

UNIVERSIDADE FEDERAL DO AMAZONAS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Identificação e análise de discurso sexista em
comentários curtos provenientes de portais de
notícia

Maria Luísa Pereira Braga

MANAUS
2022

MARIA LUÍSA PEREIRA BRAGA

IDENTIFICAÇÃO E ANÁLISE DE DISCURSO
SEXISTA EM COMENTÁRIOS CURTOS
PROVENIENTES DE PORTAIS DE NOTÍCIA

Dissertação apresentada ao Programa de
Pós-Graduação em Informática do Instituto
de Computação da Universidade Federal do
Amazonas como requisito parcial para a ob-
tenção do grau de Mestre em Informática.

ORIENTADOR: EDUARDO FREIRE NAKAMURA

Manaus

Agosto de 2022

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

B813i Braga, Maria Luísa Pereira
 Identificação e análise de discurso sexista em comentários curtos
 provenientes de portais de notícia - Manaus/AM / Maria Luísa Pereira
 Braga. - 2022.
 62 f. : il., color. ; 31 cm.

 Orientador(a): Eduardo Freire Nakamura.
 Dissertação (mestrado) - Universidade Federal do Amazonas, Programa
 de Pós-Graduação em Informática, Manaus, 2022.

 1. sexismo. 2. discurso de ódio. 3. ciência de dados. 4. machine learning.
 I. Nakamura, Eduardo Freire. II. Universidade Federal do Amazonas.
 Programa de Pós-Graduação em Informática. III. Título



PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
INSTITUTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA



UFAM

FOLHA DE APROVAÇÃO

"IDENTIFICAÇÃO E ANÁLISE DE DISCURSO SEXISTA EM COMENTÁRIOS CURTOS PROVENIENTES DE PORTAIS DE NOTÍCIA"

MARIA LUISA PEREIRA BRAGA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Freire Nakamura - PRESIDENTE

DocuSigned by:

5ED7C34E819C4FB...

Profa. Tanara Lauschner - MEMBRO EXTERNO

Profa. Jonice Oliveira - MEMBRO EXTERNO

Manaus, 11 de Agosto de 2022

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 aceitação 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 técnicas de aprendizagem de máquina para classificação de comentários sexistas/não sexistas. Os resultados preliminares mostram que, ao utilizar classificadores automáticos, é possível identificar comentários sexistas com uma medida F1 e precisão acima de 0,9 e revocação próxima a 0,8.

Palavras-chave: Sexismo, Discurso de ódio, Ciência de dados.

Abstract

Identifying the spread of hate and prejudice speech is an ongoing challenge for online social media environments. Accurate characterization and identification are key to treating and mitigating violence, as well as educating users effectively and assertively. The online dissemination 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 objective has grown as the female figure overcomes the barriers of gender inequality. In particular, sexist speech propagates and encourages derogatory and abusive behavior against women. In this work, we build a corpus of sexist speech in Portuguese collected from popularly accepted news portals, such as G1 and UOL, for example. The work presents the three main contributions: (1) the process of creating the corpus and rotating comments (sexist/non-sexist); (2) a characterization and analysis of the corpus and behavior of anonymous labelers; (3) an assessment of machine learning techniques for detecting sexist/non-sexist comments. Preliminary results show that, by using automatic classifiers, it is possible to identify sexist comments with an F1 measure and precision above 0.9 and recall close to 0.8.

Keywords: Sexism, Hate speech, Data science.

Lista de Figuras

4.1	Esquema representando as etapas de desenvolvimento do trabalho.	19
4.2	Fluxo de classificação final dos comentários a partir dos votos atribuídos. .	21
5.1	Gráfico de barras da distribuição de idades dos usuários cadastrados plataforma de votação separados por gênero.	25
5.2	Gráfico de barras da distribuição de quantidade de votos pela idade e gênero dos usuários cadastrados plataforma de votação.	27
5.3	Gráfico de barras relacionando as porcentagens dos votos corretos para a classe “sexista” separado por gênero e distribuído por idades.	28
5.4	Gráfico de barras relacionando as porcentagens dos votos incorretos para a classe “sexista” separado por gênero e distribuído por idades.	28
5.5	Gráfico de barras relacionando as porcentagens dos votos corretos para a classe “não sexista” separado por gênero e distribuído por idades.	28
5.6	Gráfico de barras relacionando as porcentagens dos votos incorretos para a classe “não sexista” separado por gênero e distribuído por idades.	29
5.7	Gráfico de barras relacionando as porcentagens dos votos femininos corretos para a classe “sexista” separado por gênero e distribuído por idades.	30
5.8	Gráfico de barras relacionando as porcentagens dos votos femininos incorretos para a classe “sexista” separado por gênero e distribuído por idades. .	31
5.9	Gráfico de barras relacionando as porcentagens dos votos femininos corretos para a classe “não sexista” separado por gênero e distribuído por idades. . .	31
5.10	Gráfico de barras relacionando as porcentagens dos votos femininos incorretos para a classe “não sexista” separado por gênero e distribuído por idades.	31
5.11	Distribuição de votos femininos e masculinos para o o conjunto de comentários com pelo menos dois votos de cada gênero.	32
5.12	Prevalência de votos femininos e masculinos para o o conjunto de comentários com pelo menos dois votos de cada gênero.	33
5.13	Nuvem de palavras de comentários sexistas.	35

5.14	Nuvem de palavras de comentários não sexistas.	35
5.15	Gráficos de pizza representando as quantidades de <i>likes</i> e <i>dislikes</i> para cada classe de comentários.	40

Lista de Tabelas

2.1	Associação de métricas nominais aos valores de Kappa.	12
3.1	Relação dos resultados principais apresentados pelos trabalhos relacionados.	18
5.1	Número de comentários que recebeu cada quantidade de votos de 1 até 9 votos e Concordância de Fleiss's Kappa calculada pelo número de votos que cada comentário recebeu	27
5.2	Número de comentários que recebeu cada quantidade de votos femininos de 1 até 7 votos e Concordância de Fleiss's Kappa calculada pelo número de votos femininos que cada comentário recebeu	30
5.3	Concordância geral, feminina e masculina para comentários com quatro votos distribuídos igualmente entre esses gêneros.	32
5.4	Relação dos quinze unigramas com maior $F_s - F_n$ positivo.	37
5.5	Relação dos quinze unigramas com menor $F_s - F_n$ negativo.	37
5.6	Relação dos quinze bigramas com maior $F_s - F_n$ positivo	38
5.7	Relação dos quinze bigramas com menor $F_s - F_n$ negativo	38
5.8	Relação das estatísticas relacionadas às características quantitativas do dis- curso sexista.	41
5.9	Relação das estatísticas relacionadas às características quantitativas do dis- curso não sexista.	41
5.10	Resultados obtidos na aplicação do modelo BERT e dos modelos SVM e LR implementados como linhas de base.	42
5.11	Exibe valores de p-values encontrados na execução do teste estatístico de Wilcoxon para os modelos descritos no trabalho.	44
5.12	Relação dos resultados obtidos para o SVM.	45
5.13	Relação dos resultados obtidos para o KNN.	45
5.14	Relação dos resultados obtidos para o RF.	46

Sumário

Resumo	iii
Abstract	iv
Lista de Figuras	v
Lista de Tabelas	vii
1 Introdução	1
1.1 Problema	3
1.2 Motivação	4
1.3 Justificativa	4
1.4 Objetivos	5
1.5 Abordagem proposta	5
1.6 Organização da dissertação	5
2 Fundamentação teórica	7
2.1 Discurso de ódio	7
2.2 Sexismo	8
2.3 Representação dos dados	9
2.3.1 TF-IDF	9
2.3.2 Bag of words	10
2.3.3 POS tagging	10
2.3.4 Índices de legibilidade	11
2.4 Índices de concordância	11
2.5 Algoritmos de Classificação	12
2.5.1 Regressão Logística	13
2.5.2 Support Vector Machine	13
2.5.3 K-Nearest Neighbors	13

2.5.4	Random Forest	14
2.5.5	Bidirectional Encoder Representations from Transformers	14
2.6	Considerações do capítulo	14
3	Trabalhos relacionados	15
3.1	Análise de sentimentos	15
3.2	Detecção de discurso de ódio	16
3.3	Criação de bases de dados em português	17
3.4	Considerações finais	18
4	Solução proposta	19
4.1	Criação da base de dados	19
4.1.1	Coleta de comentários	20
4.1.2	Classificação manual dos comentários	20
4.1.3	Caracterização do dataset	22
4.2	Classificação automática	22
5	Resultados	25
5.1	Análise dos votos atribuídos	25
5.2	Análise dos votos femininos	29
5.3	Análise de concordâncias	31
5.4	Análise dos comentários	34
5.4.1	Análise de n-gramas	34
5.4.2	Análise de quantitativa	40
5.5	Classificação automática	41
6	Considerações finais	47
6.1	Conclusões	47
6.2	Contribuições	48
6.3	Próximos passos	49
	Referências Bibliográficas	50

Capítulo 1

Introdução

O sexismo consiste em atos e discursos que ofendem, agridem ou diminuem as pessoas de um gênero [36, 15]. Direcionar o sexismo à mulheres é muito comum na Internet e normaliza indiretamente a prática de discriminação ou violência contra a mulher. Por esse motivo, é importante mitigar a difusão do sexismo online.

Apesar de ser um meio de distribuir informações, a Internet também é utilizada como propagador de opiniões e comportamentos. Atualmente é comum que pessoas adotem e repliquem pontos de vista que leram ou ouviram em fontes digitais, mesmo que esses pontos de vista sejam nocivos para outras pessoas. A cada dia que passa, o mundo virtual está mais integrado com o mundo real, de forma que acontecimentos digitais têm grande impacto no comportamento e nas crenças do ser humano.

Em julho de 2020, um influenciador conhecido no meio dos jogos digitais cometeu suicídio após receber mensagens negativas no Twitter [25]. O homem lidava com depressão há dezesseis anos e publicou um tuíte pedindo sua namorada em casamento, no entanto seus seguidores responderam de forma ofensiva ao pedido de casamento pela Internet. Os comentários negativos foram tantos que o homem, de 31 anos, decidiu tirar a própria vida.

Outro caso em que publicações na Internet geraram consequências na vida real ocorreu em maio de 2014 quando uma mulher foi espancada até a morte por conta de um boato publicado no Facebook que acusava a mulher de praticar bruxaria [33]. 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, referentes ao seu peso e aparência [27].

Um caso mais severo aconteceu em 2018, quando a justiça do Paraná condenou a 21 anos de prisão o líder de uma página da Internet utilizada para organizar crimes [35]. O homem foi acusado de ter utilizado a Internet para divulgar imagens contendo

pedofilia e racismo, liderar uma associação criminosa virtual e incentivar o cometimento de crimes ainda mais graves por parte de terceiros, como homicídio, feminicídio e terrorismo.

É comum que opiniões nocivas voltados para grupos sociais minoritários sejam propagadas pelas redes sociais. Em 2016 a ONG Comunica que Muda monitorou durante três meses dez tipos diferentes de tópicos voltados para a intolerância nas redes sociais, os resultados foram publicados no Dossiê da Intolerância [23] e indicaram que dentre as 542.781 publicações analisadas, mais de 84% dos textos sobre intolerância consistiam em abordagens negativas. Entre os tópicos analisados estavam o racismo, a política, a homofobia e a misoginia. Uma análise mais profunda das discussões sobre misoginia mostrou que 88% das publicações analisadas expressavam intolerância contra a mulher, enquanto apenas 22% eram contra a misoginia.

O Instituto Avon em conjunto com a Decode realizaram um estudo referente ao impacto da violência virtual no mundo real [9]. Na pesquisa em questão os dados coletados contemplavam o intervalo de janeiro de 2019 até março de 2020, incluindo conteúdos produzidos e compartilhados espontaneamente por usuários da Internet. A pesquisa apresenta a estimativa de que 95% de todos os comportamentos agressivos e difamadores na Internet tem mulheres como alvo.

No período de pesquisa, foram obtidos 152 mil relatos de violência contra a meninas e mulheres, incluindo situações de vazamento não autorizado de imagens íntimas, perseguição, ameaças e assédio. As principais consequências da violência indicadas na pesquisa foram isolamento social (digital e fisicamente), problemas psicológicos graves e pensamentos suicidas.

Casos como estes, mostram que a Internet precisa ter monitoramento efetivo, para evitar que crimes e outras consequências graves ocorram a partir de conteúdos publicados online.

Historicamente, a figura feminina foi vista como frágil e impotente, e esse fator prejudica a mulher até hoje em seus ambientes pessoais e profissionais. O movimento feminista luta para que esse tipo de tratativa seja extinta e mulheres sejam vistas como iguais aos homens. Neste contexto, a Internet é um dos meios utilizados para que o movimento feminista tenha mais visibilidade e alcance.

No entanto, quando publicações sobre feminismo são feitas em sites, portais de notícias e demais mídias sociais, frequentemente aparecem comentários que diminuem, ofendem, objetificam e ridicularizam as mulheres por acreditarem no movimento feminista. Esse fato atrapalha a difusão da igualdade entre gêneros, uma vez que desmoraliza o movimento e perpetua a aversão ao feminismo, estabelecendo a normalização da cultura machista [26].

Ao longo de 2020, foram identificados mais de 105 mil casos de violência contra a mulher [12] e a ausência do combate à fragilização da figura feminina, a desmoralização do movimento feminista e a cultura patriarcal estabelecida na nossa sociedade são fatores que contribuem para isso.

Identificar e combater o discurso ofensivo na Internet são formas de evitar a propagação de comportamentos discriminatórios e violentos. Segundo Banks [2] a falta da preocupação em punir os autores de discurso de ódio agrava a propagação desse tipo de pensamento na Internet.

Embora exista uma dificuldade humana de identificar e classificar o conteúdo do grande volume de opiniões publicadas na Internet diariamente, é possível mitigar a propagação do discurso ofensivo, em particular do discurso sexista voltado para a mulher, 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. [6] e Kowk&Wang [19].

1.1 Problema

O problema de detecção de discurso prejudicial na Internet é apresentando em outros trabalhos da literatura [6, 19, 29, 28] que destacam categorias distintas desse tipo de discurso, como o racismo, homofobia e o sexismo. O fato de existirem categorias distintas de discurso de ódio trás a necessidade de reconhecermos as características de cada uma dessas categorias, a fim de trazer uma classificação automática mais precisa.

