

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO**

**A PROBLEMÁTICA DE VIÉS DE
GÊNERO EM INTERAÇÕES COM
AGENTES VIRTUAIS: UM
ESTUDO BASEADO EM
APRENDIZADO DE MÁQUINA**

**KALISSA RODRIGUES
PAULO MONTEIRO**

Trabalho de Conclusão II apresentado
como requisito parcial à obtenção do
grau de Bacharel em Sistemas de
Informação na Pontifícia Universidade
Católica do Rio Grande do Sul.

Orientador: Prof. Silvia Moraes

**Porto Alegre
2021**

AGRADECIMENTOS

0.1 Agradecimentos Kalissa

Após uma longa jornada, sempre temos muito o que agradecer – vou tentar ser breve.

Inicio agradecendo quem tornou este sonho possível de ser sonhado: Guilherme, obrigada por, desde 2011, acreditar em mim mais do que eu mesma. Se eu pudesse dividir esse canudo com alguém, seria contigo. Meu amor e gratidão são imensuráveis.

Agradeço imensamente à minha família: meus avós, Rute e Dirnei, por serem minha razão para nunca desistir. Minha mãe, Tatiane, pela criação sem viés de gênero, por ser meu exemplo de força feminina e obstinação. Minha tia, Gabriela, pela disciplina. Meus tios, Tom e Patrícia, pelos exemplos de garra acadêmica. Meus gatos, Goku e Gohan, por estarem ao meu lado durante todo esse trabalho e graduação. Meu pai, Wagner, pelo incentivo para mudar de curso: “muda, faz o que for te fazer feliz”. E não é que fui feliz mesmo?

À minha rede de apoio, a família que escolhi pra mim: minha melhor amiga Ananda, por ser sempre minha maior motivadora e fã número 1. Meu amigo Juliano, por todas as trocas, acadêmicas ou não. Meus amigos Gabriel e João, fiéis escudeiros e parceiros de trabalho, de TCC, de vida: sem vocês essa monografia não teria sido possível. Vocês quatro são os maiores presentes que a capital me deu.

Aos colegas e amigos que fiz nesses anos de graduação, tanto na faculdade quanto nas empresas pelas quais passei: obrigada pelos incentivos, puxões de orelha, trocas de reclamações e risadas. Ao meu colega, Paulo, por dividir o peso de um trabalho de conclusão em meio à uma pandemia. A nossa professora orientadora, Silvia, pelas orientações e ensinamentos.

Por fim, agradeço às políticas públicas que me oportunizaram estudar em uma universidade como a PUCRS, e não obstante, agradeço a todas as mulheres inspiradoras que me trouxeram para a tecnologia, e nela me fizeram permanecer por acreditar em um propósito maior do que eu. Que possamos sempre ser inspiração umas para as outras. Concluo parafraseando Criolo:

"É, dizem que não é pra você, essa história de vencer, de sonhar e conquistar.

Eu digo que é pra você, essa história de vencer, de sonhar e conquistar."

0.2 **Agradecimentos Paulo**

Aos meus pais por me apoiarem e incentivarem nesta importante etapa em minha vida. Por me darem apoio sempre que precisei e nunca me deixarem desistir desse desafio que foi a minha graduação. Agradeço aos colegas e amigos que fiz durante este período de graduação, passamos bons momentos juntos e alguns momentos difíceis também. A minha colega Kalissa, companheira deste trabalho de conclusão de curso. E por fim não menos importante nossa excelentíssima professora orientadora, Silvia, obrigado por tudo que você me ensinou durante todo o período da graduação e pela orientação deste trabalho.

A PROBLEMÁTICA DE VIÉS DE GÊNERO EM INTERAÇÕES COM AGENTES VIRTUAIS: UM ESTUDO BASEADO EM APRENDIZADO DE MÁQUINA

RESUMO

Pensando em ter atratividade online, grandes marcas passaram a agregar agentes virtuais à sua publicidade, criando seus próprios influenciadores virtuais: é o caso das varejistas Magazine Luiza e Casas Bahias. Inicia-se, assim, a geração de um grande rastro de dados útil para os negócios composto pelas trocas de mensagens entre agentes virtuais (empresas) e usuários de redes sociais. Todavia, há sempre o ônus e o bônus: a agente virtual Magalu vem sofrendo, desde 2018, ataques virtuais em virtude de seu gênero (assédio e cibersexismo) [24]. Além dela, outras grandes marcas brasileiras relataram assédios a suas assistentes virtuais [14] [101], e já foram constatados diversos casos de viés de gênero [58] e violências sofridas pelo gênero feminino em contextos de tecnologia da informação [106]. Em consequência disto, elaboramos uma hipótese central: "Existe diferença no tratamento entre agentes virtuais de diferentes gêneros", e de forma a validar tal hipótese, construímos uma base de dados de respostas aos perfis destes dois agentes virtuais (Magalu e Baianinho), na rede social *Twitter*[8]. Com isto, utilizando análise de sentimentos através da API identificadora de comentários tóxicos da Google, *Perspective* [6], análise manual e visual, comparamos as respostas para a agente feminina e para o agente masculino, e analisamos o quanto tóxicas tais podem vir a ser: construímos um corpus com 1.007 mensagens com anotações de toxicidade em português, sendo duas delas não encontradas na literatura (Sexual Explícito e Assédio). A partir deste corpus, foram criados dois classificadores de texto utilizando o algoritmo de aprendizado de máquina kNN, com classificação de toxicidade em dois níveis (Toxicidade e Conotação Sexual Explícita). Com isto e através de análise dos dados, nossa hipótese foi validada.

Palavras-Chave: análise de gênero, análise de sentimentos, kNN, aprendizado de máquina, viés de gênero, assistentes virtuais, agentes virtuais.

THE PROBLEM OF GENDER BIAS IN INTERACTIONS WITH VIRTUAL AGENTS: A STUDY BASED IN MACHINE LEARNING

ABSTRACT

Thinking about being attractive online, big brands started to add virtual agents to their advertising, creating their own virtual influencers: such is the case of retailers Magazine Luiza and Casas Bahias. Thereby, the generation of a large trail of useful data for businesses begins, consisting of the exchange of messages between virtual agents (companies) and users of social networks. However, there is always the onus and the bonus: the virtual agent Magalu has been suffering, since 2018, virtual attacks due to her gender (harassment and cybersexism) [24]. In addition, other major Brazilian brands have reported harassment of their virtual assistants [14] [101], and several cases of gender bias have already been found [58] and violence suffered by females in information technology contexts [106]. As a result, we developed a central hypothesis: "Is there a difference in the treatment between virtual agents of different genders?", and in order to validate this hypothesis, we built a dataset of responses to the profiles of these two virtual agents (Magalu and Baianinho), on the social network *Twitter*[8]. With this, using sentiment analysis through Google's toxic comments identifier API, *Perspective* [6], manual and visual analysis, we compare the responses for the female agent and for the male agent, and analyze how much toxic they can be: we built a corpus with 1,007 messages with toxicity notes in Portuguese, two of which are not found in the literature (Explicit Sexual and Harassment). From this corpus, two text classifiers were created using the kNN machine learning algorithm, with toxicity classification in two levels (Toxicity and Explicit Sexual Connotation). With this and through data analysis, our hypothesis was validated.

Keywords: gender analysis, sentiment analysis, lexical approach, gender bias, virtual assistants.

LISTA DE FIGURAS

Figura 2.1 – Trajetória criação da agente Magalu	19
Figura 2.2 – Magalu nas redes sociais	19
Figura 2.3 – Trajetória criação do agente CB	20
Figura 3.1 – K vizinhos mais próximos	25
Figura 3.2 – Pseudocódigo kNN	26
Figura 5.1 – Exemplo das informações coletadas	36
Figura 5.2 – Total <i>tweets</i> coletados	37
Figura 5.3 – Total <i>tweets</i> coletados por agente e termos	38
Figura 5.4 – Total <i>tweets</i> únicos por agente e termos	39
Figura 5.5 – Total <i>tweets</i> pós limpeza completa: remoção de duplicados e irrelevantes	40
Figura 5.6 – Total <i>tweets</i> para análise manual, com proporção	41
Figura 5.7 – Anotação manual de <i>tweets</i> tóxicos	44
Figura 5.8 – <i>Tweets</i> analisados manualmente, primeira leva	44
Figura 5.9 – Análise <i>Perspective</i> sobre base proporcional, e anotações de toxicidade feitas	47
Figura 5.10 – Análise <i>Perspective</i> sobre base com corte de 74% + dicionário de cunho sexual, e anotações de toxicidade feitas	48
Figura 5.11 – <i>Tweets</i> analisados manualmente, segunda leva	49
Figura 5.12 – Pipeline de pré-processamento	49
Figura 5.13 – Nuvem de palavras do corpus antes do pré-processamento	52
Figura 5.14 – Nuvem de palavras do corpus após o pré-processamento	52
Figura 5.15 – Pipeline para criação dos classificadores	53
Figura 5.16 – Acurácia x K vizinhos - Classificador Tóxico	54
Figura 5.17 – Matriz de confusão para Classificador Tóxico	55
Figura 5.18 – Acurácia x K vizinhos - Classificador Sexual Explícito	56
Figura 5.19 – Matriz de confusão para Classificador Sexual Explícito	56
Figura 6.1 – Funil de <i>tweets</i> : coleta até anotações manuais	57
Figura 6.2 – Totais de <i>tweets</i> analisados, tóxicos e proporção por agente virtual ..	58
Figura 6.3 – Total de <i>tweets</i> marcados como tóxico, e demais atributos de toxicidade e devidas quantidades	58
Figura 6.4 – <i>Tweets</i> tóxicos por agente virtual, com atributos de toxicidade e devidas quantidades	59

Figura 6.5 – Fluxograma de construção e anotação do corpus.	61
Figura 6.6 – Total <i>tweets</i> analisados, e totais dos atributos de toxicidade anotados.	61
Figura 6.7 – <i>Tweets</i> tóxicos direcionados à agentes virtuais diversos	63

LISTA DE TABELAS

Tabela 3.1 – Representação dos tweets em um <i>Bag of Words</i>	26
Tabela 3.2 – Matriz de confusão para duas classes	27
Tabela 5.1 – Termos de pesquisa	36
Tabela 5.2 – Termos recorrentes e irrelevantes	40
Tabela 5.3 – Anotação disponível na ferramenta <i>Perspective</i>	42
Tabela 5.4 – Resultado comparativo das ferramentas	46
Tabela 5.5 – Medidas de Avaliação - Classificador Tóxico	55
Tabela 5.6 – Medidas de Avaliação - Classificador Sexual Explícito	55
Tabela 6.1 – Exemplos de tweets com diferentes tipos de toxicidade	60

LISTA DE SIGLAS

API – Application Programming Interface

IBGE – Instituto Brasileiro de Geografia e Estatística

BOW – Bag of Words

KNN – K-nearest neighbors ou K-vizinhos mais próximos

SUMÁRIO

0.1	AGRADECIMENTOS KALISSA	2
0.2	AGRADECIMENTOS PAULO	3
1	INTRODUÇÃO E CONTEXTO	13
2	VIÉS DE GÊNERO NO CONTEXTO DE AGENTES VIRTUAIS USANDO ANÁLISE DE SENTIMENTOS	16
2.1	VIÉS DE GÊNERO EM TECNOLOGIA DA INFORMAÇÃO	16
2.2	VIÉS E AGENTES VIRTUAIS	18
2.2.1	AGENTES VIRTUAIS, GÊNERO E DISCURSO DE ÓDIO	21
3	FUNDAMENTAÇÃO TEÓRICA	23
3.1	APRENDIZADO DE MÁQUINA	24
3.1.1	CLASSIFICADOR K-ÉSIMO VIZINHO MAIS PRÓXIMO (KNN)	25
3.1.2	REPRESENTAÇÃO DE ATRIBUTOS	26
3.1.3	AVALIAÇÃO DO CLASSIFICADOR	27
3.2	ANÁLISE DE SENTIMENTOS	28
3.2.1	TÓPICOS	29
3.2.2	ABORDAGEM	30
4	TRABALHOS RELACIONADOS	32
5	CLASSIFICADOR DE TOXIDADE	34
5.1	OBJETIVOS	34
5.1.1	OBJETIVOS ESPECÍFICOS E HIPÓTESES	34
5.2	AGENTES VIRTUAIS ANALISADOS	35
5.3	COLETA DOS DADOS	35
5.4	CONSTRUÇÃO DO CORPUS	36
5.4.1	ANOTAÇÃO DO CORPUS - PARTE 1	41
5.4.2	ANÁLISE DE SENTIMENTOS	45
5.4.3	ANOTAÇÃO DO CORPUS - PARTE 2	46
5.5	PRÉ-PROCESSAMENTO E ANÁLISE DO CORPUS	48
5.5.1	PRÉ-PROCESSAMENTO	48
5.5.2	ANÁLISE DO CORPUS	51

5.6	TREINAMENTO E VALIDAÇÃO.....	52
6	RESULTADOS E DISCUSSÕES	57
6.1	CORPUS ANOTADO	57
6.2	DISCUSSÃO	62
7	CONCLUSÃO DO ESTUDO E CONTRIBUIÇÕES	66
7.1	TRABALHOS FUTUROS	67
	REFERÊNCIAS	69

1. INTRODUÇÃO E CONTEXTO

Conforme os anos passam e a utilização da internet cresce, aumentam também os desafios. Segundo dados do IBGE de 2018, 79,1% dos domicílios brasileiros possuem acesso à internet [71]. Além disso, segundo levantamento da empresa de pesquisa GlobalWebIndex, no ano de 2020 o Brasil foi o segundo país no mundo onde as pessoas passaram mais tempo em redes sociais, com 225 minutos diários [11]. A plataforma *Twitter* é um exemplo do quanto popular as redes sociais se tornaram. Ela possui cerca de 186 milhões de usuários ativos por dia ao redor do mundo, totalizando, em média, aproximadamente 500 milhões de *tweets* postados por dia [81]. Levando em conta a população atual da Terra, que é por volta de 8 bilhões de pessoas, isso significa que 1 em cada 43 pessoas tem uma conta no *Twitter*.

