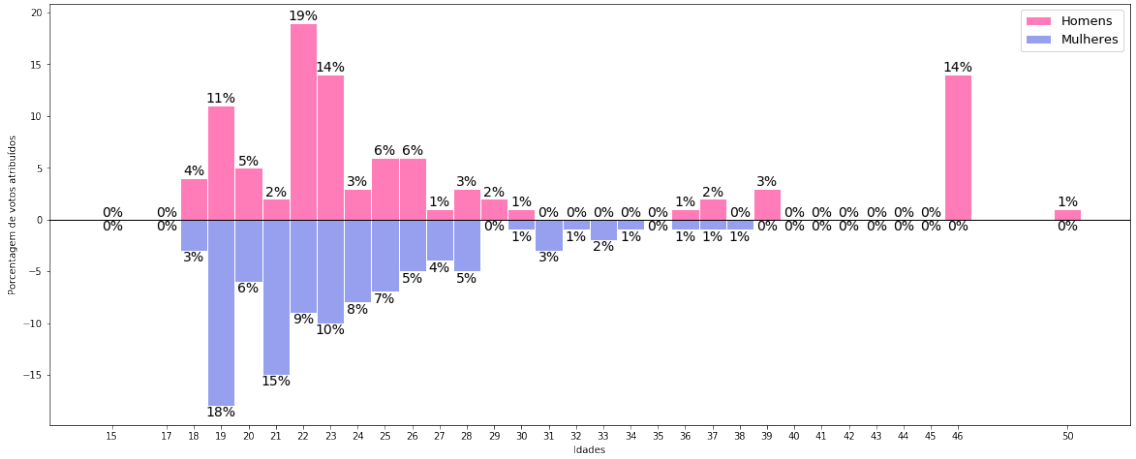


**Figura 8. Gráfico de barras relacionando as porcentagens dos votos incorretos para a classe "não sexista" separado por gênero e distribuído por idades.**

- *Term Frequency (TF)* dos 100 bigramas com menor valor de $F_s - F_n$ ($TF_{bn}$)
- Quantidade de caracteres dos comentários ($Q_c$)
- Quantidade de palavras dos comentários ($Q_p$)
- Quantidade de *likes* dos comentários ($Q_l$)
- Quantidade de *dislikes* dos comentários ($Q_d$)

Testamos cada modelo usando a validação cruzada de 10 folds, mantendo 20% da amostra para teste, os resultados obtidos estão listados na nas figuras 9, 10 e 11 com as métricas de precisão (P), revocação (R) e F1 para cada modelo. Toda modelagem foi realizada usando o scikit-learn [Pedregosa et al. 2011].

Podemos ver pela figura 9 que para o SVM, os melhores resultados foram obtidos com o uso das características (1) e (2) combinadas e com o uso características (5), (6), (7), e (8) combinadas. Pela figura 10 observamos que os melhores resultados para o KNN foram obtidos usando as características (5), (6), (7), e (8) combinadas. A figura 11

| | SVM | | | | | |
|---|---|---|---|---|---|---|
| | P | | R | | F1 | |
| classes | sexista | não sexista | sexista | não sexista | sexista | não sexista |
| TF de unigramas sexistas e não sexistas | 0.99766 | 0.91856 | 0.91929 | 0.99765 | 0.95683 | 0.95643 |
| TF de unigramas sexistas | 0.97531 | 0.86193 | 0.85750 | 0.97608 | 0.91256 | 0.91542 |
| TF de unigramas não sexistas | 0.83160 | 0.94887 | 0.96143 | 0.78588 | 0.89175 | 0.85958 |
| Caracteres | 0.67145 | 0.66840 | 0.72393 | 0.61059 | 0.69653 | 0.63792 |
| Palavras | 0.57922 | 0.57195 | 0.69643 | 0.44471 | 0.63241 | 0.50030 |
| Caractres + Palavras | 0.82822 | 0.82452 | 0.84321 | 0.80784 | 0.83557 | 0.81600 |
| Likes | 0.58900 | 0.56886 | 0.65750 | 0.49608 | 0.62127 | 0.52982 |
| Dislikes | 0.60660 | 0.56897 | 0.60929 | 0.56588 | 0.60769 | 0.56714 |
| Likes + Dislikes | 0.74964 | 0.67527 | 0.66893 | 0.75490 | 0.70689 | 0.71279 |
| Likes + Dislikes + Caracteres + Palavras | 0.98333 | 0.98625 | 0.98750 | 0.98157 | 0.98539 | 0.98388 |
| TF de bigramas sexistas e não sexistas | 0.94043 | 0.68708 | 0.60214 | 0.95804 | 0.73387 | 0.80015 |
| TF de bigramas sexistas | 0.87667 | 0.60851 | 0.45464 | 0.92980 | 0.59826 | 0.73552 |
| TF de bigramas não sexistas | 0.61844 | 0.86108 | 0.94714 | 0.35804 | 0.74823 | 0.50525 |