Outro fator relevante para possibilitar uma classificação precisa, é o uso de uma base de dados representativa do discurso analisado. Os trabalhos conhecidos que tratam a detecção automática de discursos sexista online não trabalham com a língua portuguesa. Sendo assim, para estudar e entender categorias de discurso de ódio, se faz necessária a montagem de uma base de dados que diferencie textos pertencentes a cada tipo de discurso.

Para este trabalho, focamos no discurso sexista, que é uma das categorias de discurso de ódio em ascensão na Internet e também é muito prejudicial para suas vítimas, como mostra o estudo feito pelo Instituto Avon [9]. Diante disso, aqui temos dois objetivos principais: (i) disponibilizar uma base de dados representativa do discurso sexista em português e (ii) definir as melhores combinações de características do texto com métodos de aprendizagem para classificar o discurso sexista automaticamente.

1.2 Motivação

O movimento feminista existe a fim de atingirmos a igualdade entre os gêneros, este movimento foi responsável pela atribuição da cidadania a pessoas do gênero feminino de diversos outros atos visando a melhoria das condições sociais da mulher [22].

A difusão do feminismo incentiva que mulheres busquem seus direitos quando necessário e entendam que devem ser tratadas com igualdade, no entanto ainda existem muitas pessoas que não entendem as motivações desse tipo de movimento e ridicularizam a causa, o que impede que o tópico de igualdade de gênero alcance determinados públicos. Comentários em publicações de informações sobre causas feministas em portais de notícias são grandes exemplos do comportamento discriminatório e ofensivo em relação as mulheres [21, 16].

Os portais de notícia digitais estão entre as principais fontes de informação utilizadas e a possibilidade de leitores emitirem suas opiniões em comentários nessas plataformas tornou-se comum. Mitigar a presença de comentários nocivos nesses meios de comunicação é uma tarefa difícil para o ser humano, uma vez que há um grande volume de comentários publicados em um curto espaço de tempo. Portanto, a classificação automática de comentários como parte de um grupo de discurso de ódio contribui para evitar a difusão desse tipo de pensamento.

Dentre as categorias de discurso intolerante na Internet, um dos tópicos mais presentes em 2016 era a misoginia [23], de forma que quando se fala sobre misoginia na Internet, a maioria das pessoas compactua com comentários misóginos. O crescimento desse tipo de discurso online contribui para consequências no mundo real, uma vez que incentiva a manutenção de comportamentos machistas que podem gerar danos a saúde física e mental da mulher [26]. Por esses motivos, o discurso sexista contra a mulher precisa ser mitigado no ambiente virtual.

1.3 Justificativa

A detecção automática de discurso de ódio na Internet vem sendo abordada na literatura. Os trabalhos conhecidos utilizam técnicas de aprendizagem para detectar diversos tipos de discurso de ódio em diferentes plataformas [28, 8], como a detecção de racismo em tweets [19] ou a detecção de misoginia em bases de dados estruturadas [30]. Além disso, alguns trabalhos exploram também a diferenciação entre o discurso ofensivo e o discurso de ódio [6]. Existem também trabalhos que envolvem técnicas de diferenciação de sentimentos em texto [31, 29, 5].

Os trabalhos da literatura mencionam a extração de *features* relevantes para cada discurso como um problema a ser tratado para otimizar as classificações automáticas. Desta maneira, temos como objetivo desenvolver uma base de dados representativa de comentários sexistas em portais de notícias para realizar a detecção automática desse tipo de discurso em comentários na Internet, explorando diferentes formas de detectar *features* relevantes para a diferenciação de comentários sexistas ou não sexistas a partir da análise da base de dados.

1.4 Objetivos

O objetivo geral deste trabalho é construir uma base de dados em português, composta por comentários coletados em portais de notícias, a fim de caracterizar a diferença entre discurso sexista e não sexista quando apresentados em um mesmo contexto. Para atingir o objetivo geral definimos os seguintes objetivos específicos:

- Identificar, em comentários, as principais características que diferenciam o discurso sexista do discurso não sexista;
- Demonstrar a validade da base de dados criada utilizando diferentes algoritmos de classificação automática para detectar o discurso sexista.

1.5 Abordagem proposta

Neste trabalho pretendemos construir uma base de comentários e validá-la utilizando técnicas de aprendizado de máquina, sendo assim as etapas que seguimos são:

1. Classificação manual de comentários, realizada por rotuladores de diferentes gêneros e idades;
2. Análise dos comentários rotulados e dos rótulos atribuídos para estes comentários, a fim de extrair as características relevantes para cada classe de comentários;
3. Aplicação de classificadores automáticos utilizando diferentes características dos comentários e avaliação dos resultados obtidos para cada classificador.

1.6 Organização da dissertação

O trabalho foi organizado da seguinte forma: o capítulo 2 apresenta os conceitos relevantes para o desenvolvimento do trabalho. O capítulo 3 apresenta os trabalhos da

literatura relacionados com o problema tratado. O capítulo 4 descreve a abordagem proposta para o atingimento dos objetivos deste trabalho. O capítulo 5 descreve os resultados obtidos. Por fim, o capítulo 6 apresenta as conclusões e trabalhos futuros.

Capítulo 2

Fundamentação teórica

Neste capítulo descrevemos os principais conceitos necessários para o entendimento do trabalho apresentado. Definimos o conceito de discurso de ódio e suas variações, dando foco a definição do sexismo e conceitos relacionados. Descrevemos as características do discurso que desejamos identificar, formas de representar o texto e os classificadores que utilizamos para realizar a predição da classe de um comentário.

2.1 Discurso de ódio

O discurso de ódio pode ser definido como o discurso prejudicial dirigido a pessoas ou grupos sociais, com a intenção de insultar, ameaçar ou ofender [34, 6, 28]. O discurso de ódio pode ser categorizado a partir do seu alvo, sendo assim existem diversas categorias distintas de discurso de ódio, entre elas as principais são racismo, homofobia, intolerância religiosa, intolerância política, sexismo e discriminação direcionada a aparências e tipos físicos.

Dentro do discurso de ódio, temos aqueles que utilizam de ofensas e palavras de injúria diretamente e os que utilizam sarcasmos e piadas que ofendem, sem necessariamente utilizar palavras de ódio. Além disso, existem também os casos em que frases não demonstram ódio, ofensas ou piadas, e mesmo assim geram ofensa em determinados grupos sociais.

Um exemplo disso é o caso do ator Julianos Cazarré, que em 2019 foi acusado de machismo após realizar duas publicações no seu Instagram [24]. A primeira publicação continha a legenda “a masculinidade faz do mundo um lugar mais seguro”, essa frase foi considerada machista por muitos dos seus seguidores. Na segunda publicação, Cazarré afirmou que não teve intenção de disseminar machismo anteriormente e usou a frase “homens: cuidem de suas famílias, sejam pais legais e presentes” o que também foi

considerado como machista, pois para alguns dos seguidores do ator a frase apresentava a crença de que mulheres não seriam capazes de cuidar de suas famílias.

Independente da intenção do autor ao realizar as publicações, o discurso foi prejudicial para alguns de seus seguidores e isso é um fenômeno comum na Internet atualmente.

2.2 Sexismo

O foco deste trabalho é a detecção do discurso prejudicial para mulheres, que denominamos de sexismo voltado para o gênero feminino. Sendo assim, incluímos na definição de sexismo, para este trabalho, atos de machismo, misoginia e feminicídio, além de atos e falas que impactam negativamente na construção da imagem feminina, seja ela cis ou trans.

O machismo tem como princípio a supremacia do gênero masculino sobre o feminino, o que leva as pessoas a construírem discursos baseados na ideia de que homens podem fazer coisas que mulheres não seriam capazes ou não deveriam fazer. Pensamentos machistas contribuem também para a misoginia, que é definida como a aversão e ódio às mulheres. Falas e comportamentos machistas não são sempre necessariamente misóginos, um exemplo disso é a atribuição de tarefas e atividades ao gênero feminino, como a frase “lugar de mulher é na cozinha” que é machista mas não é misógina.

Além da misoginia, existe também a misandria, definida como repulsa e ódio ao gênero masculino. Discursos misândricos estão incluídos no sexismo voltado para o gênero masculino, mas não daremos foco para esse tópico neste trabalho uma vez que a misoginia é apresentada em maior volume na Internet atualmente [23].

Tanto o machismo quanto a misoginia contribuem a difusão de pensamentos que dão base para crimes contra a mulher, como a violência doméstica, assédios, estupros e feminicídios, que consistem no assassinato de uma pessoa motivado pelo fato de ela ser mulher.

Segundo Glick&Fisk [15] 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 indesejado dirigidos às mulheres, mesmo que não gerem ofensa.

Muitas vezes o sexismo benevolente não é percebido pelo interlocutor ou até mesmo pelas mulheres que são alvo desse tipo de discurso. Exemplos de sexismo benevolente são comentários como “minha mulher é perfeita, não tem uma celulite”, “o homem deve proteger a sua mulher” ou “mulheres devem saber cozinhar”, essas frases

podem parecer inofensivas e até positivas para algumas mulheres, mas são frases que reforçam a ideia do gênero feminino como o sexo frágil, com capacidades limitadas em relação ao homem.

Smigay&Ellen [36] definem dois conceitos para sexismo. O primeiro deles é retirado do Novo Dicionário Aurélio na edição de 1975, que diz que sexismo é uma posição misógina, de desprezo frente ao sexo oposto. O segundo é que sexismo é atitude de discriminação em relação às mulheres, que pode se manifestar tanto por homens quanto por mulheres.

Nesse segundo conceito, os autores assumem que o sexismo inclui 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 que ofendem, diminuem, oprimem, ou agredem pessoas que se incluem em um determinado gênero.

Este conceito inclui o sexismo direcionado a homens, mulheres e pessoas não-binárias, mas neste trabalho focamos no sexismo direcionado ao gênero feminino, por conta da abrangência da misoginia e do machismo em redes sociais, que torna mais importante a tarefa de encontrar comentários de ódio ou ofensivos direcionados à mulheres.

2.3 Representação dos dados

A fim de determinar automaticamente se um comentário é sexista ou não, precisamos treinar um classificador. Na literatura, trabalhos envolvendo classificação de textos por meio de métodos supervisionados, utilizam três principais formas de representação dos dados: Term Frequency - Inverse Document Frequency(TF-IDF), part of speech tagging (POS *tagging*) e bag of words. Essas representações levam em consideração a estrutura do texto sem considerar a semântica do conteúdo, para suprir essa necessidade utilizamos cálculos de entropia e divergências. Existem também os índices de legibilidade, que indicam a complexidade do texto analisado.

2.3.1 TF-IDF

O TF-IDF é uma medida estatística com o objetivo de quantificar a importância de um termo em um texto. Dada uma coleção D de documentos, seja $w_{i,j}$ o termo associado

ao par (k_i, d_j) , $d_j \in D$ e $k_i \in d_j$, definimos a ponderação TF-IDF como

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & , \quad \text{se } f_{i,j} > 0 \\ 0 & , \quad \text{caso contrário} \end{cases}$$

onde N é o número de documentos da coleção, $f_{i,j}$ representa a frequência de um termo k_i em um documento d_j e n_i é o número de documentos onde k_i ocorre.

Além do TF-IDF, podemos representar o texto utilizando apenas o *Term Frequency* (TF) que representa a frequência de cada termo em um determinado documento.

2.3.2 Bag of words

O uso de n-gramas na classificação de texto é interessante quando precisamos usar palavras-chave para determinar a classificação de uma sentença. Em máquinas de busca, é possível usar n-gramas com $n=1$ (unigramas) para realizar a indexação de termos relevantes. O uso de n-gramas com $n=2$ ou $n=3$ (bigramas e trigramas, respectivamente) é útil na análise de sentimentos em um texto, uma vez que dá contexto a uma parte da sentença [8].

Para exemplificar o uso de n-gramas na classificação de texto, vamos analisar a frase “uma camiseta vagabunda”. O termo “vagabunda” é negativo e pode ser ofensivo quando usado em contextos diferentes deste. Nesta sentença, o termo se refere a camiseta.

Analizando a sentença em unigramas, cada termo será considerado individualmente, e a frase poderia ser considerada ofensiva por conter a palavra “vagabunda”. Usando bigramas, seria possível identificar a combinação “camiseta vagabunda” e concluir que, uma vez que o termo negativo se refere a camiseta, a sentença não pode ser considerada ofensiva.

2.3.3 POS tagging

A representação de dados utilizando POS *tagging* consiste em atribuir *tags* a cada parte do texto, essas *tags* representam entidades sintáticas na estrutura do texto, como verbos, substantivos e adjetivos. Usando POS *tagging* na frase “as mulheres são bonitas”, teríamos a frase representada como “artigo”, “substantivo”, “verbo” e “adjetivo”.

Utilizar POS *tagging* como representação do texto para classificadores automáticos permite que conteúdos com estruturas sintáticas similares sejam associados entre si, mas não leva em consideração a semântica do texto, o que pode ser problemático em

textos contendo sarcasmos ou ironias e por isso precisa de complementos que ajudem a extrair outras características do texto.

2.3.4 Índices de legibilidade

Os índices de legibilidade são scores calculados a fim de medir a facilidade que um leitor teria para entender determinado conteúdo, essa métrica se baseia nos tamanhos das sentenças e das palavras utilizadas no texto [11].

Dentre os índices de legibilidade temos o índice de Flesch [7], que estima a legibilidade baseado na relação entre o total de frases T_f e o total de palavras T_p , seguindo a fórmula:

$$Flesch = 206.835 - \left[1.002 \times \frac{T_p}{T_f} \right] - \left[84.6 \times \frac{T_p}{T_f} \right].$$

Para os textos brasileiros, este índice foi adaptado [20] para incluir também o total de sílabas T_s , atribuindo pesos diferentes para cada relação, utilizando a fórmula:

$$FleschBr = 248.835 - \left[1.015 \times \frac{T_p}{T_f} \right] - \left[84.6 \times \frac{T_s}{T_p} \right].$$

Outro índice de legibilidade é o de de Dale-Chall [14], que busca considerar o léxico dos textos utilizando a porcentagem de palavras difíceis P_{df} presentes no texto juntamente com o tamanho médio das frases T_{mf} , seguindo a fórmula:

$$DaleChall = [0.1579 \times P_{df}] + [0.0496 \times T_{mf}] + 3.6365.$$

Esse índice considera que as palavras difíceis do texto foram definidas previamente.