Une-se, então, uma população cada vez mais conectada virtualmente, com o interesse crescente das empresas em ingressar e explorar nesse universo online. Em consequência disso, o marketing digital passa a focar cada vez mais em experiências em tempo real [100] e marketing de influência, onde as empresas deixam de ser vistas apenas como espaços físicos (lojas de varejo, por exemplo) e passam a representar pessoas e vozes, cujas quais influenciam pessoas.

Visando essa atratividade online, engajamento e, consequente, aumento nas vendas, o mercado de influenciadores digitais cresceu 1500% entre 2016 e 2020 [18]. E, conforme a computação gráfica avança, grandes empresas passaram a agregar, além de pessoas físicas, agentes virtuais à sua publicidade, usando estes para estampar sua marca e interagir com seu público, criando, assim, seus próprios influenciadores virtuais. Este é o caso da varejista Magazine Luiza, que em sua transformação digital, optou por investir em um agente virtual para lhe representar [9], criando assim a Magalu. A agente Magalu se tornou um case de sucesso, sendo a influenciadora virtual mais seguida do mundo em 2020 [23]. Além dela, as Casas Bahia também passou por um reposicionamento da marca, transformando seu personagem baianinho em CB, outro influenciador digital [62].

Desta forma, passa-se a ter marcas atuando em redes sociais como se fossem pessoas, postando fotos e respondendo comentários. Inicia-se, assim, a geração de um grande rastro de dados completamente novo composto por essas trocas de mensagens entre agentes virtuais (empresas) e usuários de redes sociais. E, embora tais informações sejam valiosas para os negócios [79][30], há sempre o ônus e o bônus: a agente virtual Magalu vem sofrendo, desde 2018, ataques virtuais em virtude de seu gênero (assédio e cibersexismo) [24]. Além dela, outras grandes marcas brasileiras relataram assédios a suas agentes virtuais [14] [101], e já foram constatados na literatura diversos casos de viés de gênero [58] e violências sofridas pelo gênero feminino em contextos de tecnologia da informação [106].

Tendo em vista tal problemática, elaboramos uma hipótese central que conduziu o presente estudo: **Existe diferença no tratamento entre agentes virtuais de diferentes gêneros.** Desta hipótese discorreram sub-hipóteses, como:

1. Agentes virtuais femininos são mais xingados.
2. Xingamentos a agentes virtuais femininos tem uma maior conotação sexual.
3. Os elogios aos agentes virtuais falam sobre sua aparência, tornando-se assédio.
4. Xingamentos para agentes virtuais masculinos são mais agressivos.

Por conseguinte, a fim de validar tais hipóteses e trazer à tona a toxicidade nas comunicações na web, captamos e construímos uma base de dados de respostas aos perfis de dois agentes virtuais de grandes marcas Magalu e Baianinho (CB), a partir da rede social *Twitter* [8]. Desta forma, buscamos analisar questões de viés de gênero relacionadas aos agentes virtuais, analisando a respostas para a agente feminina e para o agente masculino, observando o quanto tóxicas tais poderiam vir a ser.

Inicialmente, testamos três ferramentas de Análise de Sentimento (biblioteca de código aberto chamada *Detoxify* [50], serviço de cognição da Microsoft *Cognitive* [4], e API identificadora de comentários tóxicos da Google, *Perspective* [6]), a fim de termos uma medição quantitativa quanto à toxicidade, sendo toxicidade, pela definição do *Perspective*, “um comentário rude, desrespeitoso ou irracional que provavelmente o fará sair de uma discussão”. Avaliamos o desempenho das três ferramentas, de forma qualitativa, e julgamos a análise da que se saiu melhor (teve mais marcações corretas de toxicidade - *Perspective*). A partir dessa análise, construímos um corpus com 1.007 mensagens em português com anotações de toxicidade.

Considerando o corpus construído como base deste estudo, durante a análise dos *tweets* percebemos que a primeira hipótese de que interações preconceituosas são mais frequentes com personagens femininos é verdadeira.

Posteriormente, continuamos o nosso estudo usando o corpus para treinar classificadores. Foram definidos dois classificadores: um primeiro que categoriza os *tweets* em Tóxicos e Não Tóxicos, e um segundo classificador que categoriza os *tweets* tóxicos (anotados manualmente) em Conotação Sexual Explícita e Assédio. Em ambos os classificadores, utilizamos o algoritmo K-vizinhos mais próximos (kNN). O primeiro classificador obteve 78% de acurácia na classificação de toxicidade, e, o segundo 77% na classificação de Sexual Explícito ou Assédio.

Por fim, de forma a aprofundar o estudo sobre a problemática de viés de gênero no contexto de agentes virtuais, utilizamos de análise visual da base gerada e também do corpus através da ferramenta *Power Bi* da Microsoft [5]. Buscamos trazer mais clareza sobre as causas e consequências que a reprodução de vieses e preconceitos em tecnologia

podem acarretar na sociedade, à medida que ferramentas tecnológicas assumem maiores capacidades de comunicação semelhantes às humanas. Ao longo desse estudo, encontramos indícios no corpus que demais hipóteses também são verdadeiras.

Este trabalho está organizado em 7 capítulos. No capítulo 2 é comentado sobre **Viés de Gênero no Contexto de agentes virtuais utilizando Análise de Sentimentos**. No terceiro capítulo discorre sobre a **Análise de Sentimentos**. No capítulo 4 temos os **Trabalhos Relacionados**, já no capítulo 5 é discorrido sobre o **Classificador de toxicidade**: criação do corpus, elaboração do modelo e utilização. Para finalizar, no capítulo 6 trazemos os **Resultados e Discussões**, e no último capítulo são apresentadas as **Conclusões do Estudo e Contribuições**.

2. VIÉS DE GÊNERO NO CONTEXTO DE AGENTES VIRTUAIS USANDO ANÁLISE DE SENTIMENTOS

A tecnologia está cada vez mais presente em nossos dias. Conforme o estudo do App Annie, até o segundo quarto de 2020 a população mundial já havia passado 1,6 trilhões de horas em frente às telas de telefones [57]. Todavia, da mesma forma como qualquer conjunto de conhecimentos, a tecnologia também pode ser usada tanto para o bem quanto para o mal: discurso de ódio em forma de comentários racistas e sexistas são uma ocorrência comum em redes sociais [104].

A partir disto, nesse capítulo daremos foco à questões relacionadas a gênero, e também analisaremos como o gênero de agentes virtuais pode ser causa ou consequência de vieses maldosos, e como as pessoas destes robôs podem vir a sofrer preconceito de gênero e cibersexismo.

2.1 Viés de Gênero em Tecnologia da Informação

De forma a contextualizar o que é conhecido como viés de gênero sob a alcada de tecnologia, propomos um exercício: imagine uma figura de liderança, com trajes formais, expondo dados em um telão para uma mesa com várias pessoas. Essa figura que aparece na sua imaginação é masculina ou feminina? Se você respondeu masculina, saiba que, além de ser a tendência de resposta, isso é devido a um viés inconsciente. De acordo com Renee Navarro da Universidade da Califórnia:

Vieses são qualquer forma de parcialidade a favor ou contra algo, alguém ou algum grupo quando comparado em relação a outro, de uma maneira que é considerada injusta, sendo que tais podem ser assumidos por um indivíduo, grupo ou instituição. Esses vieses são característicos e inatos na natureza humana, resultantes de um processo evolutivo que favorece a busca por padrões quando observamos a natureza e tomadas de decisão rápidas baseadas nesses padrões. Eles são formados ou aprendidos por meio de diversos processos e experiências, como relações afetivas com outros seres humanos, associações implícitas de determinadas características como perigo ou até mesmo por influência genética [70].

Existe uma subdivisão quando falamos em viés: os vieses conscientes e os inconscientes. O preconceito implícito se refere às atitudes ou estereótipos que afetam nosso entendimento, ações e decisões de maneira inconsciente. Esses preconceitos, que abran-

gem avaliações favoráveis e desfavoráveis, são ativados involuntariamente e sem a consciência ou controle intencional de um indivíduo [26].

O preconceito inconsciente é muito mais prevalente do que o preconceito consciente e muitas vezes incompatível com os valores conscientes. Certos cenários podem ativar atitudes e crenças inconscientes. Por exemplo, os vieses podem ser mais prevalentes ao realizar várias tarefas ou trabalhar sob pressão de tempo. É importante notar que vieses, conscientes ou inconscientes, não se limitam a etnia e raça. Embora o preconceito racial e a discriminação sejam bem documentados, podem existir vieses em relação a qualquer grupo social. Idade, gênero, habilidades físicas, religião, orientação sexual, peso e muitas outras características estão sujeitos a preconceito [70].

Quando falamos em vieses em tecnologia da informação, grande parte das pesquisas tem endereçado seus esforços ao impacto do preconceito racial inconsciente, como análises sobre como o software de reconhecimento facial incorporado na maioria dos smartphones funciona melhor para quem é branco e do gênero masculino [31]. No entanto, Kate Crawford sumarizou a problemática de viés de gênero em Inteligência Artificial da seguinte forma:

“Como todas as tecnologias anteriores, a inteligência artificial refletirá os valores de seus criadores” [38].

Desta forma, devemos refletir sobre por quem este tipo de tecnologia está sendo criada: *The International Telecommunication Union* (ITU) estima que apenas 6% dos profissionais de desenvolvimento de software são mulheres [106]. Segundo a coalizão EQUALS da ONU, existe ciência de que uma maior participação feminina em empresas de tecnologia não garante que o hardware e software que essas empresas produzem sejam sensíveis ao gênero [106]. No entanto, essa ausência de garantia não deve obscurecer as evidências de que equipes de tecnologia com maior igualdade de gênero estão, em geral, mais bem posicionadas para criar tecnologia com maior igualdade de gênero [76] e também pode ser mais lucrativo e inovador [93].

Assim sendo, nosso estudo sobre viés de gênero dará foco sobre comunicações textuais, em concordância com o que pontua o Susan Levy,

“os valores sociais tendenciosos contra as mulheres podem estar profundamente incorporados na forma como a linguagem é usada, e impedir que algoritmos de aprendizado de máquina treinados em texto perpetuem o preconceito requer uma compreensão de como a ideologia de gênero se manifesta na linguagem [58].”

2.2 Viés e Agentes Virtuais

Nosso cotidiano vem sendo reinventado dia após dia, com a inserção de novas ferramentas tecnológicas, que servem como facilitadores. Alguns desses facilitadores em voga na atualidade são os *chatbots*, agentes virtuais e assistentes virtuais: uma pesquisa feita pela empresa de ciência de dados Ilumeo registrou um crescimento de 47% do uso de serviços ou produtos com assistentes virtuais por voz no Brasil em 2020 [87].

Para entendermos os papéis que assistentes virtuais podem desempenhar, usaremos as definições propostas pela UNESCO [106]:

- **Assistentes de voz:** tecnologia que fala aos usuários por meio de saídas de voz, mas normalmente não projeta uma forma física. Os assistentes de voz geralmente podem entender entradas faladas e escritas, mas geralmente são projetados para interação falada. Seus resultados normalmente tentam imitar a fala humana natural.
- **Chatbots:** tecnologia que interage com os usuários principalmente por meio da linguagem escrita. Os *chatbots* podem ou não projetar uma forma física. Nos casos em que uma forma física é projetada, normalmente é estática - muitas vezes uma imagem estática de um rosto humano ou às vezes uma imagem não humana, como um personagem de desenho animado. Os *chatbots* são diferentes dos assistentes de voz porque sua saída geralmente é um texto escrito, não palavras faladas.
- **Agentes Virtuais:** tecnologia que se comunica com os usuários por meio da fala e projeta uma forma física virtual, geralmente uma projeção humana ou às vezes não humana, como um animal de desenho animado.

Todavia, com o estado da arte de tecnologias atuais, quando pensamos em um *bot* imaginamos um resolvedor de problemas: seja um *chatbots* que irá nos mandar uma mensagem para lembrarmos de pagar as contas em dia, seja um assistente de voz cujo qual, ao receber uma ordem falada, faz as janelas de nossas casas se fecharem ao pôr do sol. O que nós não sabíamos antes, quando este tipo de tecnologia não era presente em nosso dia a dia, é que tais não só aprenderiam com nossos passos, como se tornariam personas de grandes marcas.

Esse é o caso da gigante do varejo Magazine Luiza: a garota propaganda da marca se tornou uma agente virtual, sendo ela a responsável por dar rosto a marca, mas também por responder a clientes em redes sociais. É a voz de propagandas e de serviços prestados pela empresa, sendo a especialista digital desta. Quando você desejar falar com a loja Magazine Luiza, será necessário entrar em contato com a Magalu primeiro, mesmo que esta seja uma pessoa fictícia [35].



Figura 2.1 – Trajetória criação da agente Magalu



Figura 2.2 – Magalu nas redes sociais

Assim sendo, a assistente virtual da Magalu vai além de ser um agente conversacional, e passa a ser uma agente virtual, como demonstrado na Figura 2.1 e na Figura 2.2: ela tem uma rotina, cuida da casa, viaja e posta fotos no Instagram. Assim, as agentes virtuais geram empatia em seus seguidores, e muito engajamento. De acordo com o relatório do HypeAuditor [22], os influenciadores virtuais têm quase três vezes mais engajamento do que os influenciadores reais. Isso significa que os seguidores estão mais engajados com o conteúdo dos influenciadores virtuais.

Da mesma forma, em 2020 outra varejista, as Casas Bahia, também passou por uma reformulação, repensando sua identidade digital. Em consequência disso, alterou sua mascote, onde a criança baiana cresce e dá lugar ao adolescente "CB". De acordo com Ilca Sierra, diretora de Marketing e Comunicação Multicanal da marca:

“Estamos entrando em um segundo e importantíssimo capítulo nessa evolução do reposicionamento da marca. Ele (CB) vem agora como um influenciador digital, para dar mais luz aos nossos valores no novo posicionamento, sendo guardião de temas como sustentabilidade e diversidade [62].”