**Figura 9. Precisão, revoção e F1 para o SVM usando as features selecionadas.**

| | KNN | | | | | |
|---|---|---|---|---|---|---|
| | P | | R | | F1 | |
| classes | sexista | não sexista | sexista | não sexista | sexista | não sexista |
| TF de unigramas sexistas e não sexistas | 0.92413 | 0.97736 | 0.98071 | 0.91137 | 0.95153 | 0.94312 |
| TF de unigramas sexistas | 0.98123 | 0.86175 | 0.85643 | 0.98196 | 0.91454 | 0.91790 |
| TF de unigramas não sexistas | 0.81516 | 0.96888 | 0.97786 | 0.75608 | 0.88904 | 0.84913 |
| Caracteres | 0.61074 | 0.58905 | 0.65571 | 0.54118 | 0.63231 | 0.56392 |
| Palavras | 0.57190 | 0.54301 | 0.63036 | 0.48196 | 0.59961 | 0.51052 |
| Caractres + Palavras | 0.73996 | 0.72264 | 0.75179 | 0.70980 | 0.74577 | 0.71610 |
| Likes | 0.56486 | 0.52289 | 0.57000 | 0.51765 | 0.56737 | 0.52020 |
| Dislikes | 0.61660 | 0.53767 | 0.46500 | 0.68275 | 0.53002 | 0.60151 |
| Likes + Dislikes | 0.62399 | 0.58289 | 0.61214 | 0.59490 | 0.61791 | 0.58873 |
| Likes + Dislikes + Caracteres + Palavras | 0.99781 | 0.97264 | 0.97429 | 0.99765 | 0.98587 | 0.98494 |
| TF de bigramas sexistas e não sexistas | 0.86605 | 0.62822 | 0.50714 | 0.91373 | 0.63934 | 0.74446 |
| TF de bigramas sexistas | 0.79235 | 0.57491 | 0.40500 | 0.88314 | 0.53557 | 0.69636 |
| TF de bigramas não sexistas | 0.57953 | 0.74228 | 0.91429 | 0.27137 | 0.70936 | 0.39705 |

**Figura 10. Precisão, revoção e F1 para o KNN usando as features selecionadas.**

mostra que para o RFC os melhores resultados foram utilizando as características (1) e (2) combinadas.