Para a detecção de comentários sexistas, utilizar esses índices trás informações sobre o quão fácil cada comentário pode ser compreendido e devidamente rotulado.

2.4 Índices de concordância

A rotulagem manual de bases de dados consiste em diversas pessoas atribuindo opiniões sobre cada texto analisado. Para definir os rótulos finais destes textos de forma assertiva, precisamos garantir que o grau de concordância entre os rótulos atribuídos é alto o suficiente para considerarmos a classificação como válida. Para isso, podemos calcular o coeficiente de concordância entre os rótulos atribuídos para cada texto.

Para variáveis numéricas, existem os índices de concordância de Pearson e Spearman, enquanto que para variáveis nominais o índice de concordância mais utilizado é

o Kappa. Todos os índices levam em consideração os rótulos atribuídos por dois juízes diferentes sendo comparados. Para a avaliação de múltiplos rotuladores e não apenas dois, utilizamos a variação de Kappa chamada Fleiss's Kappa [13] que é definida pela fórmula:

$$kappa = \frac{P_o - P_e}{1 - P_e},$$

onde P_o representa a proporção de concordância definida por $P_o = \sum_{i=1}^k p_{ii}$ e P_e representa a proporção esperada de k vezes que os avaliadores concordam, definida por $P_e = \sum_{i=1}^k p_i + p_{+i}$. Para essas fórmulas, temos $p_{i+} = \sum_j p_{ij}$, $p_{+i} = \sum_i p_{ij}$ e $p_{ij} = \frac{n_{ij}}{N}$, onde os valores de p representam a concordância entre dois rótulos atribuídos para um mesmo objeto posicionados em uma matriz e N corresponde ao número total de amostras dos rótulos, enquanto n_{ij} corresponde ao valor das amostras na linha i e na coluna j .

Para verificar o grau de concordância entre os rótulos atribuídos de forma nominal, assumimos os valores da tabela 2.1.

Valores de Kappa	Interpretação
menor que 0	Ausência de concordância
de 0 até 0.19	Concordância pobre
de 0.20 até 0.39	Concordância leve
de 0.40 até 0.59	Concordância moderada
de 0.60 até 0.79	Concordância substantiva
de 0.80 até 1	Concordância quase perfeita

Tabela 2.1. Associação de métricas nominais aos valores de Kappa.

2.5 Algoritmos de Classificação

Na aprendizagem de máquina, os algoritmos supervisionados de classificação são aqueles que precisam ser treinados para classificar automaticamente elementos desconhecidos. Aqui, a definição desse tipo de algoritmos será baseada nas descrições de Baeza-Yates [1] a respeito de classificadores binários.

Dada uma coleção D e uma subcoleção $D_t \subset D$ de documentos, uma função de treinamento $T : D_t \times C \rightarrow 0, 1$, onde C representa um conjunto de características de D_t , atribui um valor de 0 ou 1 a cada par $[d_j, c_p]$, $d_j \in D_t$ e $c_p \in C$. A função T é responsável por treinar o algoritmo de classificação. A função T é usada para ajudar o classificador, melhorando a função de classificação $F : D \times C \rightarrow 0, 1$.

Para avaliar a eficácia do classificador que utiliza aprendizagem supervisionada, usamos uma subcoleção $D_a \subset D$, $D_a \cap D_t = \emptyset$. Ao submeter os documentos de D_a à

função F de forma que $F(D_a, C)$ será equivalente aos documentos de D_a com classes atribuídas por F . O F1 de F é calculada a partir da fórmula

$$F1(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{p(j)}},$$

onde $p(j)$ é a revocação da j -ésima classificação e $r(j)$ é a precisão na j -ésima classificação realizada.

Neste trabalho, os algoritmos de classificação utilizados foram a Regressão Logística (RL), o Support Vector Machine (SVM), o K-Nearest Neighbors (KNN), o Random Forest (RF) e o Bidirectional Encoder Representations from Transformers (BERT).

2.5.1 Regressão Logística

A Regressão Logística [38] é uma técnica estatística utilizada para realizar a predição automática de um evento baseado na probabilidade calculada para a sua ocorrência, considerando determinadas variáveis que impactam nessa probabilidade. Neste trabalho, consideramos a Regressão Logística binária, ou seja, a variável dependente calculada pela regressão pode pertencer apenas a duas classes diferentes.

2.5.2 Support Vector Machine

As Máquinas de Suporte de Vetores (SVM's) utilizam métodos estatísticos de regressão para realizar o reconhecimento de padrões em dados [17]. Nesse classificador, os dados de entrada são representados como pontos no espaço, distribuídos de forma que cada espaço representa um grupo de características dos dados. Dizemos que o SVM divide as representações dos dados utilizando um hiperplano entre os dados de duas classes diferentes. Assim, o SVM gera um modelo que ao receber novos dados, plotará a representação desses dados no espaço que corresponde às características de novos dados de entrada.

O SVM é amplamente utilizado na detecção de discurso de ódio, e normalmente apresenta os melhores resultados em comparação com outros métodos similares.

2.5.3 K-Nearest Neighbors

O KNN [18] é um algoritmo onde os dados são agrupados por vizinhanças similares a fim de atribuir uma classificação a eles. Neste algoritmo temos uma variável K que ditará a quantidade de vizinhanças geradas pelo algoritmo, essas vizinhanças são formadas

com base em padrões identificados com características similares entre os participantes de cada grupo.

2.5.4 Random Forest

No RF [4] são criadas várias árvores de decisão, que são construídas de forma similar a fluxogramas, considerando as variáveis que caracterizam um dado a ser classificado. Neste algoritmo, as variáveis a serem classificadas são escolhidas de modo aleatório com base em uma primeira variável, a partir da qual serão montadas árvores de decisão repetidas vezes até que se chegue em uma classificação final para o dado de entrada.

2.5.5 Bidirectional Encoder Representations from Transformers

O modelo BERT [10] procura trazer para a máquina uma compreensão de expressões da linguagem humana, com base em um modelo de processamento de linguagem natural chamado Transformer. Ele é um modelo pre treinado, isso indica que antes de executar a aprendizagem sobre uma base de dados, ele deve ser treinado com um corpus de texto. O BERT analisa o contexto ao redor de cada termo, tanto pela esquerda quanto pela direita de uma determinada palavra. Depois de pre treinado, o BERT executa um *fine-tuning* a fim de melhorar sua compreensão do conteúdo.

2.6 Considerações do capítulo

Neste capítulo, apresentamos os fundamentos teóricos utilizados para o desenvolvimento deste trabalho, incluindo o conceito de sexismo, formas de análise e representação dos dados e explicações sobre os algoritmos de classificação utilizados no desenvolvimento do trabalho, além disso incluímos a explicação dos mecanismos utilizados para definição do rótulo final atribuído a cada comentário da base de dados.

Capítulo 3

Trabalhos relacionados

Neste capítulo, descrevemos os trabalhos da literatura que aplicam técnicas relacionadas a detecção automática de discursos em bases de texto. Aqui exploramos a análise de sentimentos, a detecção e a diferenciação do discurso de ódio.

3.1 Análise de sentimentos

O problema de análise de sentimentos consiste, normalmente, em identificar se as opiniões sobre o alvo de um texto são positivas ou negativas. As abordagens conhecidas na literatura envolvem a classificação de tweets [29], análise de reviews de produtos [5] e filmes [31] a fim de determinar se os comentários sobre determinado item são bons ou ruins.

Em seu trabalho, Pak&Paroubek [29] tinham o objetivo de aumentar a acurácia de um classificador na identificação automática de sentimentos em tweets, segregada em sentimentos positivos e negativos a partir da presença de emoticons no texto. Para isso, os autores utilizaram a presença de n-gramas como uma característica binária de cada documento em um classificador de Naïve Bayes, descartando os n-gramas não significativos para sentimentos positivos ou negativos e tomando como métrica a entropia de Shannon, a fim de determinar a distribuição de ocorrência de cada n-grama nos datasets (positivo e negativo), n-gramas com valores menores de entropia foram considerados no treino do classificador. Além disso, os autores também empregaram uma métrica que denominaram de *saliência*, como um indicativo do quão discriminatório entre sentimentos positivos e negativos um termo é. Com o foco na presença ou não dos n-gramas definidos, o contexto de cada tweet analisado não é levado em conta, apenas a sua estrutura.

Dave et. al [5] trabalharam na classificação de reviews de produtos na Internet. Os autores utilizaram um classificador SVM usando n-gramas como features e tiveram 85% de acurácia. Neste problema, os autores identificaram que existem casos em que palavras negativas foram utilizadas em reviews que são positivas ou até mesmo comparam experiências negativas com positivas no mesmo reviews para chegar a uma conclusão, o que dificulta a classificação. Os autores relataram uma dificuldade na extração de features.

Já no trabalho de Pang et al. [31], 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.

3.2 Detecção de discurso de ódio

O problema de detecção de discurso de ódio na literatura inclui a detecção e diferenciação do discurso de ódio do discurso ofensivo [6] e a detecção específica de determinadas classes de discurso de ódio como a misoginia [30] e o racismo [19].

Davidson et al. [6] realizaram utilizaram Regressão Logística (LR) 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.

Kwok&Wang [19] utilizaram um classificador de Naive Bayes com *bag of words* 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.

A pesquisa de Pamungkas et al. [30] executa a detecção de discurso misógino em diversos idiomas a fim de identificar métodos eficientes de classificação em domínios diferentes. Os autores utilizaram o SVM como método de aprendizagem de máquina, o Bidirectional Encoder Representations from Transformers (BERT) e o Long Short Term Memory (LSTM) como métodos de aprendizagem profunda. Em seu trabalho, os autores relatam que os resultados para o SVM aplicados na língua inglesa usando variações de *bag of words* como *features* e outras características quantitativas do texto chegaram a 87% de precisão e 91% de revocação, enquanto os melhores resultados demonstrados pelos autores para os métodos de aprendizagem profunda foram para o modelo BERT na maioria dos testes, demonstrando até 83% de precisão e 87% de revocação para o Italiano e 77% de precisão e 91% de revocação para o Inglês.

3.3 Criação de bases de dados em português

Na literatura, a maioria dos trabalhos busca analisar uma base de dados, no entanto poucos trabalhos detalham o processo de criação de uma base de dados na língua portuguesa baseado em comentários curtos, como é feito nos trabalhos de Pamungkas et al. [30]. Este fato faz com que seja difícil encontrarmos bases de dados prontas na língua portuguesa, que analisem a fundo o discurso sexista.

No entanto, temos trabalhos em português que descrevem e detalham bases de dados de discurso de ódio, como o trabalho de de Pelle & Moreira[8] que exhibe a geração de uma base de comentários em português e a classificação automática para validar esta base. A base foi construída a partir de comentários do G1 visando possibilitar a detecção de discurso de ódio em textos escritos na língua portuguesa.

A base de dados de Pelle & Moreira[8] foi rotulada de forma manual por três rotuladores distintos. A base dispões de 1.250 comentários rotulados e dentro da classe ofensiva os comentários foram rotulados em subcategorias como sexismo, xenofobia, homofobia e racismo. Dos 1.250 comentários rotulados apenas 419 foram considerados ofensivos. Os autores utilizaram uma combinação do Naive Bayes com o SVM, chamada SMO, para testar a possibilidade de classificação automática utilizando essa base de dados e obtiveram um F1 de 82% usando uma seleção de features relacionadas à quantidade de informação no texto.

Neste trabalho, temos o esforço inicial da montagem de uma base de dados, no entanto a quantidade de dados rotulados é baixa para fazer com que os dados sejam representativos da classe ofensiva.

3.4 Considerações finais

Autores	Método	Features	Resultados
Kwok&Wang	Naive Bayes	BoW de unigramas	76% de acurácia
Davidson et al.	Regressão Logística	TF-IDF de n-gramas de até 3 palavras	91% de precisão, 90% de revocação e 90% de F1
Pamungkas et al.	SVM	variações de BoW e características quantitativas do texto	87.16% de precisão, 91.16% de revocação e 89.12% de F1
Pamungkas et al.	BERT		77.31% de precisão, 91.52% de revocação e 83.82% de F1
Pak&Paroubek	SVM	Presença ou não de determinados n-gramas	$F_{0.5}$ de 0.65
Dave et. al	SVM	BoW de n-gramas	85% de acurácia
Peng et al.	SVM	BoW de unigramas	82.9% de acurácia
Pelle & Moreira	SMO	Features relacionadas ao ganho de informação	F1 de 82%

Tabela 3.1. Relação dos resultados principais apresentados pelos trabalhos relacionados.

A tabela 3.1 apresenta resumidamente os principais resultados obtidos pelos trabalhos citados nesta seção, apresentando os métodos de aprendizagem utilizados como classificadores em cada trabalho, as features utilizadas e os resultados obtidos em cada caso.

Na maioria dos trabalhos citados, para os métodos de aprendizagem de máquina o uso de unigramas mostrou resultados superiores ao uso de n-gramas com mais termos. A maioria dos trabalhos relacionados ao tema utiliza bases de dados em inglês e o trabalho que monta uma base de dados em português possui um número baixo de comentários ofensivos a serem considerados, e um número menor ainda de comentários sexistas a serem analisados.

Capítulo 4

Solução proposta

Neste capítulo descrevemos a metodologia proposta para classificar e caracterizar o discurso sexista. A Figura 4.1 ilustra os componentes de nossa solução, que serão descritos em detalhes nas seções seguintes: criação da base de dados (Seção 4.1) e classificação automática (Seção 4.2).

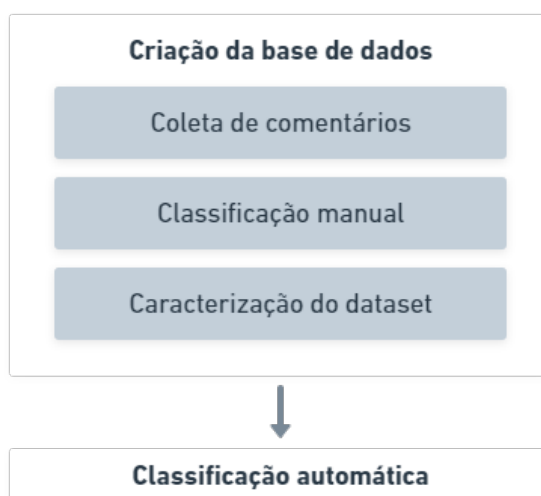


Figura 4.1. Esquema representando as etapas de desenvolvimento do trabalho.