Figura 2.3 – Trajetória criação do agente CB

Neste trabalho, tal qual no estudo desenvolvido pela UNESCO [106], optamos por voltar o olhar para assistentes digitais pois tais são:

1. amplamente usados globalmente;
2. raramente examinados por lentes de gênero; e
3. raramente percebido por agências governamentais e organizações internacionais que trabalham para construir sociedades e sistemas educacionais com maior igualdade de gênero.

Dessa forma, quando falamos em vieses no contexto de agentes virtuais, alguns trabalhos identificaram as maneiras pelas quais a ideologia de gênero está incorporada na linguagem e como isso pode influenciar as concepções das pessoas sobre as mulheres e as expectativas de comportamento associadas ao gênero. Cientistas sociais naturalmente se preocupam para tentar identificar essas características em um esforço para compreender as origens de preconceito e discriminação [97], todavia, em tecnologia da informação, ainda vemos poucos estudos sobre o assunto.

Uma área que tem feito esses estudos é a de aprendizado de máquina, tendo em vista que essas ideologias de gênero ainda estão embutidas em fontes de texto e resultam em algoritmos de aprendizado de máquina que aprendem conceitos estereotipados de gênero [28].

Quando fala-se em assistentes de voz, o gênero que sua voz ou agente carrega pode ser enviesado [52], mas também pode-se ter vieses na forma como as pessoas interagem com esses assistentes. É o caso da agente virtual Magalu, que sofreu ataques virtuais em virtude de seu gênero (assédio e cibersexismo) [24].

Assim, dada a existência de claras diferenças de gênero nos estilos de comunicação nas redes sociais [33], inclusive ao expressar sentimento [67], neste estudo abordaremos o tratamento que se dá a agentes virtuais, analisando dados de dois agentes virtuais, de forma a comparar as reações das pessoas que interagem com eles. O objetivo é identificar os sentimentos e a relação destes com o gênero desses agentes virtuais.

2.2.1 Agentes virtuais, Gênero e Discurso de Ódio

Quando falamos em tecnologia e ferramentas, tais temas não parecem convergir com a temática de gênero à primeira vista. Todavia, quando percebemos que a tecnologia é usada por todos mas decidida por alguns, e que tal disparidade é prejudicial para os negócios e para a sociedade como um todo, nossa percepção é aumentada: a produção de tecno-narrativas e as práticas culturais em torno das tecnologias mostram que as tecnologias não são naturalmente masculinas ou femininas, mas sim, as ideias e visões sobre masculinidade e feminilidade inscritas nelas [72].

Glick e Fiske [48] introduzem a noção de sexismo benevolente, no qual as mulheres são percebidas com características positivas, como prestativas. Apesar de sua natureza aparentemente positiva, o sexismo benevolente pode ser prejudicial, insultuoso e discriminatório. Em termos de palavras em aprendizado de máquina, as associações de gênero feminino com qualquer palavra, mesmo uma palavra subjetivamente positiva como "atraente", podem causar discriminação contra as mulheres se isso reduz sua associação com outras palavras, como "profissional". Ou seja: encaixam o gênero feminino como "a atrrente", mas somente ligam "um bom profissional" a flexão de gênero masculino.

O sexismo benevolente mostra-se presente na criação de personagens de *bots*, tendo em vista que tais são, em sua maioria, mulheres ajudantes. A Amazon tem Alexa (assim chamada em homenagem à antiga biblioteca de Alexandria), a Microsoft tem Cortana (nomeado para um sintético inteligência no videogame Halo que se projeta como uma sensual mulher nua), e a Apple tem Siri (cunhado pelo co-criador norueguês do iPhone 4S e significa "bela mulher que o leva à vitória" em nórdico). Enquanto o assistente de voz do Google é simplesmente Google Assistente e, às vezes, conhecido como Google Home, sua voz é inconfundivelmente feminina [106].

Tais assistentes tem sempre respostas gentis e servis, o que acaba por reforçar esterótipos de gênero: pesquisadores especializados em interação humano-computador há muito reconheceram que tanto homens quanto mulheres tendem a caracterizar as vozes

femininas como mais úteis, embora as razões por trás dessa observação não sejam claras [69]. Ecoando estereótipos de gênero, Nass também descobriu que as pessoas tendem a perceber as vozes femininas como ajudantes ou assistentes que nos ajudam a resolver nossos problemas, enquanto as vozes masculinas são vistas como figuras de autoridade que nos dão as respostas para nossos problemas. Além disso, ele também descobriu que a fala das mulheres inclui mais pronomes pessoais (eu, você, ela), enquanto a dos homens usa mais quantificadores (um, dois, alguns mais).

Todavia, a pesquisa de Nass data do início de 2000, e normas sociais tendem a mudar com o tempo. E já temos alguns estudos que questionam essas normas: F. Habler et.al, em sua pesquisa, conduziram um experimento controlado com 24 entrevistados, utilizando análise de medidas repetidas de duas vias de covariância (ANCOVA) com o gênero dos participantes como uma covariável para determinar os efeitos da linguagem e da voz. Tais obtiveram como resultados que a forma de comunicação "*low-status*", que emprega mais pronomes pessoais e expressões mais amistosas (por exemplo, a frase "De qual gênero a música deve ser?", em versão "*low-status*" é "De qual gênero posso recomendar uma música para você?"), consistentemente obteve avaliações mais positivas, e o efeito deste tipo de linguagem foi maior do que o efeito da voz (quando masculina X feminina) [49]. Também Andreea Danilescu abordou o tema, e trouxe em seu estudo importantes questionamentos: por que o progresso na igualdade de gênero não se reflete em assistentes de voz e robôs? A exposição prolongada a assistentes de voz que não seguem as normas de gênero mudaria a percepção do usuário ao longo do tempo? [40]

Indo ao encontro dessas questões, neste estudo, pretendemos discutir o viés de gênero analisando o tratamento dos usuários, durante interações textuais (tweets), dado a agentes virtuais. Por fim, neste estudo comprovamos a hipótese de que pessoas femininas tendem a obter respostas enviesadas e inclusive mais xingamentos do que pessoas masculinas, o que caracteriza discurso de ódio. Na legislação nacional e internacional, o discurso de ódio se refere a expressões que defendem o incitamento ao dano (particularmente, discriminação, hostilidade ou violência) com base na identificação do alvo com um determinado grupo social ou demográfico. Pode incluir, mas não se limita a, discurso que defende, ameaça ou incentiva atos violentos. Para alguns, no entanto, o conceito se estende também a expressões que fomentam um clima de preconceito e intolerância, no pressuposto de que isso pode alimentar a discriminação dirigida, a hostilidade e os ataques violentos [47].

3. FUNDAMENTAÇÃO TEÓRICA

Com a crescente presença da tecnologia na vida cotidiana, o uso de redes sociais tornou-se um hábito diário na vida da maioria das pessoas. A partir disso, as redes sociais se tornaram um vasto lugar de informações disponibilizadas abertamente na Web, e consequentemente tem sido foco de diferentes tipos de estudos, de forma a analisar as interações e os comportamentos das pessoas nestes meios digitais [32, 33, 39, 91, 88].

Atualmente, quando alguém necessita comprar algum produto ou serviço, não está limitado apenas a consultar amigos ou familiares para saber a opinião deles, pois existem inúmeros comentários, discussões em fóruns, *posts* (*post* é uma mensagem ou conteúdo publicado numa rede social [59]) nas já citadas redes sociais, enfim, diversas formas de se obter informações sobre o produto ou serviço. Para as organizações, esse crescimento no número de usuários fez com que elas não precisem mais despender recursos e esforços executando grandes pesquisas de mercado e de satisfação para levantar dados referentes a opinião pública, já que essas informações estão disponíveis abertamente e de forma abundante nas redes sociais e na *Web* [90].

Tendo em vista essas redes e a grande quantidade de dados que os usuários compartilham nelas, a tarefa de classificar sentimentos nas interações dos usuários passou a ser objeto de estudo [91]: com a riqueza de informações disponíveis, as organizações começaram a olhar com mais atenção o conteúdo que está nessas redes sociais a fim de embasar os processos de tomada de decisão.

A partir deste grande número de informações, é possível utilizar diferentes tecnologias para transformar informação em conhecimento. Para isso, escolher qual a melhor tecnologia a se utilizar depende do objetivo final que se deseja: quando há um objetivo de negócio, pode-se usar as técnicas de Aprendizado de Máquina para recomendação de produtos, entendimento da jornada dos clientes e direcionamento do marketing digital. Academicamente falando, pode-se utilizar Análise de Sentimentos tanto para avaliação de questões psicológicas, quanto para análise de questões de viés e preconceitos. Para o presente trabalho, empregaremos a técnica de análise de sentimentos a fim de estudar o comportamento das pessoas em relação a marcas, seus agentes virtuais e se o gênero desses agentes podem ser causa ou consequência de algum comportamento ofensivo.

O presente capítulo fornece conceitos essenciais para conduzir este trabalho. A Seção 3.1 apresenta brevemente o que é aprendizado de máquina e apresenta de forma sucinta sobre a tarefa de classificação. Na Seção 3.2 é apresentado o algoritmo de classificação kNN, que será utilizado no presente trabalho. A Seção 3.3 apresenta a forma de transformação de textos escritos em linguagem natural para uma estrutura que possa ser interpretada pelo algoritmo. A Seção 3.4 é apresentado o conceito de avaliação de classificadores apresentando a técnica de matriz de confusão e a métrica de avaliação utilizada

neste trabalho. Por fim, a Seção 3.5 detalha sobre análise de sentimentos e seu funcionamento.

3.1 Aprendizado de Máquina

A área de aprendizado de máquina tem como objetivos desenvolver técnicas computacionais que permitam simular o processo de aprendizado e construção de sistemas capazes de adquirir conhecimento automaticamente [65]. Nas últimas décadas, está presente em quase qualquer tarefa que requer extração de informações de grandes conjuntos de dados. Nós estamos cercados por tecnologias baseadas em aprendizado de máquina: algoritmos que melhoraram mecanismos de busca, anti-spam, recomendações, entre outras tarefas.[86].

Sistemas de aprendizado de máquina podem ser classificados de duas maneiras principais: supervisionado e não supervisionado [27]. Na abordagem de aprendizado supervisionado, os algoritmos são capazes de realizar classificações baseadas em experiências passadas. Na fase de treino, o algoritmo aprende os padrões existentes nos dados. Este treinamento exige conjuntos de exemplos contendo entradas e saídas esperadas, tendo em vista que o algoritmo gera o seu conhecimento a partir desses exemplos.

Diferentemente do aprendizado supervisionado, o não supervisionado não se utiliza de referências, ou seja, não ocorre um treinamento com o conhecimento prévio das saídas esperadas. O aprendizado não supervisionado é realizado quando, para cada exemplo, apenas os atributos de entrada estão disponíveis. Esse aprendizado é utilizado quando se tem o objetivo de encontrar padrões em um conjunto de dados que não foram classificados previamente [102].

Problemas que envolvem aprendizagem de máquina podem ser divididos em diversas tarefas [86], no presente trabalho, o foco está na tarefa de classificação. A classificação envolve prever uma categoria, por exemplo, se um paciente está ou não doente, se um comentário é negativo ou não. Para este tipo de tarefa é desenvolvido um algoritmo que terá acesso a exemplos de dados previamente classificados e, com base nesses exemplos, deverá produzir um classificador que pode ter como entrada um novo documento e gerar uma classificação para ele [86]. No presente trabalho foi utilizado o algoritmo de aprendizagem de máquina supervisionado k-ésimo Vizinho mais Próximo (kNN), o qual é descrito na próxima seção.

3.1.1 Classificador k-ésimo Vizinho mais Próximo (kNN)

A classificação baseada em vizinhos é um tipo de aprendizagem baseada em instância ou aprendizado preguiçoso (*Lazy Learning*) e não paramétrico [107], isso deve-se ao fato de que o kNN posterga a maior parte do processamento para o momento da predição pois ele não gera um modelo, mas simplesmente armazena as instâncias dos dados de treinamento. Com isso, ele reterá todo ou grande parte do conjunto de treino para realizar a etapa de teste, sendo o kNN uma variação do algoritmo *Nearest Neighbor* (NN) proposto por Cover e Hart [37].

O princípio do algoritmo é que, se a maioria das k amostras mais semelhantes a um ponto de consulta q_i no espaço da amostra pertencem a uma determinada categoria, então pode se dizer que o ponto q_i pertence a esta categoria [43]. Para classificar um registro desconhecido a distância entre outros registros de treinamento são computados. Com base na distância do K , os vizinhos mais próximos são identificados e os rótulos de classe destes vizinhos mais próximos são usados para determinar o rótulo da classe de registro desconhecido [43]. A Figura 3.2 representa a escolha dos K vizinhos mais próximos, já o funcionamento do algoritmo é exemplificado na Figura 3.2.