## 6. 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

| | RFC | | | | | |
|---|---|---|---|---|---|---|
| | P | | R | | F1 | |
| classes | sexista | não sexista | sexista | não sexista | sexista | não sexista |
| TF de unigramas sexistas e não sexistas | 0.98126 | 0.86450 | 0.85964 | 0.98196 | 0.91637 | 0.91945 |
| TF de unigramas sexistas | 0.97370 | 0.82049 | 0.80536 | 0.97608 | 0.88150 | 0.89151 |
| TF de unigramas não sexistas | 0.94166 | 0.77868 | 0.75429 | 0.94863 | 0.83754 | 0.85525 |
| Caracteres | 0.55058 | 0.57762 | 0.82857 | 0.25725 | 0.66151 | 0.35561 |
| Palavras | 0.56951 | 0.57801 | 0.74821 | 0.37882 | 0.64670 | 0.45753 |
| Caractres + Palavras | 0.62207 | 0.74657 | 0.87071 | 0.41882 | 0.72564 | 0.53638 |
| Likes | 0.56790 | 0.56256 | 0.71357 | 0.40392 | 0.63239 | 0.47003 |
| Dislikes | 0.58662 | 0.55767 | 0.62929 | 0.51294 | 0.60705 | 0.53415 |
| Likes + Dislikes | 0.57617 | 0.58174 | 0.73214 | 0.40863 | 0.64480 | 0.47987 |
| Likes + Dislikes + Caracteres + Palavras | 0.59760 | 0.60715 | 0.72714 | 0.46235 | 0.65595 | 0.52471 |
| TF de bigramas sexistas e não sexistas | 0.58630 | 0.90654 | 0.97750 | 0.24235 | 0.73294 | 0.38201 |
| TF de bigramas sexistas | 0.55571 | 0.83842 | 0.97464 | 0.14431 | 0.70783 | 0.70783 |
| TF de bigramas não sexistas | 0.57314 | 0.85786 | 0.96893 | 0.20745 | 0.72023 | 0.33385 |

**Figura 11. Precisão, revoção e F1 para o RFC usando as features selecionadas.**

e pronomes femininos como palavras mais relevantes do discurso sexista. Além disso, identificamos que mulheres acusam mais falsos positivos do que homens, e que homens identificam mais falsos negativos para o discurso sexista, esse fato trás o questionamento sobre qual seria a forma mais balanceada de distribuir os rótulos atribuídos para cada comentário sem enviesar a base para nenhuma das classes.

Alguns dos comentários coletados caiam no conceito de sexismo benevolente, como o comentário “Ela é muito linda. Merece ser paquerada e amada. Em todos os sentidos” que pode causar desconforto em mulheres que o escutam, mas homens comumente não veem sexismo nele. Esse fato torna relevante a distinção entre sexismo hostil e sexismo benevolente para classificações futuras.

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 100 palavras mais frequentes no discurso sexista e das 100 palavras mais frequentes no discurso não sexista como características textuais da base ou considerando a quantidade de *likes*, *dislikes*, palavras e caracteres em cada comentário como características quantitativas.

O uso das características quantitativas é mais barato do que o uso das características textuais, considerando que para obter as características textuais é necessário processar todo o texto da base de dados. Obtivemos precisão de aproximadamente 0,99 na classe sexista para o SVM usando características textuais e para o KNN usando características quantitativas. Para o SVM usando características quantitativas e para o RFC usando características textuais, obtivemos aproximadamente 0,98 de precisão na classe sexista. Isso mostra que podemos obter resultados eficientes em determinados classificadores com um processamento reduzido.

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 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

Badjatiya, P., Gupta, S., and Gupta, M. (2017). Deep learning for hate speech detection in tweets. pages 759–760.

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers Technology*, pages 233–239.

Davidson, T., Warmsley, D., and Macy, M. (2017). Automated hate speech detection and the problem of offensive language. *Eleventh International AAAI Conference on Web and Social Media*.

Fuglede, B. and Topsoe, F. (2004). Jensen-shannon divergence and hilbert space embedding. page 31.

Glick, P. and Fiske, S. T. (2018). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In *Social Cognition*, pages 116–160. Routledge.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Marques, J. J. and dos Santos, J. L. (2018). Mapa da violência contra a mulher.

Oliveira, S. (2018). Adolescente vítima de bullying se suicida por 'não aguentar mais'.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Rossi, M. (2014). Mulher espancada após boatos em rede social morre em guarujá, sp.

Vianna, J. and Hising, E. (2018). Homem é condenado a 41 anos de prisão por crimes como racismo, terrorismo e divulgação de pedofilia na internet.

Von Smigay, K. E. (2002). Sexismo, homofobia e outras expressões correlatas de violência: desafios para a psicologia política. *Psicologia em revista*, 8(11):32–46.