4.1 Criação da base de dados

Este estudo tem início na criação da base de dados que foi iniciada em um trabalho anterior a esse [3] e que se dá em três etapas: a coleta de comentários, a classificação

manual dos comentários e a caracterização dos discursos a fim de detectar características relevantes para a classificação dos comentários. Os detalhes de cada etapa estão descritos nesta seção.

4.1.1 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, pois consideramos que essa quantidade é o suficiente para indicar que uma notícia incitou o engajamento dos leitores. 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 3.588 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 encontrado na 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... Aham que é mimimi. Comparam a vida da mulher há 50 anos atrás com agora”.

4.1.2 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. No site permitimos que pessoas diversas se cadastrassem, preenchendo suas datas de nascimento e gênero para realizar a classificação manual dos comentários. Na primeira versão da plataforma, consideramos apenas os gêneros feminino e masculino, posteriormente incluímos a opção “outro” acompanhada de um campo de texto para que o usuário pudesse nomear o gênero a qual ele pertence.

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 as opções “sexista” e “não sexista” para que o rotulador avalie o comentário. Nos casos em que o comentário responde a algum outro comentário na mesma notícia, exibimos os dois comentários, para que o rotulador tenha o 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 por ele. 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.

Nessa etapa, decidimos coletar rótulos atribuídos por pessoas que não fazem parte do gênero feminino para analisar a diferença entre os rótulos atribuídos por pessoas de gêneros diferentes. Por conta dos resultados das análises realizadas nas seções 5.3 e 5.2 decidimos considerar apenas os votos femininos ao definirmos os rótulos finais do *dataset*.

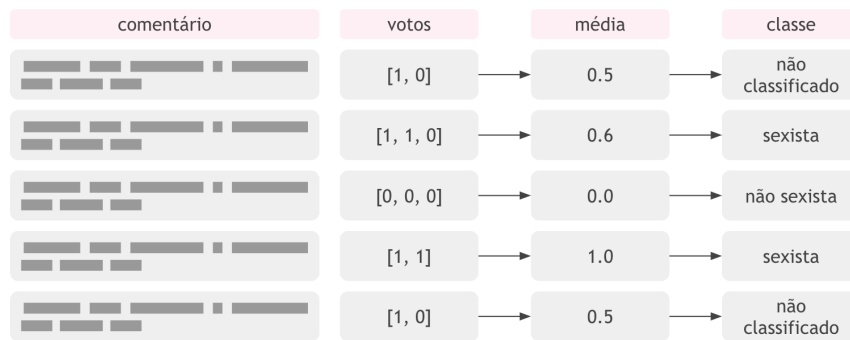


Figura 4.2. Fluxo de classificação final dos comentários a partir dos votos atribuídos.

Cada comentário recebe rótulos de mais de um usuário, o rótulo final dos comentários é atribuído considerando os votos “sexista” com peso 1 e os votos “não sexista” com peso 0, a média dos votos define a qual classe o comentário pertence, como mostra a figura 4.2. Comentários com a média de votos acima de 0,5, foram considerados como sexistas, enquanto 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, pois isso indica que houve um empate entre os rótulos atribuídos.

Esse processo gerou uma base de dados rotulada manualmente com 1.821 comentários sexistas, 1.712 comentários não sexistas e 55 comentários não rotulados.

4.1.3 Caracterização do dataset

A caracterização da base de dados envolve a identificação dos motivos pelo qual cada comentário foi atribuído a uma classe. Então, realizamos uma análise dos votos atribuídos e uma análise dos comentários.

Durante a análise dos votos atribuídos, procuramos entender características dos rotuladores que influenciaram na classificação manual, para isso observamos as relações de gênero, idade e voto atribuído por rotulador. Nesta etapa também analisamos a quantidade de votos atribuídos a cada comentário e a concordância de Fleiss's kappa [13] entre esses votos.

Na análise dos comentários, estudamos o conteúdo dos comentários por meio de características textuais e quantitativas.

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 n -gramas presentes em D , calculamos as frequências normalizadas F_s e F_n para cada n -grama $t_i \in \sigma$, considerando $n = 1$ e $n = 2$.

Consideramos que 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 n -grama em busca das palavras relevantes na diferenciação das duas classes de comentários. A tabela 5.4 mostra os quinze unigramas com maior valor de $F_s - F_n$ que encontramos, sendo os unigramas mais relevantes para o discurso sexista.

Uma vez que o trabalho tem como foco o estudo do discurso sexista direcionado a mulheres, decidimos fazer uma análise separada da base de dados considerando apenas os votos femininos, buscando identificar diferenças entre os resultados encontrados quando consideramos votos de pessoas que não são alvo desse discurso.

4.2 Classificação automática

Utilizamos no nosso dataset métodos de aprendizagem de máquina, a fim de validar a possibilidade da classificação automática na base de dados. Os métodos escolhidos foram: o Support Vector Machine (SVM) com kernel RBF e parâmetro de regularização igual a 1.0 e gamma igual a scale; o K-Nearest Neighbors (KNN) com $N=3$; o Random Forest (RF) com profundidade máxima igual a 8 e 200 estimadores. Para todos os

modelos, utilizamos as implementações disponibilizadas pela biblioteca scikit-learn [32]. As implementações realizadas para o trabalho estão disponíveis em <https://github.com/mlpbraga/sexism-detection-notebooks>.

Executamos cada método escolhido para combinações entre as características do texto listadas:

- *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});
- *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);
- Índices de legibilidade (IL);
- POS tagging (POS).

Uma vez que o foco da pesquisa era o sexismo direcionado ao gênero feminino, optamos por não realizar técnicas de stemming e nem a remoção de stopwords do texto, uma vez que artigos, preposições e variações de gênero das palavras são relevantes para a identificação do alvo dos comentários.

Além dos modelos de aprendizagem de máquina, utilizamos o BERT com um modelo pre treinado em português disponibilizado no *huggingface*¹. A escolha desse modelo se baseia na abordagem adotada por Pamungkas et al.[30], o modelo utilizado começa com a camada BERT, a saída desta camada se conecta a uma camada densa com ativação RELU, passando por uma camada densa com ativação sigmóide. A rede foi treinada com o otimizador Adam e taxa de aprendizagem de 0,0001. Para otimizar os resultados, ajustamos o modelo em até 5 épocas.

Além dos métodos citados, implementamos também a abordagem utilizada no trabalho de Davidson et al., onde foram utilizados o Support Vector Machine (SVM) e a Regressão Logística (LR) como métodos de aprendizagem. O pre-processamento do

¹O modelo pre treinado utilizado neste trabalho está disponível em <https://huggingface.co/neuralmind/bert-base-portuguese-cased>

texto envolveu a transformação de todas as palavras em letras minúsculas, técnicas de stemming, uso de n-gramas de uma até três palavras sendo representados por TF-IDF, técnicas de Part-of-Speech (POS) tag para cada n-grama e características quantitativas relacionando a contagem de números, caracteres especiais, número de palavras, sílabas e caracteres em cada documento do texto.

Para cada um dos algoritmos acima, dividimos a base de dados em treino e teste. Separamos 90% da base para treino e os 10% restantes para teste, repetindo o treino e o teste dez vezes para realizar a validação cruzada. As métricas analisadas para cada modelo foram: precisão para a classe sexista (P_s), precisão para a classe não sexista (P_n), revocação para a classe sexista (R_s), revocação para a classe não sexista (R_n), F1 para a classe sexista ($F1_s$) e F1 para a classe não sexista ($F1_n$).

Capítulo 5

Resultados

Neste capítulo, apresentaremos os resultados obtidos na execução da abordagem proposta. O capítulo segue a estrutura a seguir: na Seção 5.3 descrevemos a análise dos votos atribuídos aos comentários da nossa base; na Seção 5.4 apresentamos uma análise dos comentários após terem seus rótulos atribuídos; por fim, a Seção 5.5 apresentamos a aplicação da classificação automática sob a base de dados.

5.1 Análise dos votos atribuídos

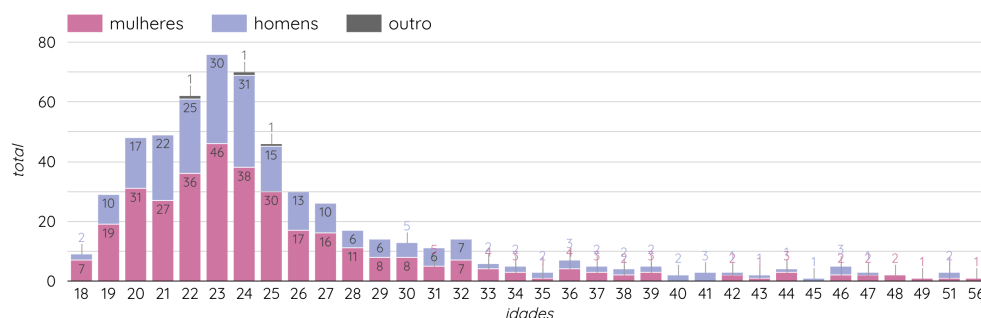


Figura 5.1. Gráfico de barras da distribuição de idades dos usuários cadastrados plataforma de votação separados por gênero.

Uma vez que a plataforma para classificar comentários foi divulgada em diversas redes sociais, 579 pessoas mostraram interesse em se cadastrar no site para rotular comentários, das quais aproximadamente 58,9% declararam ser do gênero feminino, 40,6% declararam ser do gênero masculino e os 0,5% restantes declararam ser não binários.

Esse dado mostra que a maior parte dos interessados em colaborar com uma pesquisa sobre “sexismo” foram mulheres, também tivemos pouco engajamento de pessoas

não binárias, o que torna a base de dado menos representativa quando se trata da opinião de pessoas desse último grupo.

Os usuários da plataforma online poderiam rotular quantos comentários quisessem, e no final coletamos 12.055 votos. Todos os comentários da base receberam pelo menos dois votos e a tabela 5.1 mostra a distribuição da quantidade de votos recebida pelos comentários. A tabela 5.1 também mostramos a concordância entre os votos atribuídos para os comentários, segregando a concordância pelo número de votos dos comentário.

Tendo como referência as interpretações descritas na tabela 2.1, observamos que comentários com seis votos ou mais tiveram ausência de concordância, os comentários com quatro votos tiveram concordância moderada entre os votos, os comentários com três votos tiveram concordância leve e os comentários com apenas dois votos tiveram concordância substantiva.

Sendo assim, verificamos que 93,64% dos comentários tiveram presença de concordância no rótulos atribuídos a eles, enquanto o restante dos comentários não apresentou concordância dos rotuladores entre os rótulos atribuídos.

O comentário que recebeu nove votos era “Ouvi dizer que se eu encostar a minha mão na mão de uma mulher na condução no momento que estou me segurando para não cair quando está cheio, já é considerado assédio sexual. Procede?”. Este comentário teve cinco votos para “sexista” e quatro votos para “não sexista”. Todos os votos para a classe sexista foram atribuídos por mulheres, enquanto os votos para não sexista foram divididos igualmente entre pessoas do gênero feminino e masculino.

A plataforma de rotulação fazia com que os comentários com empate no número de votos, fossem priorizados na exibição para os rotuladores que ainda não viram aquele comentário, a fim de desempatar o rótulo final para o comentário. Considerando que comentários com empate de votos, eram comentários difíceis de rotular, é esperado que a concordância entre votos decresça a medida que mais pessoa opinam em um comentário.

No trabalho original, mencionado previamente [3], tivemos cerca de quinhentos comentários com empate de votos para as duas classes, o que fez com que esses comentários não recebessem rótulo, neste trabalho submetemos os comentários a mais rotuladores e conseguimos reduzir esse número para 55.

A figura 5.2 mostra a relação de quantidades de votos atribuídos segregados por gêneros e idades. É notável que o volume de votos de pessoas na faixa etária de 19 até 28 anos é majoritariamente composto por mulheres. A média de votos atribuídos por usuário é de 68 votos e nenhum dos usuários teve zero votos, todos os rotuladores votaram em pelo menos um comentário.

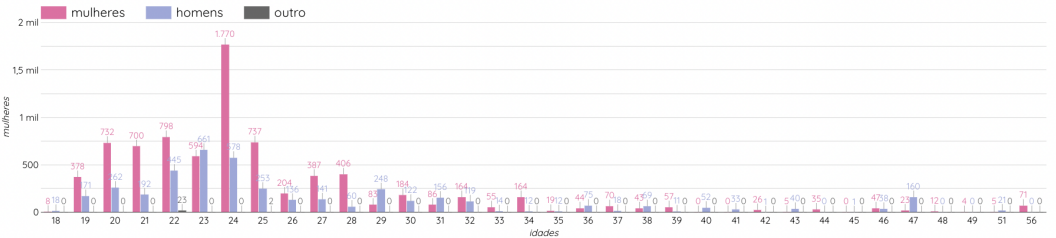


Figura 5.2. Gráfico de barras da distribuição de quantidade de votos pela idade e gênero dos usuários cadastrados plataforma de votação.

Número de votos	Quantidade de comentários	Fleiss's Kappa
9	1	0.00000
7	12	-0.07692
6	16	-0.08130
5	199	0.07008
4	809	0.44187
3	2533	0.38938
2	18	0.77777

Tabela 5.1. Número de comentários que recebeu cada quantidade de votos de 1 até 9 votos e Concordância de Fleiss's Kappa calculada pelo número de votos que cada comentário recebeu

Pelo gráfico da figura 5.1, podemos ver a distribuição de idades dos rotuladores, nele vemos que a faixa etária de 19 à 26 anos tem a maior quantidade de rotuladores, isso pode ser reflexo do fato de termos divulgado a plataforma de rotulação por redes sociais e as faixas etárias mais jovens tendem a interagir mais com o meio virtual.

Podemos notar também que a quantidade de mulheres é superior a quantidade de pessoas pertencentes a outros gêneros na maioria das faixas etárias, apenas nas faixas a partir de 29 anos os números são mais equilibrados.