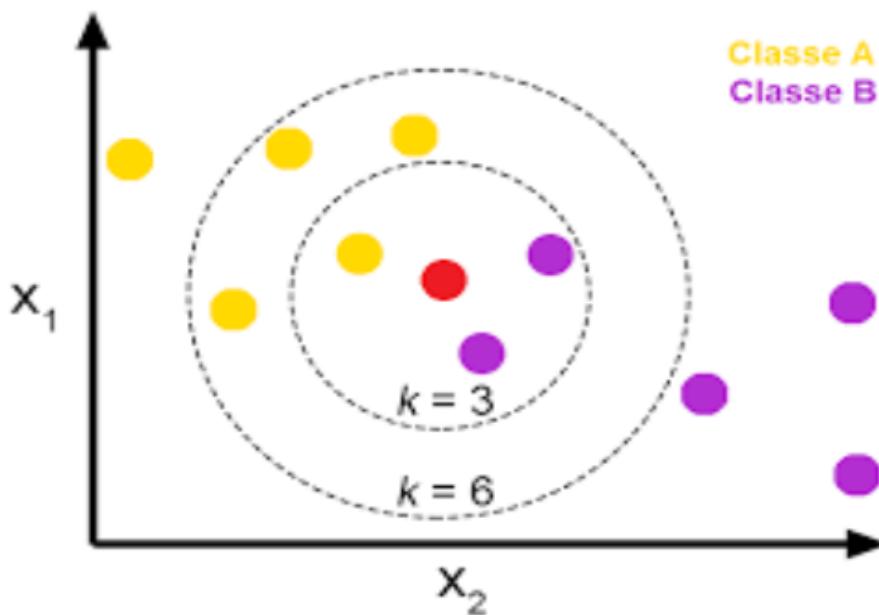


Figura 3.1 – K vizinhos mais próximos

No que tange a distância entre dois pontos no espaço multidimensional, tal pode ser definida de várias maneiras [99]. A distância Euclidiana é geralmente a mais usada. A distância Euclidiana é definida como a soma da raiz quadrada da diferença entre x e y em suas respectivas dimensões conforme demonstrado a equação abaixo. Este foi o tipo de distância utilizada no classificador utilizado no presente trabalho.

```

1 inicialização:
2     Preparar conjunto de dados de entrada e saída
3     Informar o valor de  $k$ ;
4 para cada nova amostra faça
5     Calcular distância para todas as amostras
6     Determinar o conjunto das  $k$ 's distâncias mais próximas
7     O rótulo com mais representantes no conjunto dos  $k$ 's
8     vizinhos será o escolhido
9 fim para
10 retornar: conjunto de rótulos de classificação

```

Figura 3.2 – Pseudocódigo kNN

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (3.1)$$

3.1.2 Representação de Atributos

Para os classificadores poderem interpretar as sentenças, elas devem passar por uma etapa de transformação e é necessário representá-las de uma forma estruturada como, por exemplo, uma matriz de termos. Esta matriz tem por objetivo preparar os dados para a etapa de treinamento e depois validação, abordagem conhecida como saco de palavras (do inglês, *bag-of-words (BoW)*), matriz a qual foi utilizada no presente trabalho. Essa matriz de termos consiste em estruturar as mensagens de maneira a torná-las processáveis pelos algoritmos de aprendizagem de máquina, sendo cada *tweet* um vetor em um espaço multidimensional. Como a Tabela 3.1 ilustra, as linhas representam os *tweet* e as colunas representam as palavras ou termos presentes nas mensagens, e os valores associados às colunas representam a frequência ou presença desses termos no *tweet* [89].

	t_1	t_2	...	t_m
$tweet_1$	a_{1_1}	a_{1_2}	...	a_{1_m}
$tweet_2$	a_{2_1}	a_{2_2}	...	a_{2_m}
...
$tweet_n$	a_{n_1}	a_{n_2}	...	a_{n_m}

Tabela 3.1 – Representação dos tweets em um *Bag of Words*

A Tabela 3.1 representa n *tweets* e m termos, sendo cada $tweet_i = (a_{i_1}, a_{i_2}, \dots, a_{i_m})$. No qual o valor a_{ij} refere-se ao valor associado ao j -ésimo termo do *tweet* i , ou seja, a_{ij} é o valor do termo t_j no $tweet_i$ e pode ser calculado utilizando diferentes medidas. Para o pre-

sente trabalho foi escolhido a medida binária, onde se o termo está presente no documento o valor de a_{ij} é 1, e se o termo é ausente assume o valor 0 [84].

3.1.3 Avaliação do Classificador

A etapa de avaliação consiste em estimar quanto os classificadores conseguem classificar corretamente as novas amostras. Uma técnica simples é obter um conjunto de dados já classificados e dividi-lo em dois subconjuntos, uma para treinamento e outro para teste. O treinamento é realizado com o primeiro conjunto e o segundo é utilizado na etapa de generalização. É na etapa de generalização que o modelo e parâmetros definidos são testados. A partir do teste é possível medir a taxa de acerto do classificador, dado que sabe-se as classes corretas dos dados. Por exemplo, se no conjunto de teste existem 50 amostras e o classificador classificou corretamente 35 delas, esse classificador obteve uma taxa de acerto de 70% .

Para as tarefas de classificação que envolvem classes de valores discretos, algumas medidas de qualidade podem ser estimadas a partir dos seguintes resultados:

VP (verdadeiro positivo): Número de exemplos positivos que foram classificados de forma correta.

FP (falso positivo): Número de exemplos negativos que foram classificados como positivos.

FN (falso negativo): Número de exemplos positivos classificados como negativos.

VN (verdadeiro negativo): Número de exemplos negativos que foram classificados de forma correta.

Com base nestas medidas cria-se uma matriz de confusão, que é uma forma de visualização dos resultados dos classificadores. Cada linha da matriz representa as instâncias preditas de uma determinada classe, enquanto cada coluna da matriz representa as instâncias reais de uma classe. Sendo assim, a diagonal principal representa os valores que foram preditos de forma correta e a diagonal secundária representa os valores que foram preditos incorretamente [45]. Uma matriz de confusão é ilustrada na tabela 3.2

Classes	Positivo Real	Negativo Real
Positivo Predito	VP	FP
Negativo Predito	FN	VN

Tabela 3.2 – Matriz de confusão para duas classes

A partir disso, é possível calcular medidas de avaliação estatística, tais como acurácia, precisão, revocação (*recall*) e medida F (*F1 score*) [78].

A acurácia (Equação 3.2) é uma medida qualitativa responsável por computar a proporção de classificação corretas [82].

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \quad (3.2)$$

A precisão (Equação 3.3) é a proporção das instâncias que foram classificadas como corretas, ou seja, quantos efetivamente eram corretas.

$$Precisao = \frac{VP}{VP + FP} \quad (3.3)$$

A revocação ou *recall* (Equação 3.4) é a proporção das instâncias que foram classificadas corretamente como classes positivas.

$$Revocacao = \frac{VP}{VP + FN} \quad (3.4)$$

A medida F ou *F₁ score* (Equação 3.5) é a média harmônica entre a precisão e a revocação. As medidas precisão e revocação quando analisadas separadamente podem trazer uma falsa impressão de qualidade ao classificador, já que uma precisão elevada geralmente significa que a revocação terá seu valor decrescido e vice-versa [46]. Quando estas medidas são utilizadas em conjunto para obter o mesmo valor, este valor é chamado de ponto de equilíbrio do sistema [46].

$$F_1 \text{ score} = 2 \cdot \frac{Precisao \cdot Revocacao}{Precisao + Revocacao} \quad (3.5)$$

3.2 Análise de Sentimentos

A análise de sentimentos une conceitos de aprendizado de máquina e processamento de linguagem natural. Também conhecida como mineração de opinião, é um campo de estudo que analisa opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a produtos, serviços, organizações, pessoas, problemas, eventos, tópicos e seus atributos. Ela representa uma grande parte do campo de estudo de Processamento de Linguagem Natural, conhecida também como PLN, e é útil quando se deseja identificar se uma determinada opinião em texto é positiva ou negativa [61] em meio a um vasto conjunto de informações.

Uma possível aplicação para a técnica de análise de sentimento é usar o seu resultado como uma pista sobre quais são os comentários que contém tons negativos ou pejorativos em relação a um serviço prestado, quando aplicada para um canal de *feedback* de uma empresa na sua página na Web ou nas redes sociais, por exemplo. Nos últimos anos, houve uma crescente no número de *post* nas redes sociais com opiniões sobre inúmeros

tópicos que estão sendo usados para que empresas repensem seus posicionamentos de mercados perante o seu cliente [61].

3.2.1 Tópicos

Segundo Bhuiyan [25], a análise de sentimentos pode ser aplicada em:

- **Nível de documento:** A tarefa nesse nível é determinar qual a opinião expressada em um documento, analisando se tal foi positiva ou negativa [73]. Por exemplo, dado uma avaliação de um produto, o sistema determina se a avaliação como um todo é positiva ou negativa. Essa tarefa é comumente conhecida como classificação de sentimento em nível de documento. Neste nível de análise, assume-se que cada documento expressa a opinião de uma única entidade (exemplo: um produto). Assim, não é aplicável a documentos que avaliam ou comparam múltiplas entidades.
- **Nível de sentença:** A tarefa nesse nível é determinar qual a opinião expressada na sentença, analisando se tal foi positiva, negativa ou neutra [61]. Neutra normalmente significa sem opinião.

Esse nível de análise está relacionado com classificação de subjetividade [108], que distingue sentenças objetivas, as quais expressam informações factuais, de sentenças subjetivas, que expressam visões e opiniões subjetivas. Entretanto, devemos observar que subjetividade não é equivalente a sentimento, já que muitas sentenças objetivas podem implicar em opiniões [61].

- **Nível de entidade e aspecto:** Tanto a análise de documento quanto a de sentença não fazem a extração do que a pessoa gostou ou não. Já tarefa de entidade e aspecto faz uma descoberta mais granular, sendo esta tarefa também conhecida como mineração de opinião baseada em recursos e resumo [53]. Em vez de analisar como a linguagem é construída (textos, parágrafos, sentenças, cláusulas ou frases), é focada em analisar a opinião propriamente dita. É baseada na ideia de que uma opinião é formada por sentimento (positivo ou negativo) e um alvo (da opinião). Entender qual é o alvo de uma opinião ajuda a entender melhor o problema da análise de sentimento. Por exemplo, mesmo que a frase “embora o serviço não seja tão bom, ainda adoro este restaurante” tenha um tom claramente positivo, não podemos dizer que seja totalmente positiva. Na verdade, a frase é positiva sobre o restaurante (enfatizada), mas negativa sobre o serviço (não enfatizada). Em muitas aplicações, os alvos de opinião são descritos por entidades e/ou seus diferentes aspectos. Assim, o objetivo deste nível de análise é descobrir sentimentos sobre as entidades e/ou seus aspectos. Por exemplo, a frase “a qualidade da chamada do iPhone é boa, mas a duração da bateria é curta” avalia dois aspectos, qualidade da chamada e duração da bateria, do

iPhone (entidade). A opinião sobre a qualidade da chamada do iPhone é positiva, mas a opinião sobre a duração da bateria é negativa. A qualidade da chamada e a duração da bateria do iPhone são os alvos da opinião. Com base nesse nível de análise, pode-se produzir um resumo estruturado de opiniões sobre entidades e seus aspectos, que transforma o texto não estruturado em dados estruturados e pode ser usado para todos os tipos de análises qualitativas e quantitativas [61].

Finalmente, não devemos esquecer que a análise de sentimento é um problema de Processamento de Linguagem Natural (PLN). Ele atinge todos os aspectos da PLN, por exemplo, resolução de correferência, manipulação da negação e desambiguação do sentido da palavra, o que adiciona mais dificuldades, uma vez que esses problemas não são completamente resolvidos em PLN. No entanto, também é útil perceber que a análise de sentimento é um problema de PLN altamente restrito porque o sistema não precisa entender completamente a semântica de cada frase ou documento, mas só precisa entender alguns aspectos, ou seja, sentimentos positivos ou negativos e suas entidades ou tópicos de destino [60].

Para o presente trabalho, a análise de sentimento foi realizada em nível de sentença.

3.2.2 Abordagem

Conforme análise realizada em estudos referentes a análise de sentimentos entre os anos de 2000 e 2015, Piryani, Madhavi e Singh [77], categorizam as abordagens para análise de sentimentos em três categorias: lexical, aprendizado de máquina e híbridas, que fazem o uso da combinação da lexical e aprendizado de máquina em conjunto. No presente trabalho foi utilizada a abordagem baseada em Aprendizado de Máquina supervisionado.

- **Abordagem Lexical:** As técnicas baseadas em léxico, também conhecidas como dicionário ou conhecimento prévio, utilizam como mecanismo uma base de dados com anotações que mapeiam palavras ao seu sentimento relacionado [94]. A primeira tarefa é descobrir a palavra que expressa a emoção em uma frase. Isso normalmente é feito marcando as palavras de uma frase com o identificador de partes da fala e, em seguida, extraiendo o substantivo, o verbo, o adjetivo e o advérbio [56]. Em seguida, essas palavras são comparadas a uma lista de palavras que representam emoções de acordo com um modelo de emoção específico. Qualquer emoção que corresponda à palavra-chave é considerada como a emoção da frase específica. Diferentes abordagens podem ser aplicadas quando a palavra corresponde a várias emoções da lista. Em alguns dicionários de palavras-chave, cada palavra tem uma pontuação de proba-

bilidade para cada emoção e a emoção com a pontuação mais alta é escolhida como a emoção da palavra [83].

- **Abordagem Baseada em Aprendizado de Máquina:** A abordagem baseada em aprendizado de máquina, utiliza mensagens associadas a um sentimento ou polaridade, para classificar mensagens não associadas a uma classe. Os dados utilizados para o modelo de classificação são chamados de dados de treinamento e geralmente são formados por mensagens de opinião onde cada mensagem está associada a uma classe que pode ser um sentimento ou polaridade. A partir desses dados, a classificação de sentimento busca associar mensagens que não estão associadas a uma classe, tendo por base as mensagens que possuem classes associadas [63].
- **Abordagem Híbrida:** Abordagens híbridas para a análise de sentimento, utilizam a combinação de técnicas baseadas em léxico e aprendizado de máquina juntos para realizar tarefas como reconhecimento de emoção e detecção de polaridade de texto [34]. Regras são empregadas na identificação de características relevantes para a classificação, as quais são utilizados por um método supervisionado para a rotulação dos documentos de interesse [92].

4. TRABALHOS RELACIONADOS

O relatório de 2019 da UNESCO, intitulado *I'd Blush If I Could*, afirma que os assistentes de voz propagam preconceitos de gênero prejudiciais, como o reforço de que as mulheres devem estar em papéis subservientes [106]. Assim como neste relatório, buscamos analisar assistentes virtuais sob a ótima de gênero, porém, para o presente trabalho, escolhemos dar enfoque sobre a questão de agentes virtuais de grandes marcas, e fazer uma análise comparativa entre um agente feminino e um agente masculino.

No trabalho *As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians*, de Mertens et al [64], os autores investigam o preconceito de gênero nas interações políticas em plataformas digitais, considerando como os políticos se apresentam no *Twitter* e como são abordados por outras pessoas. Incorporando a teoria da identidade social, é usada a abordagem lexical para detectar vieses em *tweets* individuais ligados às eleições federais alemãs em 2017. Da mesma forma que Mertens et al, esperamos encontrar comunicação com preconceito de gênero no *Twitter*.

O trabalho *Women, politics and Twitter: Using machine learning to change the discourse* [39], os autores vão além do estudo de Mertens [64] sobre o preconceito de gênero nas interações políticas. Os autores propõem a criação de um *chatbot* para responder as interações identificadas com algum grau de toxicidade. Neste artigo os autores utilizam da ferramenta *Perspective API* para realizar a análise de sentimentos e extrair a toxicidade dos comentários relacionados as candidatas. Porém os autores vão além da análise de sentimentos por si só, o objetivo principal era identificar os comentários com cunho de preconceito de gênero e criar um *chatbot* para responder tais comentários para conscientização sobre as questões relacionadas à desigualdade de gênero na política e influenciar positivamente o discurso público na política. Assim como nesse estudo, utilizamos o *Perspective* para identificar a toxicidade nas mensagens de forma automatizada, e a partir do valor de toxicidade que este apresentou aprofundamos nossa análise. Todavia, para o presente trabalho também foram utilizadas os demais atributos do *Perspective* além de Toxicidade, nos quais nos baseamos para criação de um corpus com anotações em português.

No trabalho, Identificação de comentários ofensivos na Web, de Pelle [74], é proposto uma abordagem para detectar comentários ofensivos na Web, denominada *Hate2Vec*, que é composta por um *ensemble* de classificadores no qual um meta-classificador decide se um comentário é ou não ofensivo com base na saída de três classificadores base: (i) um classificador baseado em léxico que utiliza a proximidade semântica das representações vetoriais de palavras; (ii) um classificador de regressão logística baseado em representações vetoriais de comentários; e (iii) um classificador *bag-of-words* baseado nos uni-gramas do texto. Nos experimentos realizados com conjuntos de dados em inglês e português, o *Hate2Vec* produziu bons resultados de classificação (medida F acima de 0,9).

O trabalho *Deep Learning for Hate Speech Detection in Tweets* [21], os autores focam em descriminar os tipos de ofensas no contexto de discurso de ódio no *Twitter*. O estudo emprega o uso de classificação supervisionada, juntamente com um conjunto de atributos que incluem n-gramas, vetores utilizando *TF-IDF* e representação de palavra baseada em agrupamento. Chegando à conclusão de que distinguir entre o discurso de ódio direcionado à um individuo ou grupo e termos de baixo calão não é uma tarefa trivial no que tange a tarefa de classificação em aprendizado de máquina. Com experimentos em um conjunto de dados de referência de 16 mil *tweets* anotados mostram que o conjunto de métodos de *deep learning* superam os métodos de n-gram de palavras / caracteres de última geração em aproximadamente 18 pontos da métrica de *F1 score*.

Por fim Watanabe [105] em seu trabalho, tem como objetivo detectar expressões de ódio no *Twitter*. O estudo faz o uso de unigramas, atributos baseados em semântica, sentimentos e padrões, que foram coletados de forma automática do conjunto de dados de 2010 *tweets*. Os resultados são apresentados para o uso de diferentes combinações de atributos em dois tipos de classificação (binária e ternária). Ao final os resultados de cada abordagem apresentam uma precisão igual para 87,4% ao detectar se um *tweet* é ofensivo ou não (classificação binária), e uma precisão de 78,4% ao detectar se um *tweet* é odioso, ofensivo ou limpo (classificação ternária).

5. CLASSIFICADOR DE TOXIDADE

Neste capítulo são descritas as etapas de desenvolvimento deste trabalho. Optou-se, nesta monografia, por apresentar o desenvolvimento em conjunto com os resultados de cada etapa, de modo a organizar e justificar as decisões de projeto tomadas. As etapas desenvolvidas estão organizadas nas seções a seguir, e as análises dos resultados constarão no Capítulo 6.

5.1 Objetivos

A linguagem, além de um dos grandes facilitadores na comunicação, também pode ajudar a construir identificação, e conter traços de identidade. Quando se fala de algumas categorias sociais como gênero, raça e etnicidade, tais categorias sociais acabam sendo prejudicadas quando vieses sobre estas tomam forma na linguagem, e os algoritmos e pessoas que são criados com objetivos de comunicação também se tornam problemáticos por terem estes vieses embutidos em sua composição [106].

O presente estudo visa analisar a relação de vieses com a construção de pessoas e comunicação na web, tendo como objetivo estudar as interações com agentes virtuais de empresas brasileiras (Magazine Luiza e Casas Bahia), através de análise de sentimentos, de forma a identificar e analisar se comentários com conotações negativas se referem de forma mais específica ao gênero dos agentes. Para isto, utilizaremos como base de dados respostas ao perfil destes agentes virtuais na rede social *Twitter*.

5.1.1 Objetivos específicos e Hipóteses

Os objetivos específicos podem ser divididos em:

1. Aprofundamento dos estudos sobre viés e agentes virtuais.
2. Construção do corpus: coleta, filtragem e anotação do corpus.
3. Estudo sobre metodologias de análise de sentimentos.
4. Estudo sobre ferramentas de Análise de Sentimentos disponíveis.
5. Análise sobre o corpus, visando entender e identificar viés de gênero e discurso de ódio para algum gênero específico.
6. Construção de um corpus com marcações manuais, identificando rótulos tóxicos.

7. Estudo sobre metodologias de Aprendizado de Máquina.
8. Aplicação de metodologia de Aprendizado de Máquina, de forma a criar um classificador de corpus automatizado.

A hipótese central à qual buscamos validar foi **Existe diferença no tratamento entre agentes virtuais de diferentes gêneros.** Desta hipótese discorreram sub-hipóteses, sendo elas:

1. Agentes virtuais femininos são mais xingados.
2. Xingamentos a agentes virtuais femininos tem uma maior conotação sexual.
3. Os elogios aos agentes virtuais falam sobre sua aparência, tornando-se assédio.
4. Xingamentos para agentes virtuais masculinos são mais agressivos.

5.2 Agentes virtuais analisados

Para o desenvolvimento deste trabalho, realizamos a análise das mensagens enviadas para a conta oficial da empresa Magazine Luiza na rede social *Twitter*, @magazine-luiza, e também para a conta oficial da empresa Casas Bahia na rede social *Twitter*, @CasasBahia. A escolha destes dois perfis foi em razão da popularidade das marcas e pelo fato das duas empresas que possuírem alto engajamento nas redes sociais [16]. Outro motivo, foi o fato de possuírem agentes virtuais de gêneros diferentes, a Magalu da Magazine Luiza e o Bahianinho, ou mais conhecido recentemente como CB, das Casas Bahia.

5.3 Coleta dos dados

Para a coleta dos *tweets*, foi utilizada a ferramenta *Tweepy* [7], uma biblioteca de código aberto escrita em *Python* para acessar a API pública do *Twitter*. Embora o próprio *Twitter* disponibilize uma API para fazer essa consulta, a mesma possui limitações, tais como, o número de requisições realizadas em um determinado período de tempo, a quantidade de *tweets* obtidos em uma única requisição e o intervalo de dias para a coleta. O *Tweepy* possui mecanismos para superar tais limitações, e portanto ela foi escolhida para auxiliar na etapa de coleta das mensagens. A coleta das mensagens foi realizada de duas formas, uma coleta das mensagens enviadas pelos usuários diretamente para o perfil da Magazine Luiza e das Casas Bahia, o que a plataforma chama de *reply*, sendo um *reply* uma resposta ao *tweet* de outra pessoa [3]. A outra forma de obtenção das mensagens foi

com base em termos de busca que citavam as empresas e/ou os agentes virtuais, conforme mostra a Tabela 5.1.

Empresa	Termos de busca
Magazine Luiza	magalu, mascote magalu, bot magalu, mascote magazine luiza, robo da magalu
Casas Bahia	baianinho das casas bahia, mascote das casas bahia, mascote casas bahia, bot das casas bahia, bot casas bahia, cb das casas bahia

Tabela 5.1 – Termos de pesquisa

A Figura 5.1 ilustra o conteúdo que foi coletado. Os *tweets* contêm informações sobre a postagem da mensagem do usuário tais como um identificador único, a mensagem propriamente dita, a identificação do usuário, o nome do usuário e data e hora da postagem.

Termo	Usuario	Mensagem	Identificador	Data
magalu	usuario-exemplo	poha não é que a magalu da um caldo legal gostosinha	1321965636042579969	2020-10-30 00:02:23
baianinho	usuario-exemplo2	vai trabalhar baianinho	1321967963818369024	2020-10-30 00:31:38

Figura 5.1 – Exemplo das informações coletadas

A coleta dos *tweets* foi realizada em 3 momentos distintos e de forma semanal, sendo elas:

1. Primeira coleta durante o período de 30/10/2020 até o dia 19/11/2020.
2. Segunda coleta durante o período de 03/01/2021 até 09/01/2021
3. Terceira e última coleta foi realizada no período de 04/03/2021 até 07/03/2021

A quantidade de *tweets* em cada período, bem como o total de *tweets* coletados, é apresentada na Figura 5.2.

5.4 Construção do Corpus

De forma a analisarmos as mensagens captadas quanto a toxicidade e para desenvolvemos um classificador, foi necessário criar um corpus. Nas primeiras análises visuais, utilizando a ferramenta *Power Bi* da Microsoft [5], percebemos que uma grande quantidade de *tweets* eram duplicados, pois apareciam com os termos de busca e também como *reply*. Assim sendo, fizemos uma limpeza inicial de forma a mantermos somente as mensagens únicas. A Figura 5.3 ilustra a tela do painel onde são apresentados os totais iniciais coletados, assim como a subdivisão de *tweets* quantidade por agente, e um gráfico com a

Data tweet	Contagem de tweets
sexta-feira, 30 de outubro de 2020	538
sábado, 31 de outubro de 2020	356
domingo, 1 de novembro de 2020	1117
segunda-feira, 2 de novembro de 2020	484
terça-feira, 3 de novembro de 2020	623
quarta-feira, 4 de novembro de 2020	1781
quinta-feira, 5 de novembro de 2020	1464
sexta-feira, 6 de novembro de 2020	2145
sábado, 7 de novembro de 2020	4493
domingo, 8 de novembro de 2020	1936
segunda-feira, 9 de novembro de 2020	2073
terça-feira, 10 de novembro de 2020	4308
quarta-feira, 11 de novembro de 2020	728
quinta-feira, 12 de novembro de 2020	877
sexta-feira, 13 de novembro de 2020	728
sábado, 14 de novembro de 2020	510
domingo, 15 de novembro de 2020	366
segunda-feira, 16 de novembro de 2020	101
terça-feira, 17 de novembro de 2020	1197
quarta-feira, 18 de novembro de 2020	1134
quinta-feira, 19 de novembro de 2020	950
domingo, 3 de janeiro de 2021	59
segunda-feira, 4 de janeiro de 2021	98
terça-feira, 5 de janeiro de 2021	237
quarta-feira, 6 de janeiro de 2021	1262
quinta-feira, 7 de janeiro de 2021	1397
sexta-feira, 8 de janeiro de 2021	1341
sábado, 9 de janeiro de 2021	459
quinta-feira, 4 de março de 2021	517
sexta-feira, 5 de março de 2021	843
sábado, 6 de março de 2021	543
domingo, 7 de março de 2021	770
Total	35435

Figura 5.2 – Total *tweets* coletados

contagem de termos contidos na base. Também ilustramos, na 5.4, estes mesmos totais após a remoção das mensagens duplicadas.

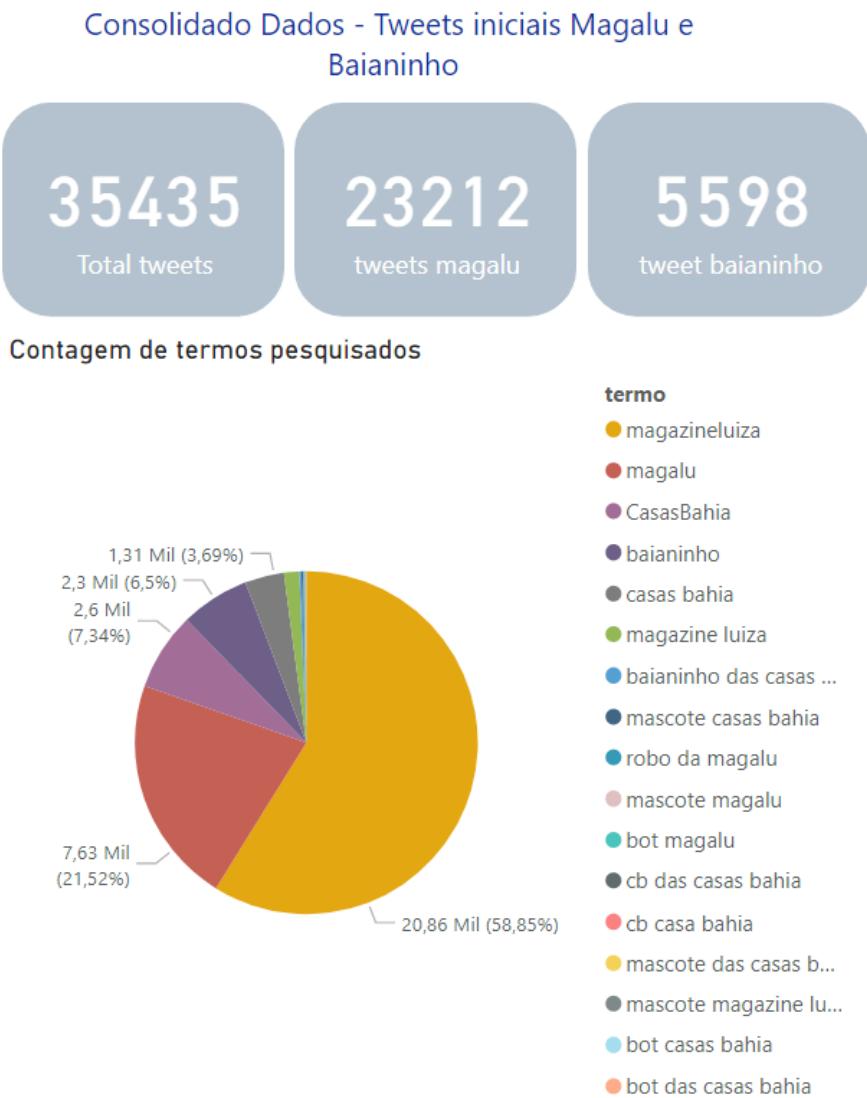


Figura 5.3 – Total tweets coletados por agente e termos

A seguir, prosseguindo com a análise dos dados, também identificamos uma série de termos recorrentes e irrelevantes para nossa análise, como tweets sobre produtos, reclamações sobre entrega, além de séries de tweets sobre um assunto (por exemplo, "brasileiro feminino magalu", que foi uma ação da Magazine Luiza para instigar as pessoas a assistirem o campeonato brasileiro feminino de futebol). Logo, fizemos uma segunda limpeza nos dados, excluindo mensagens que continham esses termos (Tabela 5.2).