Dos votos atribuídos a cada comentário, 65,6% foram de pessoas do gênero feminino, 34,2% foram atribuídos por pessoas do gênero masculino, e os 0,2% restantes foram atribuídos por pessoas não binárias. Esse fato reforça a hipótese de que mulheres se sentiram mais motivadas do que os demais gêneros a colaborar com a pesquisa.

As figuras 5.3, 5.4, 5.5 e 5.6 mostram o percentual de votos distribuídos por idade e gênero dos rotuladores do gênero masculino e feminino. Na análise dessas imagens consideramos que votos “correto” os verdadeiro-positivos ou verdadeiro-negativos, enquanto que os votos “incorretos” são os falso-negativos ou falso-positivos.

Na figura 5.3 vemos os votos corretos para a classe sexista. Pelo gráfico, podemos observar que os verdadeiro-positivos em comentários sexistas são majoritariamente femininos. Nas faixas etárias de 18, 23, 29, 31, 36 e 47 anos, os homens tiveram um

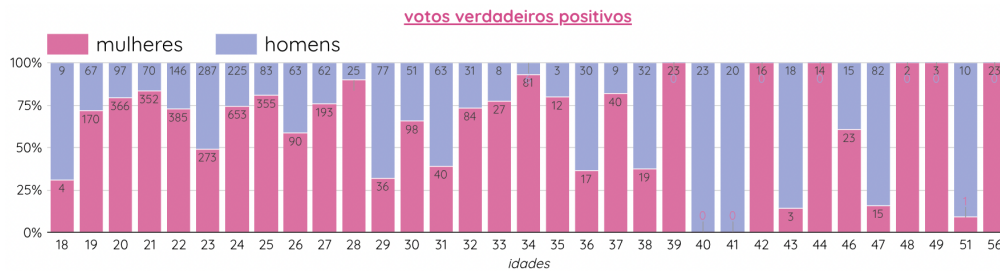


Figura 5.3. Gráfico de barras relacionando as porcentagens dos votos corretos para a classe “sexista” separado por gênero e distribuído por idades.

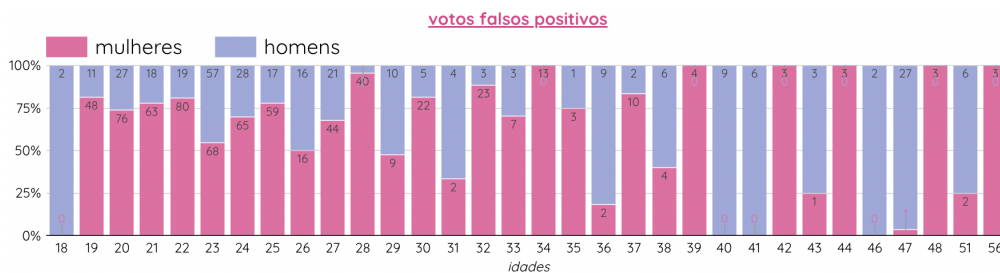


Figura 5.4. Gráfico de barras relacionando as porcentagens dos votos incorretos para a classe “sexista” separado por gênero e distribuído por idades.

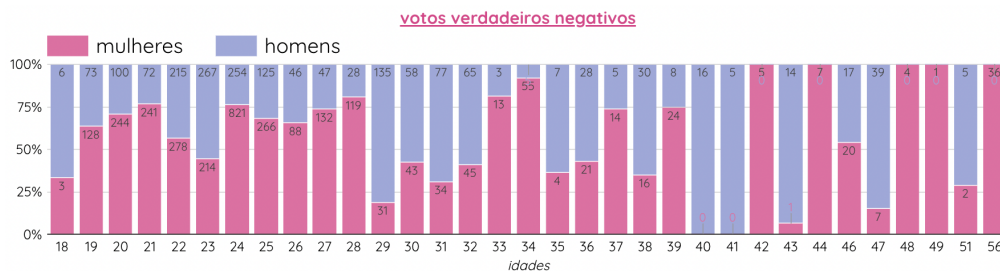


Figura 5.5. Gráfico de barras relacionando as porcentagens dos votos corretos para a classe “não sexista” separado por gênero e distribuído por idades.

volume maior de verdadeiro-positivos do que mulheres, embora o gráfico da figura 5.1 mostre que o volume de homens nessas faixas é superior ao número de mulheres.

A figura 5.4 mostra os votos incorretos para a classe sexista. Na imagem observamos que as mulheres possuem um maior volume de falsos positivos, esse dado indica que a sensibilidade de algumas rotuladoras ao sexismo pode ser maior do que a de rotuladores, uma vez que o sexismo tratado neste trabalho tem mulheres como alvo e alguns dos comentários da base se encaixam no conceito de sexismo benevolente [15], que raramente é percebido por pessoas do gênero masculino como discurso sexista.

O gráfico da 5.5 exibe o volume de verdadeiro-negativos, nele observamos a distribuição entre homens e mulheres é equilibrada até a faixa etária de 23 anos. Nas faixas

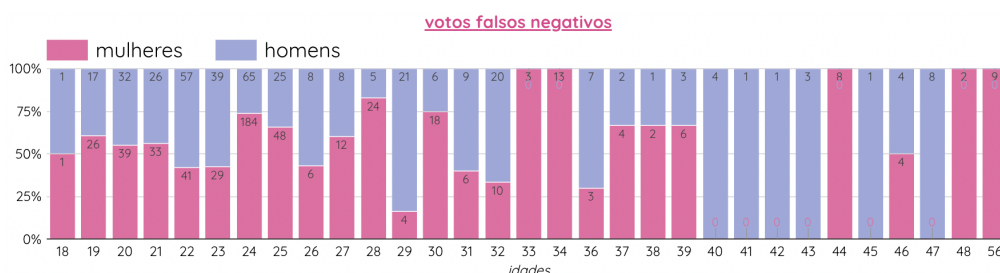


Figura 5.6. Gráfico de barras relacionando as porcentagens dos votos incorretos para a classe “não sexista” separado por gênero e distribuído por idades.

etárias de 24 a 28 as mulheres tiveram maior volume de votos corretos para a classe não sexista e a partir dos 29 anos, os verdadeiro-negativos atribuídos por homens são mais frequentes.

Nas figuras 5.6, são exibidos os falso-negativos. Nas faixas etárias de 18 a 23 anos, houve um equilíbrio entre os rótulos atribuídos por cada gênero. A partir da faixa etária de 24 a 28 anos podemos ver que as mulheres tiveram maior volume de votos incorretos para a classe não sexista. Em relação aos outros gráficos, é possível notar que neste houve um volume maior de votos relacionados ao gênero masculino, o que indica que o conceito de sexismo adotado por esses rotuladores é menos sensível que aquele adotado pelos demais rotuladores.

É importante ressaltar que temos um número superior de votos atribuídos pelo gênero feminino em relação ao gênero masculino, então é esperado que, em alguns casos, as mulheres apareçam mais que os homens nas análises acima. Mesmo com esse fato, os homens aparecem com uma frequência maior em determinados casos, como em falso-negativos.

Uma vez que os votos femininos são mais frequentes que votos masculinos, seria ideal ter mais rotuladores homens, bem como mais rotuladores de outros gêneros, com objetivo de ter comparações justas e evitar o viés emocional que votos femininos podem trazer, uma vez que mulheres são alvo direto do discurso sexista tratado neste trabalho.

5.2 Análise dos votos femininos

Analizamos a distribuição e detalhamento dos votos femininos, a fim de verificar o impacto gerado nos rótulos causado pela participação de pessoas que não são alvo direto do discurso sexista.

Utilizando novamente a tabela 2.1 para interpretar os dados da tabela 5.2, identificamos que, entre os votos femininos atribuídos aos comentários, todos tem concor-

Número de votos	Quantidade de comentários	Fleiss's Kappa
7	1	0.00000
6	3	0.00000
5	19	0.10722
4	242	0.39120
3	1058	0.44780
2	1474	0.35501
1	701	—

Tabela 5.2. Número de comentários que recebeu cada quantidade de votos femininos de 1 até 7 votos e Concordância de Fleiss's Kappa calculada pelo número de votos femininos que cada comentário recebeu

dância presente, mesmo que pobre para os comentário com cinco, seis e sete votos. Os comentários com quatro e três votos tiveram concordância moderada, os comentários com cinco votos tiveram concordância pobre e os comentários com dois votos tiveram concordância leve.

Pela análise da tabela 5.2, afirmamos que a concordância entre os votos é mais presente quando olhamos apenas para votos de pessoas do mesmo gênero, uma vez que, em comparação com a tabela 5.1 não tivemos nenhuma ausência de concordância.

As análises dos gráficos 5.7, 5.8, 5.9 e 5.10 exibem análises dos votos femininos em relação às classes de comentários disponíveis. Observamos que não há diferenças nas quantidades de votos femininos em cada gráfico em relação aos gráficos 5.3, 5.4, 5.5 e 5.6.

No entanto, observamos que 1.482 comentários tiveram uma mudança no valor da média de votos, que não foi grande o suficiente para alterar o rótulo final de cada comentário. Com isso, definimos que os votos de outros gêneros não foram significativos para alterar o rótulo final de comentários da base, atribuímos esse fato ao grande volume de votos femininos presente na base de dados.

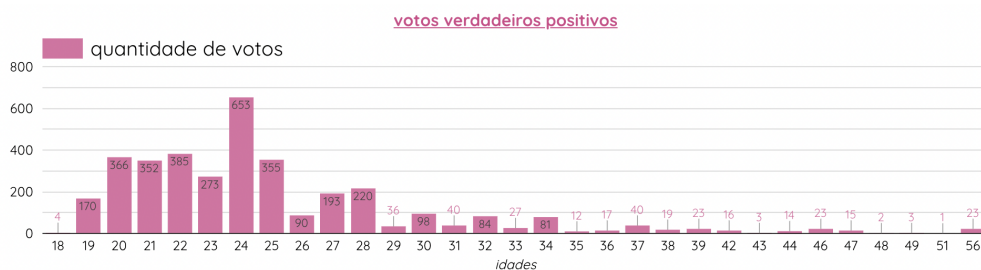


Figura 5.7. Gráfico de barras relacionando as porcentagens dos votos femininos corretos para a classe “sexista” separado por gênero e distribuído por idades.

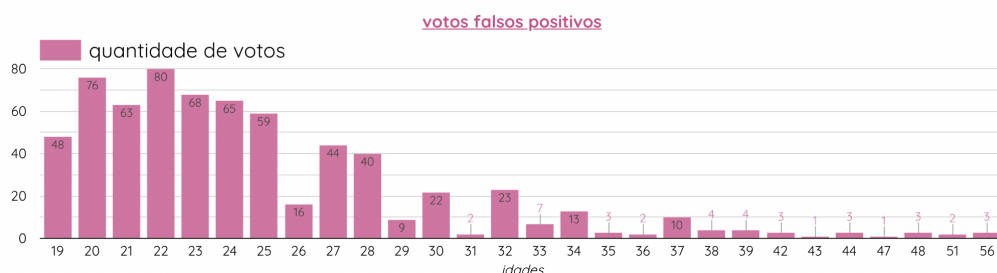


Figura 5.8. Gráfico de barras relacionando as porcentagens dos votos femininos incorretos para a classe “sexista” separado por gênero e distribuído por idades.

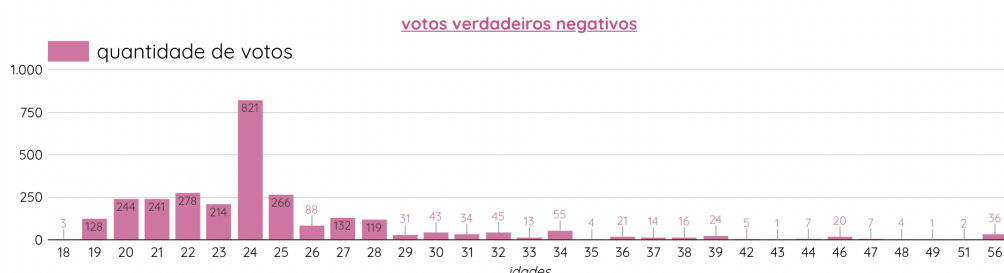


Figura 5.9. Gráfico de barras relacionando as porcentagens dos votos femininos corretos para a classe “não sexista” separado por gênero e distribuído por idades.

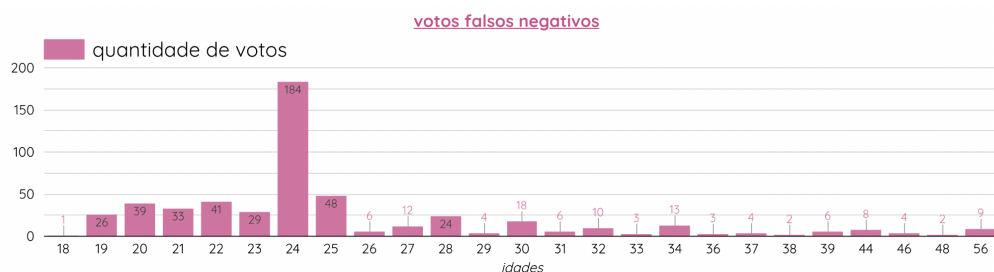


Figura 5.10. Gráfico de barras relacionando as porcentagens dos votos femininos incorretos para a classe “não sexista” separado por gênero e distribuído por idades.

5.3 Análise de concordâncias

Além da análise realizada na tabela 2.1, realizamos a análise da concordância entre os comentários com distribuição balanceada de votos por gênero.

Separamos os comentários que receberam quatro votos, sendo dois de pessoas do gênero masculino e dois do gênero feminino, independente da classe votada. Após a separação, calculamos Fleiss’s Kappa para o conjunto de votos apenas desses comentários considerando ambos os gêneros e obtivemos a concordância K . Para esses mesmos comentários, calculamos a concordância considerando apenas os votos femininos K_f e

Fleiss's Kappa	
K	0.42565
K_f	0.51234
K_m	0.43958

Tabela 5.3. Concordância geral, feminina e masculina para comentários com quatro votos distribuídos igualmente entre esses gêneros.

apenas os votos masculinos K_m , os resultados estão listados na tabela 5.3.

Pela tabela 5.3 verificamos que todos os resultados obtiveram concordância moderada segundo a interpretação da tabela 2.1, no entanto a concordância entre os votos femininos K_f foi maior que a concordância entre os votos masculinos K_m , o que indica que o consenso entre mulheres é levemente maior que entre homens ao rotular um comentário como sexista ou não.