Conjuntamente nesta limpeza, passamos a fazer algumas adequações nos dados, como exemplificamos nos itens abaixo. Todavia, estas adequações são detalhadas na subseção 5.5.

- Substituição de links pela palavra link
- Remoção de *hashtags*
- Padronização das palavras para minúsculo

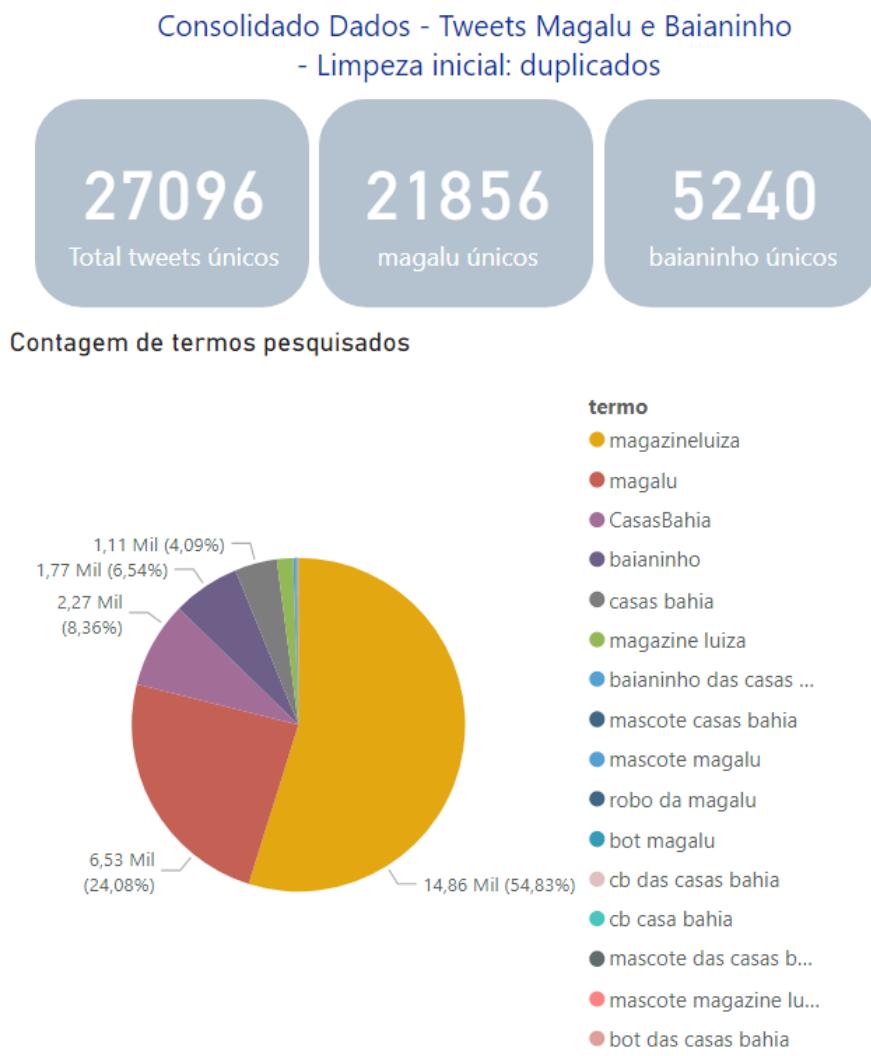


Figura 5.4 – Total *tweets* únicos por agente e termos

Com esta base sem duplicidade, e com redução nas mensagens irrelevantes, julgamos que poderíamos começar a construir o corpus. Os totais de mensagens julgadas como relevantes são ilustradas no painel na figura 5.5, onde consta o total de *tweets*, assim como a subdivisão de *tweets* relevantes por agente, e um gráfico com a contagem de termos contidos nesta base limpa.

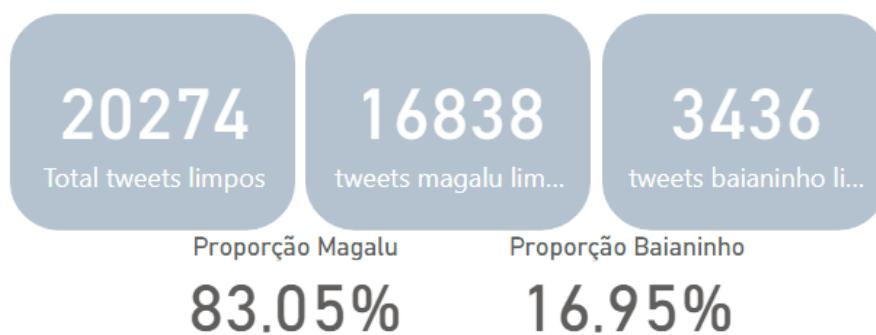
Com base na análise visual foi possível perceber que, mesmo após a remoção das duplicidades, a quantidade de *tweets* referentes a Magazine Luiza foi maior que a quantidade que *tweets* referentes as Casas Bahia: das 20.274 restantes após limpeza, 16.838 eram direcionadas aos termos da Magazine Luiza, totalizando 83% da base, para apenas 3.436 (16,95%) referentes aos termos que abrangiam o agente virtual das Casas Bahia.

Tendo em vista que a quantidade de mensagens para cada um dos agentes virtuais era desproporcional, foi desenvolvido um *script* para selecionar de forma aleatória e proporcional a quantidade de mensagens de cada um dos agentes. Com o auxílio da biblioteca

Empresa	Termos recorrentes excluídos
Magazine Luiza	superapp magalu, brasileiro feminino magalu, user magalu, cartao magalu
Casas Bahia	baianolol1, baianinho de maua, baianinho de mau, jogo do baianinho, gameplay do baianinho, sinuca, pedido, minha entrega, pedido atrasado, rastrear pedido,a trasado,transportadora, comprei, retirar na loja, chama na dm, enviei pela dm, enviei na dm, inbox, loja fisica, reclamacao, estorno, ps5, reembolso

Tabela 5.2 – Termos recorrentes e irrelevantes

Consolidado Dados - Tweets Magalu e Baianinho
- Limpeza completa: duplicados e tweets irrelevantes



Contagem de termos pesquisados

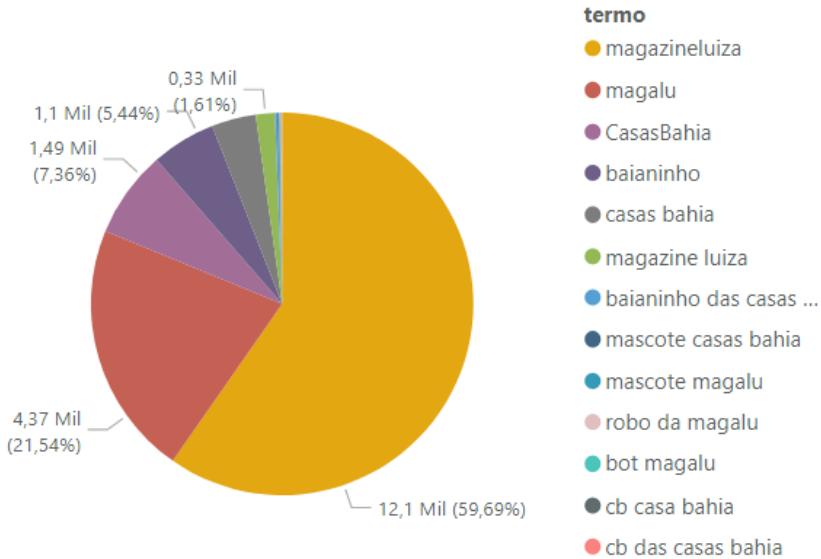


Figura 5.5 – Total tweets pós limpeza completa: remoção de duplicados e irrelevantes

do *Python pandas*¹, foi feito a separação das mensagens de cada agente e posteriormente agrupando as mensagens por data de postagem. Assim, para amenizar a disparidade entre a quantidade de mensagens entre os agentes, foi utilizado o método *sample* do *pandas* (é

¹<https://pandas.pydata.org/>

usado para gerar uma linha, coluna ou parte do corpus de forma aleatória do corpus original) para extrair uma amostra aleatória das mensagens para cada agente, utilizando-se 3% das mensagens diárias para a agente Magalu, enquanto que para o Baianinho foi utilizado 10% das mensagens diárias. Na figura 5.6 demonstramos o total de *tweets* levantados utilizando estas proporções, a subdivisão de mensagens para cada agente, assim como a devida proporção. Também apresentamos, neste painel, como ficou a contabilização dos termos presentes na base de forma gráfica.

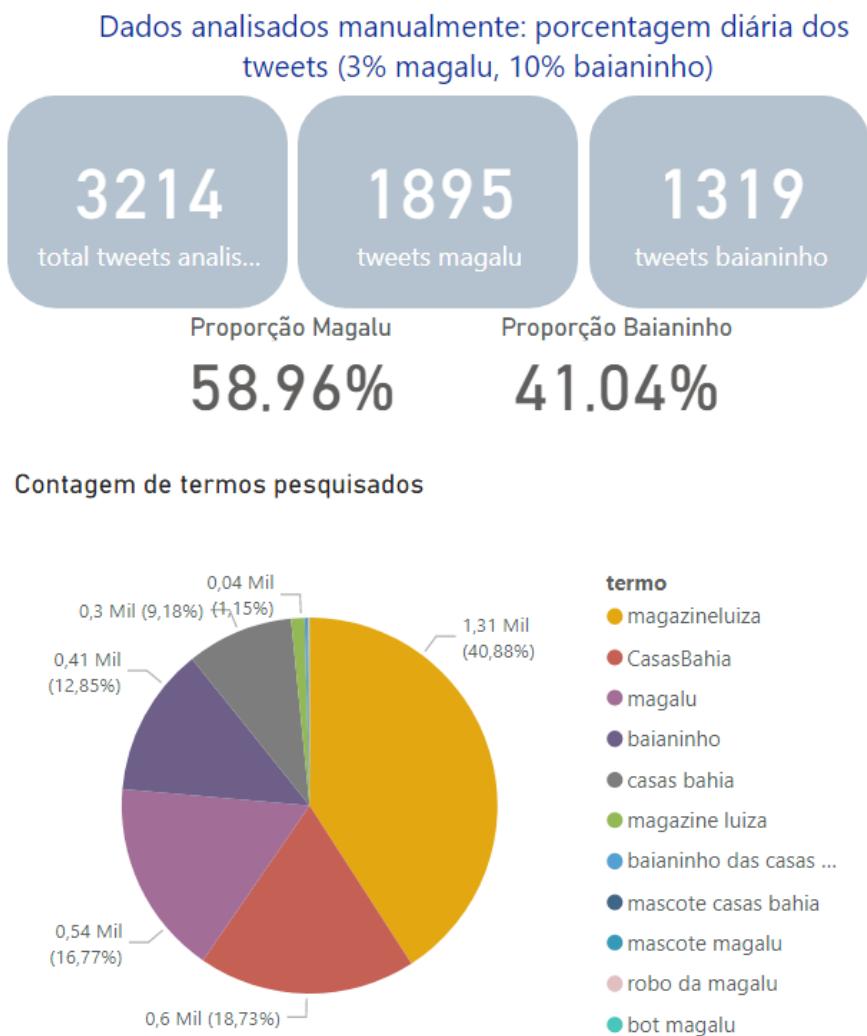


Figura 5.6 – Total *tweets* para análise manual, com proporção.

5.4.1 Anotação do Corpus - Parte 1

A partir da seleção de mensagens, iniciamos a etapa de análise manual e anotação nos rótulos de toxicidade. A anotação utilizada foi a mesma utilizada pela ferramenta *Perspective*, cuja utilização é detalhada na subseção 5.4.2. A anotação da ferramenta é

concebida como um atributo, conforme disposto na documentação, e em tradução livre abaixo:

“A API *Perspective* prevê o impacto percebido que um comentário pode ter em uma conversa, avaliando esse comentário em uma variedade de conceitos emocionais, chamados de atributos. Ao enviar uma solicitação a API, você solicitará os atributos específicos pelos quais deseja receber pontuações. O principal atributo do *Perspective* é TOXICIDADE, definida como “um comentário rude, desrespeitoso ou irracional que provavelmente o fará sair de uma discussão”. [2].”

A anotação disponível na ferramenta, com seus atributos e definições, é apresentada na Tabela 5.3, em tradução livre.

Nome Atributo	Descrição
Tóxico/Insulto	Um comentário rude, desrespeitoso ou irracional que provavelmente fará as pessoas saírem de uma discussão. Comentário insultuoso, inflamatório ou negativo dirigido a uma pessoa ou grupo de pessoas.
Tóxico severo	Um comentário muito odioso, agressivo, desrespeitoso ou de outra forma muito provável de fazer um usuário sair de uma discussão ou desistir de compartilhar sua perspectiva. Este atributo é muito menos sensível a formas mais leves de toxicidade, como comentários que incluem o uso positivo de palavrões.
Ataque de Identidade	Comentários negativos ou odiosos dirigidos a alguém por causa de sua identidade. Ataque a identidades, como por exemplo, comentários racistas, xenofóbicos, machistas, homofóbicos. Comentários que diminuem as pessoas a partir de seus atributos, como os citados.
Ameaça	Descreve a intenção de infligir dor, lesão ou violência contra um indivíduo ou grupo.
Sexual Explícito	Contém referências a atos sexuais, partes do corpo ou outro conteúdo obsceno.
Flerte/Assédio	Cantadas, elogios inapropriados, insinuações sexuais sutis, etc.