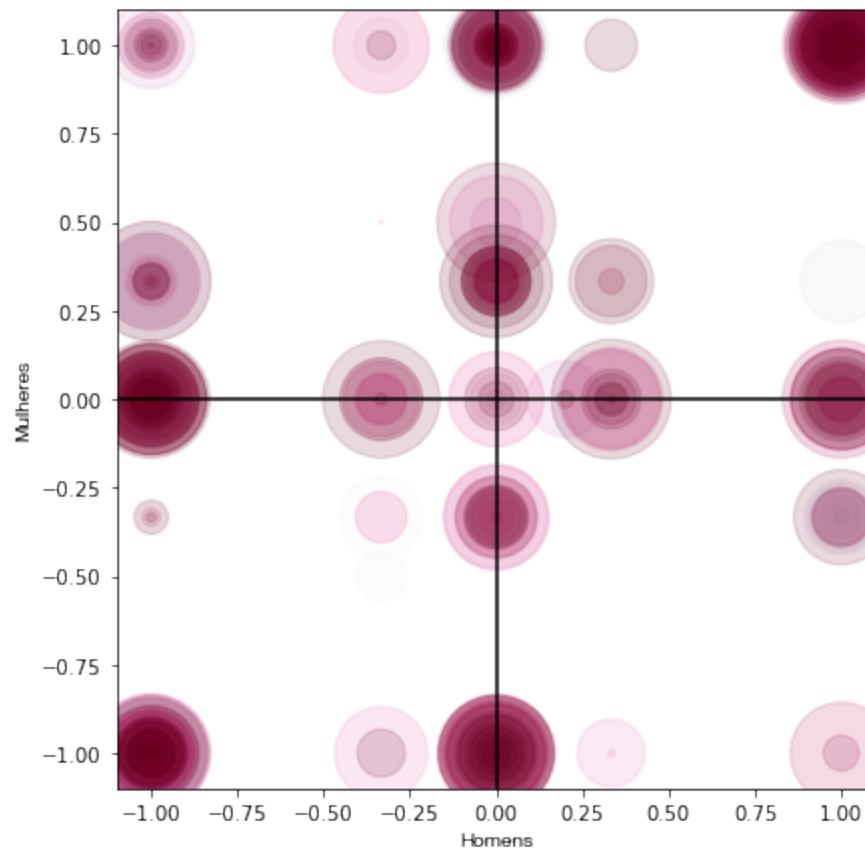


Figura 5.11. Distribuição de votos femininos e masculinos para o conjunto de comentários com pelo menos dois votos de cada gênero.

Para entender melhor a distribuição de votos entre os gêneros nos comentários

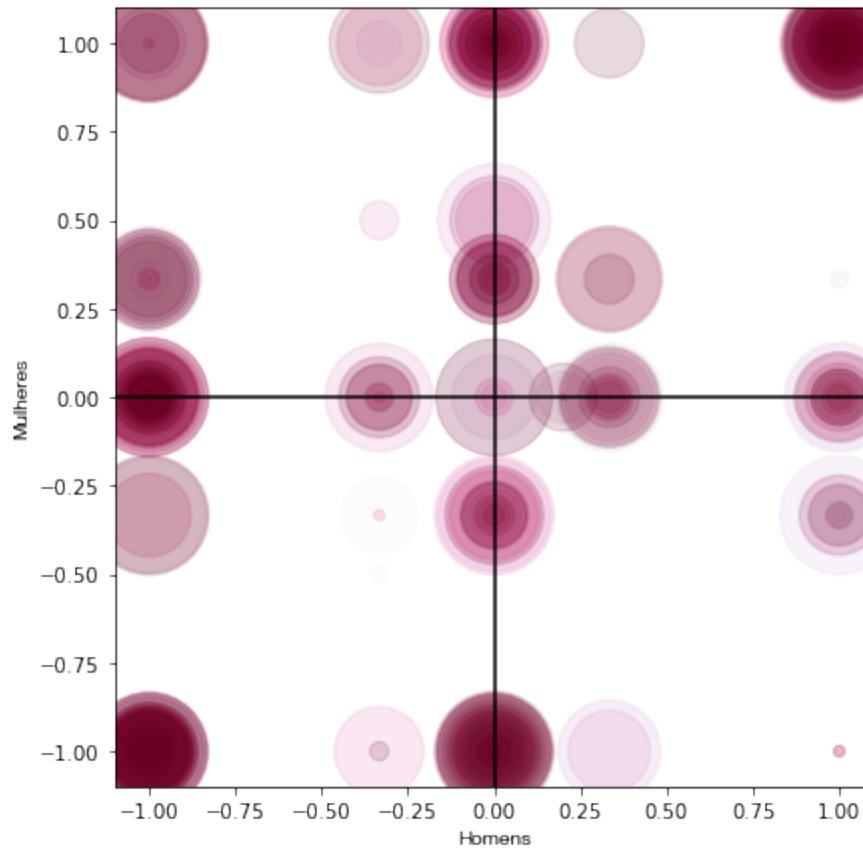


Figura 5.12. Prevalência de votos femininos e masculinos para o conjunto de comentários com pelo menos dois votos de cada gênero.

com pelo menos dois votos atribuídos por cada gênero, plotamos os gráficos das figuras 5.11 e 5.12 onde consideramos que os votos para a classe sexistas possuem o valor -1 e os votos para não sexista possuem valor 1. Para cada comentário, calculamos a média de votos de homens e mulheres separando os resultados respectivamente nos eixos X e Y do gráfico das figuras 5.11 e 5.12.

Sendo assim, o primeiro quadrante representa médias que tendem para o rótulo sexista para os dois gêneros. O segundo quadrante mostra médias nas quais mulheres tenderam para o rótulo não sexista e homens tenderam para o sexista. O terceiro quadrante mostra médias em que ambos os gêneros tenderam para o rótulo não sexista. O quarto quadrante mostra médias onde as mulheres tenderam para o rótulo sexista e os homens tenderam para o não sexista.

Observando o gráfico da figura 5.11, notamos que há uma grande quantidade de médias de votos concentradas nos eixos, isso ocorre quando um dos gêneros teve

empate entre os seus rótulos, os pontos no eixo X representam empates entre os votos femininos, e os pontos no eixo Y representam o mesmo para os votos masculinos. Com isso, observamos que o volume de empates entre votos é maior no eixo X, o que indica que homens a discordância dos homens pode ser maior em relação às mulheres.

Sabendo que a prevalência representa o impacto da ocorrência de um voto levando o total dos votos observados em consideração, calculamos também a prevalência para os votos considerados na figura 5.11, os resultados podem ser visualizados em 5.12. É notável que o primeiro e o quarto quadrante possuem maior impacto em relação ao restante dos padrões de votos.

5.4 Análise dos comentários

Considerando todos os votos coletados, dos 3.588 comentários coletados, 1.821 foram classificados como sexistas, 1.712 como não sexistas e 55 não receberam rótulo. O vocabulário referente à base de dados era composto por 13.628 palavras, incluindo palavras que contêm caracteres numéricos, erros de digitação e variações das mesmas palavras.¹

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 detecte ofensa em comentários com esses termos alfanuméricos.

O vocabulário dos comentários sexistas possui 57 termos alfanuméricos, enquanto o vocabulário dos comentários não sexistas possui apenas 38. Nos vocabulário sexista, esses termos são majoritariamente palavras ofensivas, enquanto no vocabulário não sexista, são majoritariamente erros de digitação. Sendo assim, assumimos a hipótese de que termos compostos por números e letras provavelmente são sexistas.

5.4.1 Análise de n-gramas

Afim de entender quais eram as palavras que caracterizavam cada discurso, geramos as nuvens de palavras dos comentários sexistas e não sexistas, que são exibidas nas imagens 5.13 e 5.14 respectivamente.

¹A base de comentários está disponível em <https://github.com/mlpbraga/sexism-detection-notebooks>.

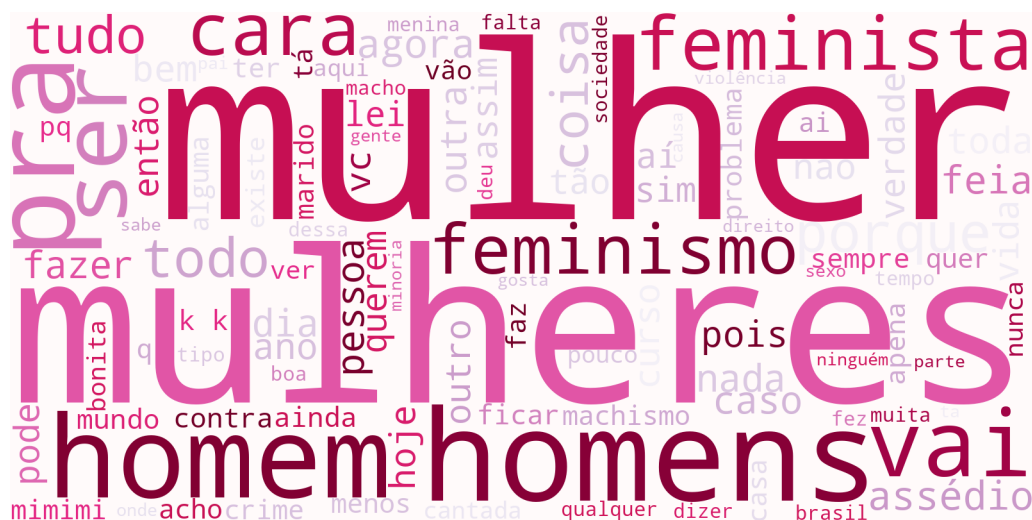


Figura 5.13. Nuvem de palavras de comentários sexistas.



Figura 5.14. Nuvem de palavras de comentários não sexistas.

Pelas imagens 5.13 e 5.14, notamos que os termos “mulher”, “mulheres”, “homens” e “homem” são muito presentes em ambos os discursos. Inferimos que isso se dá pelo fato de que as notícias que coletamos estavam relacionadas a discussões sobre gênero, o que faz com que muitos comentários relacionem e comparem mulheres e homens tanto de forma negativa quanto de forma positiva.

No entanto, observamos que as referências a feminismo são mais frequentes no discurso sexista. Além disso, no discurso sexista temos a presença marcante dos termos “mimimi”, “assédio”, “feia” e “macho”, que são termos comumente usado em discursos que visam ofender ou diminuir mulheres ou o movimento feminista.

Sabendo que determinadas palavras aparecem em ambos os discursos, analisamos o corpus buscando quais termos eram mais importantes para cada discurso, então calculamos a diferença de frequências entre as palavras em cada classe de texto.

Formalmente, 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 n -gramas presentes em D , calculamos as frequências normalizadas F_s e F_n para cada n -grama $t_i \in \sigma$.

Definimos que 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 n -grama em busca das palavras relevantes na diferenciação das duas classes de comentários. A tabela 5.4 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 5.4, notamos que algumas das palavras mais relevantes no discurso sexista são tradicionalmente *stopwords*, como “uma”, “ela”, “elas”, “na”, “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”, “na” e “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 5.4 indica que em contextos de notícias onde o tópico se refere a mulheres, assédio, feminismo ou feminicídio, temos uma presença relevante de comentários com adjetivos pejorativos que buscam ofender mulheres.

Além do adjetivo “feia”, observamos que entre as palavras mais relevantes para o discurso sexista também temos o seu antônimo “bonita”, isso acontece porque é comum que comentários sexistas busquem ofender mulheres por conta de sua aparência, como em “E a moça da notícia nem é bonita”. Também encontramos comentários em que as pessoas atribuem o sucesso das mulheres apenas a sua aparência, como o comentário “Só chegou lá porque é bonita”.

O adjetivo “linda” também aparece como relevante no discurso sexista, e segue os mesmos padrões de aparições do termo “bonita”. Além disso, assim como ofensas a aparência de mulheres são comuns no discurso sexista hostil, também é comum que o discurso sexista benevolente traga elogios a mulheres, como é o caso em “Essa cientista

Unigrama	F_s	F_n	$F_s - F_n$
mulheres	0.009441	0.006051	0.003389
ela	0.007559	0.005015	0.002544
de	0.041283	0.038780	0.002503
mulher	0.010230	0.007790	0.002440
uma	0.012203	0.009829	0.002374
as	0.009016	0.006954	0.002062
homens	0.006071	0.004012	0.002059
feia	0.002216	0.000201	0.002015
elas	0.002520	0.000535	0.001985
assédio	0.002398	0.000836	0.001562
feministas	0.001912	0.000501	0.001411
homem	0.005100	0.003711	0.001389
feminismo	0.002398	0.001070	0.001328
na	0.007801	0.006486	0.001316
com	0.012021	0.010865	0.001156

Tabela 5.4. Relação dos quinze unigramas com maior $F_s - F_n$ positivo.

Unigrama	F_s	F_n	$F_s - F_n$
ser	0.005525	0.008224	-0.002699
que	0.044015	0.046269	-0.002254
Brasil	0.001002	0.002741	-0.001740
sua	0.001457	0.003042	-0.001585
você	0.001518	0.002875	-0.001357
em	0.008651	0.009963	-0.001311
pessoas	0.001093	0.002340	-0.001247
sobre	0.000516	0.001738	-0.001222
comentários	0.000364	0.001571	-0.001207
não	0.022008	0.023101	-0.001093
do	0.011414	0.012470	-0.001056
está	0.002368	0.003410	-0.001042
lei	0.000729	0.001738	-0.001010
presidente	0.000304	0.001304	-0.001000
como	0.003187	0.004179	-0.000992

Tabela 5.5. Relação dos quinze unigramas com menor $F_s - F_n$ negativo.