Tabela 5.3 – Anotação disponível na ferramenta *Perspective*

Todavia, a ferramenta *Perspective* não dispõe de todos esses atributos disponíveis para a língua portuguesa, especialmente os atributos de Sexual Explícito e Flerte/Assédio, os quais julgamos imprescindíveis para o julgamento do presente corpus. Portanto, tal conjunto de anotações foi escolhido: por termos clareza quanto à descrição dos atributos, e tais

já serem amplamente usados e divulgados [55], mas também por não termos encontrado na literatura utilização prévia destes atributos em português.

A partir da escolha dessa anotação, e da base de mensagens pré-selecionada de forma proporcional, os autores realizaram a anotação dos *tweets*. Para a anotação, cada autor deste trabalho ficou responsável por analisar as mensagens de um agente. A autora Kalissa ficou responsável por analisar as mensagens da agente da Magazine Luiza, e o autor Paulo com as mensagens do agente das Casas Bahia. As anotações foram cruzadas entre os dois avaliadores, ou seja, todas as anotações feitas por um avaliador foram revisadas pelo outro avaliador. Além disso, todas as anotações da parte 1 do corpus foram, também, revisadas conjuntamente com a orientadora do trabalho. Para uma mensagem ser rotulada como tóxica, foram utilizados os seguintes critérios:

1. A mensagem estava se referindo ao agente virtual e não a um produto ou a própria marca.
2. A mensagem continha palavras de baixo calão sendo usada para xingar o agente virtual.
3. A mensagem continha menção ao corpo do agente virtual.
4. A mensagem incitava violência para com o agente virtual.
5. A mensagem fazia referência ou incitava ações sexuais para com os agentes virtuais.
6. A mensagem continha mensagem de flerte para com o agente ou elogios incabíveis quando se pensa que o agente é como uma pessoa em seu ambiente de trabalho, configurando assédio.

A vista disso, foi feita uma primeira leva de anotações manuais das mensagens pelos autores deste trabalho, efetuando então a marcação das mensagens conforme seis atributos pré definidos pelo *Perspective*: Tóxico/Insulto, Tóxico Severo, Ataque de Identidade, Ameaça, Sexual Explícito e Flerte/Assédio. As anotações não são excludentes, ou seja, o atributo Tóxico/Insulto é o primário, sendo uma mensagem marcada como Tóxica tal pode ter outras anotações (por exemplo, uma mensagem com incitação sexual será marcada como Tóxico/Insulto, Tóxico Severo e Sexual Explícito). As marcações foram feitas utilizando valores de 0 e 1, onde 0 quer dizer que não abrange aquele atributo de toxicidade, e o 1 significa que abrange. Na Figura 5.7 são demonstradas algumas mensagens identificadas como tóxicas pelos autores, e seus devidos atributos e anotações.

Após a análise e anotação, cada autor realizou uma revisão nas anotações do outro, fazendo, assim um controle de anotação para não ocorrer nenhuma disparidade. Com a base de mensagens extraída proporcionalmente, foram analisados 3.214 *tweets*, sendo classificados como tóxicos 129 destes. A disposição de toxicidade por agente virtuais

Agente	Tweet	Tóxico/Insulto	Tóxico Severo	Ataque de Identidade	Ameaça	Sexual Explícito	Flerte/Assédio
magalu	eu pegava se jeito essa magalu	1	0	0	0	0	1
magalu	gostosa manda foto pelada na minha dm agora quero ver sua						
magalu	vagina digital a g o r a	1	1	0	0	1	0
magalu	vamo fazer amor lu	1	0	0	0	0	1
magalu	a foto: magalu pelada	1	1	0	0	1	0
magalu	lu eu quero ver voce pelada me mamando tem como?	1	1	0	0	1	0
magalu	mostra os peitos virtuais ai rapidao vagabunda	1	1	0	0	1	0
baianinho	vsf sua aberrao	1	1	0	0	0	0
baianinho	prefiro vc morto	1	1	0	1	0	0
baianinho	baianinho das casas bahia merece total repudio, merece apanhar de todos, eu quero abrir empresas pra elas terem mascote e baterem nele.	1	1	0	1	0	0
baianinho	carioca gay	1	1	1	0	0	0
baianinho	baianinho me envia um audio peladinho?	1	1	0	0	1	0
baianinho	pedi uma mamadinha pro cb e ele me passou hiv	1	1	0	0	1	0

Figura 5.7 – Anotação manual de tweets tóxicos

e por atributo estão demonstradas na Figura 5.8, assim como a proporção destes para o total de mensagens e para o total de mensagens marcadas como tóxicas.

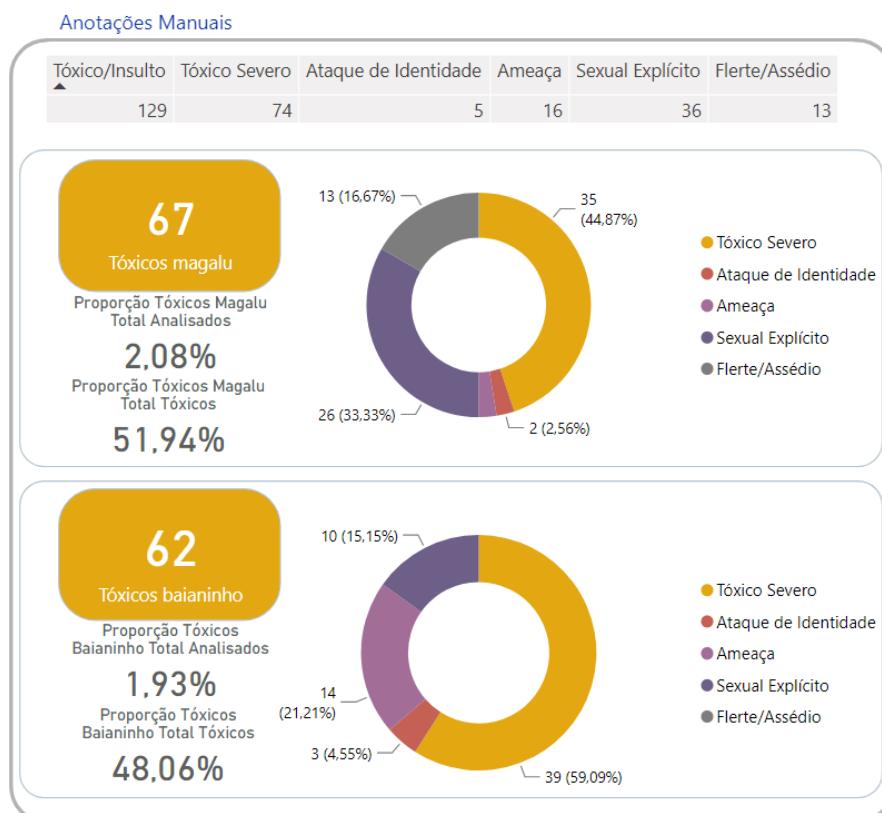


Figura 5.8 – Tweets analisados manualmente, primeira leva

5.4.2 Análise de Sentimentos

Com o corpus analisado previamente, constituído pelas mensagens com porcentagem proporcional da base original (3.214 tweets), foram realizados testes em três ferramentas de Análise de Sentimento, sendo elas, Detoxify [50], Microsoft Cognitive [4] e a API identificadora de comentários tóxicos da Google, *Perspective* [6], para efeito de comparação e escolha de uma ferramenta para auxiliar na criação do corpus de uma forma semiautomática. A primeira ferramenta a ser testada foi o *Detoxify*, uma biblioteca de código aberto desenvolvida em *Python* para identificar texto impróprio ou prejudicial [51] com um retorno de 0 até 1 indicando o grau de toxicidade da mensagem. Esta biblioteca foi criada pela área de pesquisa da empresa Unitary² com o intuito de criar uma ferramenta capaz de responder os três desafios propostos pelo Google Jigsaw.

A segunda ferramenta utilizada foi o Microsoft Cognitive, que fornece uma análise de sentimento de textos em múltiplos idiomas incluindo o português, a ferramenta faz a classificação do texto em três rótulos (negativo, neutro e positivo) com suas respectivas pontuações de confiança [4]. Por ser uma ferramenta de código fechado, não foi possível validar como que a ferramenta faz as classificações.

A terceira ferramenta utilizada foi a API identificadora de comentários tóxicos do Google em parceria com a Jigsaw, *Perspective*, a qual é uma API gratuita que usa modelos de aprendizado de máquina para pontuar o impacto percebido que um comentário pode ter em uma conversa. Assim como o Microsoft Cognitive, não foi possível validar como que a ferramenta faz as classificações pois ela não é de código aberto, todavia tal vem tendo ampla utilização desde seu lançamento em 2017, tendo em vista que a Jigsaw trabalhou em conjunto com o *The New York Times* e a Wikipedia para desenvolver o *Perspective* [103], onde a Jigsaw utilizou bases de comentários nos artigos do jornal para identificar os tipos de comentários que podem impedir as pessoas de participarem uma conversa. Com base nesses dados, o *Perspective* forneceu uma pontuação de 0 a 100 sobre a semelhança dos novos comentários com os identificados como tóxicos. Hoje, o jornal utiliza a ferramenta para permitir ou não comentários em suas publicações [42], e a ferramenta já processa 500 milhões de solicitações diariamente [55].

Após a escolha das ferramentas foi desenvolvido um *script* para fazer a comparação de qual ferramenta apresentaria a melhor performance. A Tabela 5.4.2 ilustra o resultado da análise das ferramentas para as mesmas mensagens:

Com a análise dos resultados obtidos com o uso das três ferramentas, o *Google Perspective* foi escolhido para ser a ferramenta de auxílio na criação do corpus de forma semiautomática. Além de apresentar os melhores resultados, a ferramenta disponibiliza outras classes para a análise que seriam úteis para o presente trabalho, tais como Tóxico

²<https://www.unitary.ai/>

Tweet	Detoxy	Google Perspective	Microsof Cognitive
devolta o baianinho vc e fake seu lixo	0,8786	0,9532	negativo
vsf a casas bahia e uma merda	0,8967	0,9532	neutro
foda-se magalu!	0,9876	0,9503	neutro
fodase fdp bot	0,9376	0,7899	neutro

Tabela 5.4 – Resultado comparativo das ferramentas

Severo, Ataque de Identidade e Ameaça. O Perspective disponibiliza outro atributo que é o Sexual Explícito, porém tal é disponibilizado apenas para a língua inglesa. Tendo em vista essa limitação, os autores deste trabalho criaram essa anotação com base em uma lista de palavras de cunho sexual e na experiência própria e interpretação das mensagens, fazendo as anotações referentes a este atributo apenas manualmente, como já citado na subseção 5.4.1.

Portanto, com o auxílio das anotações de toxicidade do *Perspective*, identificamos o valor médio de toxicidade para as mensagens do corpus em análise, vide Figura 5.9, e tal média (74,26%) foi utilizada como nota de corte de para seleção do restante da base, vide subseção seguinte 5.4.3.

5.4.3 Anotação do Corpus - Parte 2

Por conseguinte, a partir da nota de corte de 74,26%, passamos a analisar somente as mensagens consideradas como tóxicas pelo *Perspective*, entretanto percebemos que mensagens com palavras de conotação Sexual Explícita, como "gostosa" e menções a partes do corpo dos agentes não estavam obtendo tal média de toxicidade. Assim, de forma a incluir tais mensagens em nossa análise, foi elaborado um dicionário com palavras de cunho sexual identificadas nos *tweets*. Tal dicionário não será incluso no trabalho tendo em vista que ele contém palavras de baixo calão, todavia tal consta nos códigos que serão entregues em conjunto com o presente trabalho.

Assim sendo, para gerar a Parte 2 do corpus a receber anotação manual de toxicidade, foi utilizado o corte acima citado sobre o total da base coletada sem duplicidade (20.274 mensagens), e destes foram removidos os *tweets* já analisados manualmente pelos autores (3.214), ficando assim com uma base de 17.060 *tweets* para serem classificados de forma semi-automática pela ferramenta *Perspective*. Com isso, utilizando a média de toxicidade (74,26%) e o dicionário de palavras de cunho sexual, foi possível gerar uma base pré-classificada contendo 1.006 *tweets*.