Bigrama	F_s	F_n	$F_s - F_n$
as mulheres	0.009876	0.005088	0.004788
que as	0.004444	0.002165	0.002279
uma mulher	0.006222	0.004113	0.002108
um homem	0.002864	0.000866	0.001998
os homens	0.004938	0.003247	0.001690
as feministas	0.001975	0.000541	0.001434
do que	0.004543	0.003139	0.001404
são minoria	0.001580	0.000216	0.001364
mulheres que	0.002963	0.001732	0.001231
por causa	0.001975	0.000758	0.001217
das mulheres	0.002271	0.001082	0.001189
que ela	0.004543	0.003464	0.001079
em um	0.001580	0.000541	0.001039
gosta de	0.001975	0.000974	0.001001
com um	0.002074	0.001082	0.000991

Tabela 5.6. Relação dos quinze bigramas com maior $F_s - F_n$ positivo

Bigrama	F_s	F_n	$F_s - F_n$
que não	0.007900	0.010284	-0.002383
ser humano	0.000691	0.002381	-0.001690
no Brasil	0.001679	0.003356	-0.001677
tem que	0.003456	0.004979	-0.001523
que você	0.000889	0.002273	-0.001384
ou não	0.000593	0.001948	-0.001356
todos os	0.001679	0.003031	-0.001352
que ser	0.001185	0.002490	-0.001305
as pessoas	0.000988	0.002273	-0.001286
não sabe	0.000691	0.001840	-0.001149
pessoas que	0.000296	0.001407	-0.001111
Tem que	0.000198	0.001299	-0.001101
todo mundo	0.000494	0.001515	-0.001022
sabe que	0.000395	0.001407	-0.001012
de sua	0.000395	0.001407	-0.001012

Tabela 5.7. Relação dos quinze bigramas com menor $F_s - F_n$ negativo

[...] é linda, nunca iria querer ter aula com ela, jamais prestaria atenção nas aulas”.

A palavra “feminazi” também aparece entre as palavras relevantes para o discurso sexista, termo comumente usado para ofender mulheres feministas. Além deste termo, outras variações de “feminismo” e “feminista” aparecem com maior frequência no discurso sexista, isso indica que a maioria dos comentários sobre o movimento feminista são pejorativos.

Termos que representam risadas, como variações de “kkk”, “haha” ou “rs” também aparecem na lista de palavras mais relevantes para o discurso sexista, normalmente em comentários que desenham dos casos sobre feminicídio, assédio e tópicos relacionados a mulheres nas notícias apresentas.

Algumas notícias eram relacionadas a conquistas e ideais de mulheres feministas, nestas notícias encontramos frequentemente o termo “mimimi” e variações dele, este termo foi o quarto adjetivo mais frequente no discurso sexista. Consideramos “mimimi” como adjetivo por ele ser usado majoritariamente para atribuir uma característica aos ideais feministas.

Na tabela 5.5 são exibidos os unigramas mais relevantes no discurso não sexista. Os termos presentes nessa listagem não apontam referência à gênero e não possuem relação semântica.

Além dos unigramas, calculamos a diferença das frequências para os bigramas mais frequentes em cada discurso, exibidos nas tabelas 5.6 e 5.7 respectivamente. A tabela 5.6 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 5.7 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 *“as feministas saíram do século de um homem e isso não impede que elas preguem ódio contra homens...”*.

5.4.2 Análise de quantitativa

Para obter informações sobre o tipo de comentário que gera mais engajamento, montamos um gráfico com a porcentagem de *likes* e *dislikes* para cada classe de comentários, que pode ser visualizado na figura 5.15.

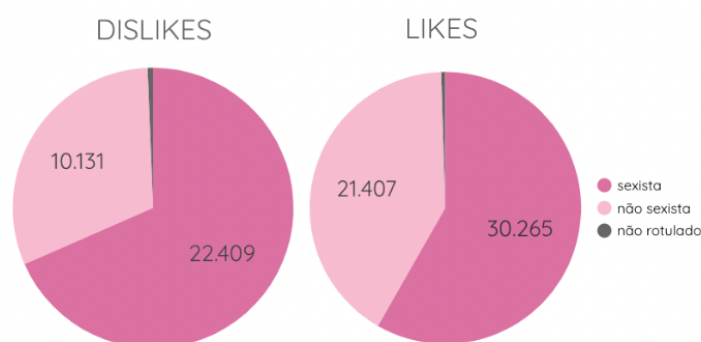


Figura 5.15. Gráficos de pizza representando as quantidades de *likes* e *dislikes* para cada classe de comentários.

O gráfico da figura 5.15 mostra que os comentários sexistas geram mais engajamento do que os demais comentários uma vez que a soma de *likes* e *dislikes* dessa classe é maior, 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.

Ainda no gráfico da figura 5.15, 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, uma vez que comentários sexistas recebem mais engajamento negativo.

As tabelas 5.8 e 5.9 exibem as estatísticas de valores, máximos, mínimos, médios, desvios padrões, primeiro, segundo e terceiro quartil para cada uma das características quantitativas da base, para o discurso sexista e não sexista respectivamente.

Comparando os dados das duas tabelas, podemos afirmar que para a classe sexista, o desvio padrão de *likes* é maior do que para *dislikes*, enquanto para a classe não sexista, temos a situação inversa. Podemos ver que para a maioria das situações, os valores de engajamento são baixos para ambas as classes, sendo diferenciais apenas em casos específicos.

Para todos os comentários, calculamos o índice de legibilidade FleschBr, a fim de identificar o quão fáceis de entender são considerados os textos levando em conta as quantidades de palavras e frases presentes no texto. Observamos que os valores de

	likes	dislikes	qtd. palavras	qtd. caracteres	legibilidade
mínimo	0	0	1	2	-5.00
25%	2	0	9	51	223.62
mediana	6	2	17	96	228.58
75%	15	7	31.75	179	229.90
máximo	688	1196	819	7050	230.30
média	17.32	13.04	25.15	144.17	223.95
desvio padrão	38.99	50.19	30.10	210.70	13.72

Tabela 5.8. Relação das estatísticas relacionadas às características quantitativas do discurso sexista.

	likes	dislikes	qtd. palavras	qtd. caracteres	legibilidade
mínimo	0	0	1	2	0.00
25%	1	0	8	44	223.38
mediana	4	1	16	91	227.97
75%	12	4	32	950	229.71
máximo	729	842	159	950	230.30
média	13.55	6.41	23.92	137.17	222.31
desvio padrão	40.66	29.84	23.24	134.94	17.51

Tabela 5.9. Relação das estatísticas relacionadas às características quantitativas do discurso não sexista.

legibilidade são muito similares nos dois discursos, bem como para as demais métrica, exceto pelo desvio padrão, que é mais alto no discurso não sexista.

Ainda nas tabelas 5.8 e 5.9, constatamos que as distribuições de valores para índices de legibilidade, quantidades de palavras e caracteres são muito similares para ambas as classes, no entanto os comentários sexista tendem a ter o conteúdo maior em relação à outra classe de comentários em alguns casos.

Sendo assim, assumimos que as características quantitativas são bons complementos, mas não são suficientes para diferenciar cada classe de comentários sozinhas.

5.5 Classificação automática

Testamos cada modelo usando a validação cruzada de 10 folds, os resultados obtidos para os modelos de base estão na tabela 5.10, enquanto os resultados para as nossas implementações do SVM, KNN e RF estão listados nas tabelas 5.12, 5.13 e 5.14 respectivamente.

Método	P_s	P_n	R_s	R_n	$F1_s$	$F1_n$
BERT	0.46	0.54	0.55	0.45	0.50	0.49
SVM	0.63	0.57	0.57	0.64	0.59	0.60
LR	0.66	0.60	0.60	0.66	0.63	0.63

Tabela 5.10. Resultados obtidos na aplicação do modelo BERT e dos modelos SVM e LR implementados como linhas de base.

Pela tabela 5.10 observamos que os modelos de base tiveram desempenho baixo na detecção de discurso sexista usando a base criada neste trabalho, dentre eles o que mostrou melhores resultados foi o modelo LR, atingindo até 0,63 de F1 para ambas as classes de discurso. Os resultados para o modelo BERT podem ser justificados pela baixa quantidade de dados que temos em relação ao trabalho de Pamungkas et al [30], uma quantidade maior de comentários rotulados poderia melhorar os resultados para esse modelo, o que indica que a adição de mais comentário rotulados à base é necessária para fazer bom uso deste modelo de deep learning.

Para os modelos LR e SVM implementados seguindo as abordagens de Davidson et al [6]. acreditamos que o pre processamento do texto impactou na aprendizagem do modelo, uma vez que nessa abordagem aplicamos técnicas de *stemming* e remoção de *stopwords* e, por se tratar de um trabalho focado no sexismo direcionado a mulheres, as variações de gênero, pronomes e artigos femininos são relevantes para caracterizar o texto como sexista ou não.

Observando a tabela 5.12, identificamos que os melhores resultados para o SVM utilizaram as características textuais de POS tagging, TF de unigramas e bigramas. Concluímos que para o SVM os resultados foram muito similares, uma vez que o teste de Wilcoxon resultou em 0.00506 para a maioria das comparações entre eles, o que significa que entre utilizar o SVM com as características quantitativas ou textuais, é favorável o uso de características textuais, dado que os resultados para o uso de POS tagging e de unigramas e bigramas também trouxe bons resultados.

Na tabela 5.13, observamos que a combinação de características com melhor resultado para o KNN é o uso das características textuais e quantitativas combinadas. Ao executarmos o teste de Wilcoxon para cada um dos modelos, observamos que o resultado para as comparações também foi 0.00506 na maioria dos casos, exceto nas comparações entre o uso de características quantitativas e o uso de bigramas, o que indica que entre os melhores resultados para o KNN, não podemos afirmar que um dos modelos é melhor que o outro, mas sabemos que entre o uso de bigramas e o uso das características quantitativas isoladas, temos um que performa melhor.

Para o RF, a tabela 5.14 mostra que, assim como no SVM, os melhores resultados

foram obtidos utilizando características textuais, um dos resultados inclui o uso de características quantitativas combinadas com características textuais. Ao executarmos o teste de Wilcoxon para os resultados desses modelos, vimos que as comparações foram similares aos casos anteriores.

Para avaliar e comparar os resultados obtidos pelos nossos modelos com os modelos de base, executamos o teste de Wilcoxon [37] com os valores de F1 obtidos por cada modelo, os resultados com p-value maior que 0.05 podem ser observados na tabela 5.11.

Assumindo um grau de confiança de 5% podemos concluir pelos resultados exibidos na tabela, que as distribuições de resultados são diferentes nas comparações de todos os modelos, o que quer dizer que nas comparações em que o p-value atinge este grau de confiança, podemos assumir que um dos modelos foi superior ao outro.

Observando a tabela 5.11 juntamente com os resultados das tabelas 5.10, 5.12, 5.13 e 5.14, vemos que a maioria dos modelos implementados neste trabalho tiveram resultados superiores às implementações dos modelos de base para o dataset criado. Concluimos que a melhor combinação de features e algoritmo, encontrada foi o KNN com características textuais e quantitativas, além disso todos os modelos performaram bem utilizando o POS tagging do texto sexista.

Comparação de modelos	p-value
SVM $Q_l + Q_d + Q_c + Q_p$ x Baseline SVM	0.07446183141740546
SVM $Q_l + Q_d + Q_c + Q_p$ x Baseline LR	0.07446183141740546
SVM Q_{il} x Baseline LR	0.09260069782067576
KNN Q_p x BERT	0.13941397332153205
SVM $TF_{us} + TF_{un} + Q_l + Q_s + Q_c + Q_p$ x Baseline LR	0.1688069535565081
KNN TF_{bs} x Baseline LR	0.20262160771231197
SVM TF_{bn} x Baseline LR	0.20262160771231197
KNN Q_d x BERT	0.24112136827741826
KNN Q_d x Baseline SVM	0.24112136827741826
RF Q_{il} x Baseline LR	0.24112136827741826
KNN $Q_c + Q_p$ x BERT	0.2845026979112075
RF $Q_l + Q_d$ x Baseline LR	0.2845026979112075
KNN $Q_l + Q_d + Q_c + Q_p$ x BERT	0.2845026979112075
KNN $TF_{bs} + TF_{bn}$ x Baseline LR	0.3328797545683888
SVM TF_{bs} x BERT	0.3862707203664827
SVM $Q_l + Q_d$ x BERT	0.3862707203664827
KNN TF_{bn} x BERT	0.4445867389117455
RF Q_p x Baseline LR	0.4445867389117455
KNN TF_{bn} x Baseline SVM	0.5076243443095237
KNN TF_{bn} x Baseline LR	0.5076243443095237
KNN Q_l x BERT	0.6464622026641696
SVM $Q_c + Q_p$ x Baseline SVM	0.7212766990291557
KNN TF_{bs} x Baseline SVM	0.7988593499960496
SVM Q_l x BERT	0.7988593499960496
KNN Q_{il} x BERT	0.7988593499960496
KNN $TF_{bs} + TF_{bn}$ x Baseline SVM	0.7988593499960496
KNN Q_c x BERT	0.8784817434328712
SVM Q_c x Baseline SVM	0.9593536340401635
SVM Q_p x Baseline SVM	0.9593536340401635
KNN $Q_l + Q_d$ x BERT	0.9593536340401635

Tabela 5.11. Exibe valores de p-values encontrados na execução do teste estatístico de Wilcoxon para os modelos descritos no trabalho.

features	P_s	P_n	R_s	R_n	$F1_s$	$F1_n$
TF_{us}	0.97962	0.82470	0.80247	0.98267	0.88171	0.89652
TF_{un}	0.97415	0.78424	0.74723	0.97911	0.84533	0.87067
TF_{bs}	0.65011	0.54815	0.41304	0.76116	0.50481	0.63712
TF_{bn}	0.56423	0.58837	0.74869	0.38278	0.64318	0.46323
TF_{pos}	0.99449	0.97910	0.97950	0.99419	0.98688	0.98652
Q_l	0.58947	0.53586	0.47640	0.64563	0.52662	0.58534
Q_d	0.62954	0.52031	0.29826	0.81273	0.40425	0.63412
Q_c	0.55371	0.54144	0.64913	0.44130	0.59720	0.48547
Q_p	0.55048	0.54062	0.65991	0.42520	0.59984	0.47519
Q_{il}	0.52879	0.54341	0.83716	0.20392	0.64782	0.29507
$Q_c + Q_p$	0.55480	0.54068	0.63982	0.45172	0.59384	0.49146
$Q_l + Q_d$	0.60868	0.53848	0.43703	0.70063	0.50808	0.60850
$Q_l + Q_d + Q_c + Q_p$	0.57021	0.56511	0.66458	0.46471	0.61346	0.50937
$TF_{us} + TF_{un}$	0.99039	0.85523	0.84231	0.99124	0.91011	0.91807
$TF_{bs} + TF_{bn}$	0.59686	0.78747	0.91413	0.34076	0.72194	0.47519
$TF_{us} + TF_{un} + TF_{bs} + TF_{bn}$	0.99040	0.84623	0.83074	0.99116	0.90331	0.91284
$TF_{us} + TF_{un} + Q_l + Q_s + Q_c + Q_p$	0.57649	0.57443	0.67181	0.47244	0.62010	0.51771

Tabela 5.12. Relação dos resultados obtidos para o SVM.

features	P_s	P_n	R_s	R_n	$F1_s$	$F1_n$
TF_{us}	0.91838	0.87924	0.87434	0.90247	0.88971	0.88436
TF_{un}	0.88614	0.87018	0.86577	0.86857	0.86744	0.86052
TF_{bs}	0.65024	0.66373	0.62735	0.56906	0.59408	0.52569
TF_{bn}	0.69669	0.61484	0.58157	0.61368	0.55347	0.56819
TF_{pos}	0.98961	0.99187	0.99236	0.98903	0.99093	0.99039
Q_l	0.53295	0.49358	0.52256	0.50085	0.51975	0.49088
Q_d	0.55631	0.46832	0.60094	0.43017	0.54285	0.40650
Q_c	0.52512	0.49595	0.53082	0.48975	0.52672	0.49143
Q_p	0.53692	0.50566	0.56011	0.48079	0.54345	0.48760
Q_{il}	0.51614	0.48343	0.52187	0.47660	0.51412	0.47523
$Q_c + Q_p$	0.51313	0.48171	0.49421	0.50072	0.50258	0.49016
$Q_l + Q_d$	0.52566	0.48346	0.51507	0.49481	0.51507	0.48415
$Q_l + Q_d + Q_c + Q_p$	0.54790	0.51485	0.51676	0.54596	0.53138	0.52945
$TF_{us} + TF_{un}$	0.95331	0.93260	0.93471	0.94992	0.94207	0.93914
$TF_{bs} + TF_{bn}$	0.71842	0.65800	0.60082	0.66463	0.60107	0.60040
$TF_{us} + TF_{un} + TF_{bs} + TF_{bn}$	0.94126	0.92760	0.93035	0.93658	0.93379	0.92984
$TF_{us} + TF_{un} + Q_l + Q_s + Q_c + Q_p$	1.00000	0.99361	0.99400	1.00000	0.99699	0.99678

Tabela 5.13. Relação dos resultados obtidos para o KNN.

features	P_s	P_n	R_s	R_n	$F1_s$	$F1_n$
TF_{us}	0.93701	0.78296	0.75325	0.94556	0.83463	0.85640
TF_{un}	0.95543	0.77490	0.73832	0.96299	0.83270	0.85857
TF_{bs}	0.65411	0.53595	0.35293	0.79887	0.45796	0.64123
TF_{bn}	0.56587	0.69297	0.88483	0.27544	0.69004	0.39355
TF_{pos}	0.97651	0.99436	0.99434	0.97424	0.98528	0.98412
Q_l	0.59377	0.55878	0.56580	0.58688	0.57891	0.57193
Q_d	0.60720	0.55656	0.52323	0.63879	0.56166	0.59443
Q_c	0.58993	0.62155	0.74423	0.44734	0.65786	0.51950
Q_p	0.56355	0.56920	0.69817	0.42317	0.62313	0.48434
Q_{il}	0.56306	0.58955	0.74943	0.37837	0.64127	0.45611
$Q_c + Q_p$	0.62261	0.69546	0.80040	0.48173	0.69998	0.56786
$Q_l + Q_d$	0.63175	0.59801	0.61061	0.61917	0.62055	0.60795
$Q_l + Q_d + Q_c + Q_p$	0.70743	0.69519	0.71997	0.68203	0.71341	0.68828
$TF_{us} + TF_{un}$	0.95145	0.84198	0.83176	0.95429	0.88738	0.89449
$TF_{bs} + TF_{bn}$	0.57663	0.83837	0.95501	0.25131	0.71888	0.38628
$TF_{us} + TF_{un} + TF_{bs} + TF_{bn}$	0.94973	0.84399	0.83455	0.95268	0.88818	0.89488
$TF_{us} + TF_{un} + Q_l + Q_s + Q_c + Q_p$	0.95603	0.84687	0.83722	0.95837	0.89247	0.89905

Tabela 5.14. Relação dos resultados obtidos para o RF.

Capítulo 6

Considerações finais

Nesta dissertação, apresentamos abordagens para a seleção de características relevantes na detecção automática de sexismo em um dataset de comentários em portais de notícia. Descrevemos detalhadamente a montagem da base de dados e a análise dos rótulos atribuídos a cada comentário para a obtenção dos rótulos finais utilizados na base de treino. Realizamos também a comparação do uso de diferentes algoritmos de classificação automática, combinando características do discurso sexista, a fim de identificar qual a melhor forma de realizar a classificação automática desse discurso. Neste capítulo, descrevemos as principais conclusões obtidas através da análise dos dados na seção 6.1, exibimos as principais contribuições na seção 6.2 e descrevemos os trabalhos futuros na seção 6.3.

6.1 Conclusões

Identificamos que a maior parte das pessoas interessadas em contribuir com a pesquisa, foram pessoas do gênero feminino (58,9%), e do total de votos atribuídos a comentários a maioria (65,6%) também foi de pessoas do gênero feminino. Esse dado mostra que as pessoas mais engajadas em colaborar com um trabalho sobre sexismo direcionado a mulheres foram as próprias mulheres. Apesar da diferença entre o engajamento dos diferentes grupos de gêneros, observamos que 93,64% dos votos atribuídos apresentaram concordância entre eles. Isso mostra que os rótulos atribuídos são representativos da opinião majoritária de rotuladores. Considerando apenas os votos femininos, o resultado dos rótulos finais se manteve igual, no entanto a concordância entre as mulheres foi superior, o que faz com que a classificação final dos comentários seja mais confiável.

Neste trabalho, tivemos a participação majoritária de pessoas dentro das faixas etárias de 20 a 30 anos. Seria interessante, executarmos as análises em cima de uma

base de dados balanceada entre diversas faixas etárias, afinal a mulher ocupou espaços diferentes com o passar dos anos e ultimamente, muito já foi conquistado pelo gênero feminino [26]. Percebemos também, que a maioria dos falsos positivos foi atribuído por mulheres, bem como os verdadeiros positivos. A partir desses dados, podemos assumir que mulheres são mais sensíveis à presença de sexismo em sentenças do que os demais gêneros, considerando principalmente a presença de sexismo benevolente em alguns comentários da base de dados.

Durante análise dos comentários, verificamos que pronomes, artigos e variações de gênero são muito relevantes para identificar o discurso sexista na língua portuguesa, bem como a atribuição de algumas características utilizando adjetivos pejorativos (em caso de sexismo hostil) ou não (sexismo benevolente). Comparando os resultados da aplicação do pre processamento do texto implementado por Davidson et al. [6] com o pre processamento que fizemos, confirmamos esse fato, portanto esses elementos caracterizam o texto como sexista ou não, por esse motivo as implementações utilizando POS tagging como características performam bem em todos os algoritmos.

A presença de palavras, ou conjuntos de palavras consideradas sexistas também se mostra uma boa característica para a identificar o discurso, já características quantitativas são bons complementares, mas não são decisivos na classificação automática.

Para os três classificadores que escolhemos implementar, os melhores resultados foram obtidos pelo KNN, que obteve 0,99 de F1 utilizando POS tagging e também combinando o uso dos dois tipos de características. Em todos os modelos que implementamos, tivemos bons resultados utilizando apenas POS tagging, com 0,98 de F1 tanto para o SVM quanto para o RF. O modelo BERT implementado como baseline não apresentou resultados tão bons quanto os demais métodos, o que atribuímos ao tamanho da base de dados, que possui apenas 3.588 comentários. Para que algoritmos mais complexos tenham um desempenho melhor, a base de precisaria ser maior e ter mais conteúdo rotulado.

A classificação automática do sexismo se mostrou possível nesta base de dados utilizando características textuais, performando melhor nos casos em que analisamos a estrutura sintática do texto analisado.

6.2 Contribuições

Este trabalho gerou publicações do artigo *Criação e Caracterização de um Corpus de Discurso Sexistas em Português*, realizada na conferência BraSNAM (*BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING*) em 2020, publi-

cado também no journal ISYS - BRAZILIAN JOURNAL OF INFORMATION SYSTEMS em 2021. Na primeira publicação descrevemos a montagem da base de dados e a comparação entre o uso de alguns classificadores automáticos, enquanto na segunda incluímos o detalhamento do estudo das características relevantes para o discurso sexista para a classificação automática.

Apesar de termos exemplos de bases de dados de discurso de ódio em português, não encontramos bases focadas no discurso sexista ou misógino. Sendo assim, uma das contribuições deste trabalho é a disponibilização de uma base de 3.533 comentários em português, coletados de portais de notícias, que estão rotulados como sexistas ou não sexistas.

Além disso, o trabalho apresenta o estudo das características relevantes para a classificação desses comentários como sexistas ou não.

6.3 Próximos passos

A fim de tornar a base de dados mais representativa do discurso sexista, desejamos aumentar a quantidade de rótulos atribuídos a cada comentário, considerando idades mais diversas e buscando equilibrar os rótulos atribuídos entre os gêneros e idade. O aumento de rótulos atribuídos aos comentários, faz com que o viés entre o votos divergentes seja reduzido tornando a base de dados mais representativa de cada um dos discursos. Além disso, também é interessante aumentar a quantidade de comentários na base de dados para que possamos tentar reaplicar algoritmos como o BERT.

Comentários que estão inclusos na classe sexista ainda podem ser separados entre o sexismo hostil e o sexismo benevolente, é interessante analisarmos quanto da base de dados se enquadra em cada uma dessas subclasses, uma vez que o sexismo benevolente pode passar despercebido por muitas pessoas, tornando a classificação automática ainda mais importante nesses casos.

Comentários em portais de notícias apresentam contextos similares, é interessante aplicar a classificação automática utilizando as mesmas características em uma base de dados composta por conteúdos não relacionados diretamente, como tweets ou publicações no Facebook.

Referências Bibliográficas

- [1] Baeza-Yates, R. e Ribeiro-Neto, B. (2013). *Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora.
- [2] Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers Technology*, pgs. 233–239.
- [3] Braga, M. L. P., Nakamura, F. G., e Nakamura, E. F. (2020). Criação e caracterização de um corpus de discurso sexista em português. Em *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pgs. 97–107. SBC.
- [4] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [5] Dave, K., Lawrence, S., e Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Em *Proceedings of the 12th international conference on World Wide Web*, pgs. 519–528.
- [6] Davidson, T., Warmusley, D., e Macy, M. (2017). Automated hate speech detection and the problem of offensive language. *Eleventh International AAAI Conference on Web and Social Media*.
- [7] de Fátima Moreira, M. e da Silva, M. I. T. (2005). Legibilidade do material educativo escrito para pacientes diabéticos. *Online Brazilian Journal of Nursing*, 4(2):3–12.
- [8] de Pelle, R. P. e Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. Em *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.
- [9] Decode, I. A. . (2021). Muito alÉm do cyberbullying: A violÊncia real do mundo virtual.

- [10] Devlin, J., Chang, M.-W., Lee, K., e Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [11] DuBay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text*. ERIC.
- [12] Federal, G. (2021). Canais registram mais de 105 mil denúncias de violência contra mulher em 2020.
- [13] Fleiss, J. L. e Cohen, J. (1973). The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.
- [14] Froese, V. (1971). Cloze readability versus the dale-chall formula.
- [15] Glick, P. e Fiske, S. T. (2018). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. Em *Social Cognition*, pgs. 116–160. Routledge.
- [16] Gomes, H. S. (2015). Mulheres criam tecnologia para combater assédio sexual.
- [17] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., e Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.
- [18] Kramer, O. (2013). K-nearest neighbors. Em *Dimensionality reduction with unsupervised nearest neighbors*, pgs. 13–23. Springer.
- [19] Kwok, I. e Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. Em *Twenty-seventh AAAI conference on artificial intelligence*.
- [20] Martins, T. B. F., Ghiraldelo, C. M., Nunes, M. d. G. V., e Oliveira Junior, O. N. d. (1996). Readability formulas applied to textbooks in brazilian portuguese.
- [21] Mendes, L. (2015). Youtuber jout jout conta que ela e o feminismo se encaixam perfeitamente.
- [22] Mendes, R. S., VAZ, B. d. O., e Carvalho, A. F. (2015). O movimento feminista e a luta pelo empoderamento da mulher. *Gênero & Direito*, 3:88–99.
- [23] Nova/SB (2016). Dossiê da intolerância.
- [24] OGlobo (2019). Juliano cazarré rebate críticas de machismo em post com outra publicação considerada machista.

- [25] Oliveira, G. (2020). Streamer reckless, star of world of warcraft, dies at 31 years.
- [26] Oliveira, R. C. d., Lima, J. d. C. P., e Gomes, R. F. (2019). Machismo e discurso de Ódio nas redes sociais: uma análise das “opiniões” sobre a violência sexual contra as mulheres. *Revista Feminismos*, 6(1).
- [27] Oliveira, S. (2018). Adolescente vítima de bullying se suicida por ‘não aguentar mais’.
- [28] Paiva, P. D., da Silva, V. M., e Moura, R. S. (2019). Detecção automática de discurso de ódio em comentários online. Em *Anais da VII Escola Regional de Computação Aplicada à Saúde*, pgs. 157–162. SBC.
- [29] Pak, A. e Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Em *LREc*, volume 10, pgs. 1320–1326.
- [30] Pamungkas, E. W., Basile, V., e Patti, V. (2020). Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.
- [31] Pang, B., Lee, L., e Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. Em *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pgs. 79–86. Association for Computational Linguistics.
- [32] 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.
- [33] Rossi, M. (2014). Mulher espancada após boatos em rede social morre em guarujá, sp.
- [34] Udoh-Oshin, G. (2017). Hate speech on the internet: Crime or free speech?
- [35] Vianna, J. e Hising, E. (2018). Homem é condenado a 41 anos de prisão por crimes como racismo, terrorismo e divulgação de pedofilia na internet.
- [36] 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.

- [37] Wilcoxon, F. (1992). Individual comparisons by ranking methods. Em *Breakthroughs in statistics*, pgs. 196–202. Springer.
- [38] Wright, R. E. (1995). Logistic regression.