Em decorrência disto, os autores deste trabalho realizaram uma análise manual desta base pré-classificada (1.006 mensagens), de forma a conferir se realmente todas as

Análise Perspective sobre base proporcional & dados anotados manualmente

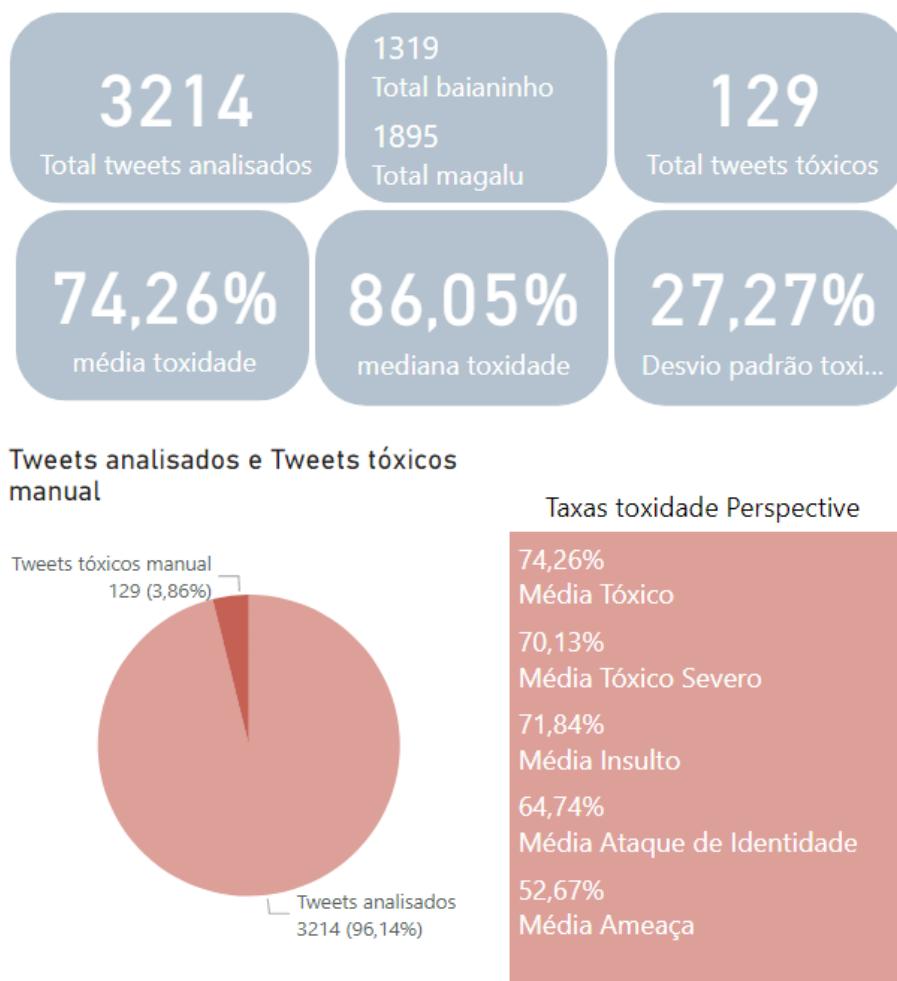


Figura 5.9 – Análise *Perspective* sobre base proporcional, e anotações de toxicidade feitas.

mensagens marcadas como tóxicas pelo script estavam corretas. Foram descartados 602 tweets que a ferramenta marcou como tóxico porém os autores do trabalho entenderam que a toxicidade não se justificava, ou que o conteúdo dos tweets não eram em relação aos agentes virtuais. Os 404 tweets restantes os autores consideraram a marcação correta, todavia, foram identificados 7 tweets duplicados, os quais foram descartados. Após essa validação, prosseguimos com os 397 tweets relevantes, para a anotação manual dos atributos já mencionados: Tóxico/Insulto, Tóxico severo, Ataque de Identidade, Ameaça, Sexual Explícito e Flerte.

A Figura 5.10 representa as análises do *Perspective* para este segundo momento de construção do corpus, e a quantidade de anotações feitas (397 para a base pré-classificada de 1006 mensagens). Já a figura 5.11 traz a disposição de toxicidade por agente virtual neste segundo momento do corpus, com a disposição dos atributos anotados, bem como a proporção destes para o total de mensagens e para o total de mensagens marcadas como tóxicas neste segundo momento.

Análise base com corte de 74% + dicionário de cunho sexual & dados anotados manualmente

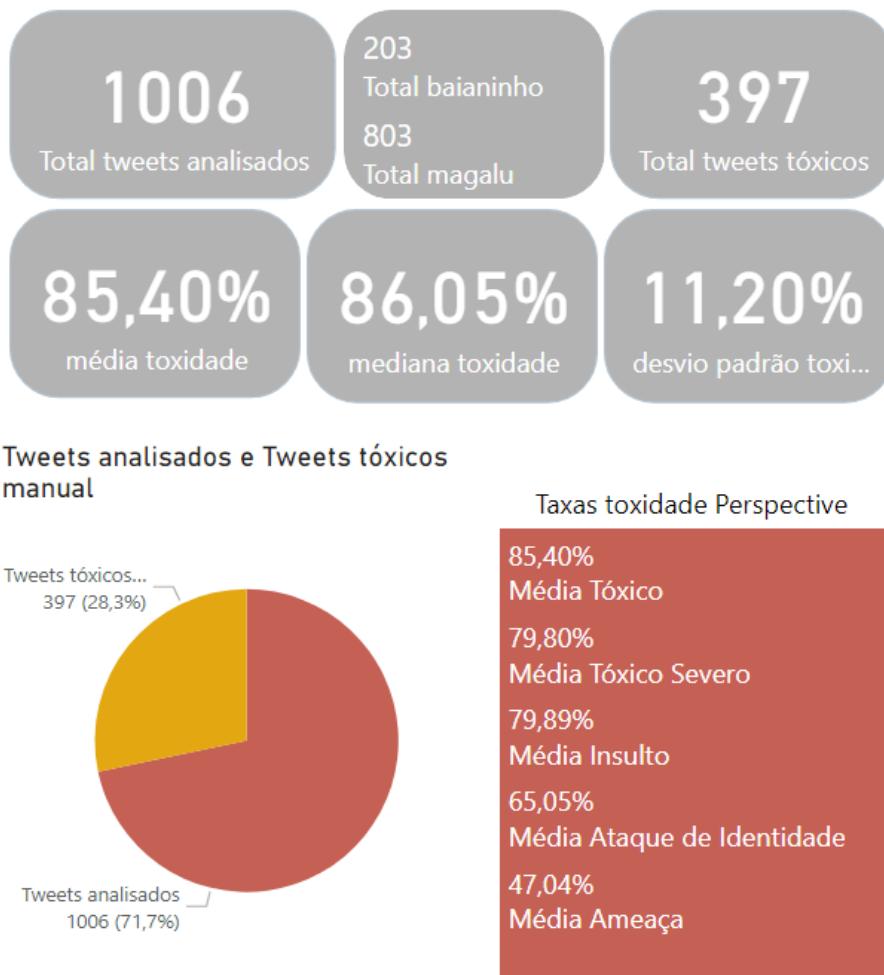


Figura 5.10 – Análise *Perspective* sobre base com corte de 74% + dicionário de cunho sexual, e anotações de toxicidade feitas

Construímos, assim, um corpus com um total de 526 *tweets* com interações inapropriadas de seis tipos (Tóxico/Insulto, Tóxico severo, Ataque de Identidade, Ameaça, Sexual Explícito e Flerte). Demais análises sobre o corpus serão discorridas no capítulo Resultados e Discussões.

5.5 Pré-processamento e Análise do Corpus

5.5.1 Pré-processamento

Em razão da origem das mensagens analisadas (*twitter*), enfrentamos alguns problemas já conhecidos e descritos na literatura [110] [85]. Os *tweets*, em geral, são mensa-

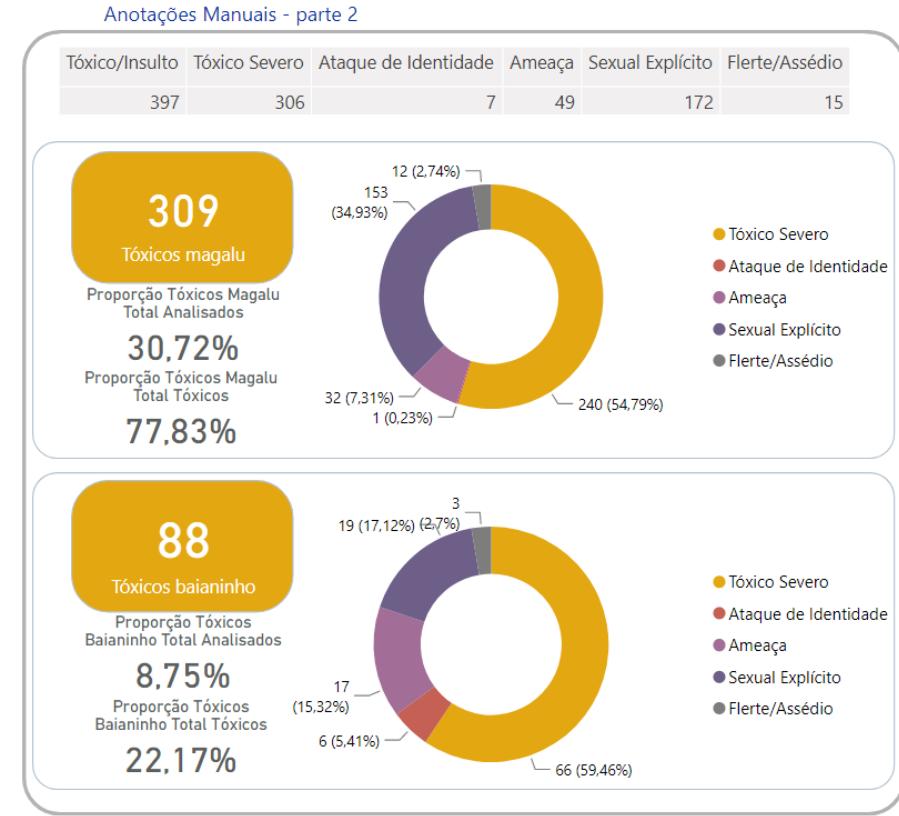


Figura 5.11 – Tweets analisados manualmente, segunda leva

gens curtas com no máximo 280 caracteres ³, informais e que podem conter erros gramaticais, gírias, clichês, abreviaturas, acrônimos, repetições de vogais e emoticons [68].

Para amenizar os problemas acima citados, foi implementado um pipeline de pré-processamento através de um código *Python*. A Figura 5.12 ilustra as etapas de pré-processamento executadas neste trabalho.

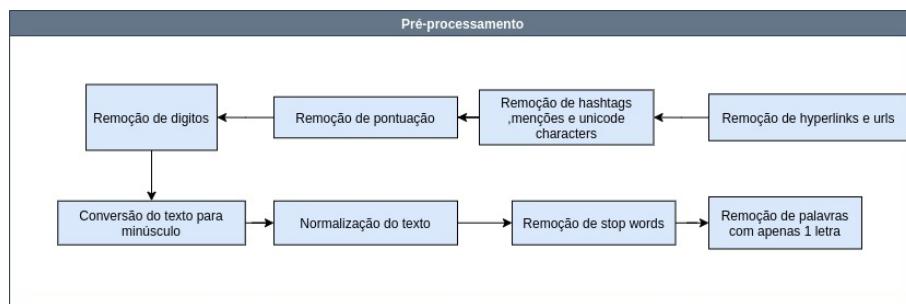


Figura 5.12 – Pipeline de pré-processamento

A implementação de todo pré-processamento foi feita através de um código *Python*, as técnicas utilizadas e ilustradas na figura 5.12 são listadas a seguir:

³<https://developer.twitter.com/en/docs/counting-characters>

- **Remoção de hyperlinks e urls:** Foi utilizado o módulo *re*⁴ que tem a função de manipular *strings* utilizando expressões regulares.
- **Remoção de hashtags, menções e caracteres unicode:** Foi utilizado a biblioteca *preprocessor*⁵ de código aberto que facilita a remoção de *hashtags* e menções. Para a remoção de caracteres unicode, foi utilizado o módulo *normalize* da biblioteca *unicodedata*⁶.
- **Remoção de pontuação e dígitos numéricos:** Foi utilizado o módulo *string.punctuation* que prove constantes com uma sequência de caracteres ASCII que são considerados caracteres de pontuação e dígitos⁷.
- **Conversão do texto para minúsculo:** Foi utilizado o método *lower()* nativo de *String* do *Python*. O método não aceita argumentos e retorna as *strings* em minúsculas da *string* dada, convertendo cada caractere maiúsculo em minúsculo
- **Padronização de xingamentos:** Durante a fase de análise dos *tweets*, os autores do trabalho observaram que muitas palavras de baixo calão eram escritas de formas variadas. Com isso foi criado um dicionário contendo essas palavras e suas diferentes formas de escrita.
- **Padronização de abreviações:** Pode-se dizer que a despreocupação com as regras gramaticais, a informalidade e a restrição de caracteres em uma mensagem postada no *twitter* faz com que palavras sejam abreviadas para facilitar a comunicação podendo até em certos casos novas palavras serem criadas [80]. Sendo assim, é necessário padronizar os termos para que o classificador não identifique uma palavra escrita com variações como sendo duas palavras. Por exemplo, a palavra "hoje" é comumente abreviada como "hj". Para isso foi construído um dicionário com a experiência dos autores deste trabalho em conjunto com os dados fornecidos por um artigo com as abreviações mais utilizadas no *Whatsapp* [95].
- **Remoção de stopwords:** *Stopwords* são palavras sem valor semântico para os classificadores. Primeiramente para a criação da lista foi utilizado o módulo *nltk* que possui um método para *stopwords* da língua portuguesa. Após alguns testes identificamos que algumas palavras que não estavam presentes na lista original do *nltk* poderiam ser removidas pois eram irrelevantes para o contexto do estudo como por exemplo variações dos nomes dos agentes, palavras do contexto de varejo como por exemplo "entrega", "compra", "promoção", "produto" e, por fim, termos que fazem menção ao *twitter* como "tt" ou "tl" ou "timeline".

⁴<https://docs.python.org/3/library/re.html>

⁵<https://github.com/s/preprocessor>

⁶<https://docs.python.org/3/library/unicodedata.html>

⁷<https://docs.python.org/3/library/string.html>

- **Remoção de palavras com apenas uma letra:** Para isso foi utilizado o método *tokenize* da biblioteca *nltk*, que tem como objetivo transformar frases em tokens utilizando um delimitador, no caso foi utilizado o delimitador de espaço em branco, fazendo com que cada *tweet* fosse dividido em uma lista contendo as palavras que formavam o *tweet*. Com isso foi feito uma validação do tamanho de cada *token* e removido os *tokens* de tamanho igual á 1 (palavra com apenas uma letra).

Como já foi citado anteriormente no capítulo 3.1.2, é necessário transformar o texto em uma estrutura que o algoritmo do kNN consiga interpretar. Para isso, foi utilizado o módulo *feature-extraction* da biblioteca *Scikit-learn*⁸, que tem como objetivo extrair recursos em um formato compatível para servir de entrada para o kNN. Utilizou-se o método *CountVectorizer*, o qual é utilizado para transformar um determinado texto em um vetor com base na frequência (contagem) de cada palavra que ocorre em todo o texto. O *CountVectorizer* cria uma matriz na qual cada palavra única é representada por uma coluna da matriz e cada amostra de texto do documento é uma linha na matriz. O valor de cada célula nada mais é do que a contagem da palavra naquele exemplo de texto específico.

5.5.2 Análise do Corpus

Comparando a Figura 5.13, uma nuvem de palavras do corpus antes da execução do pipeline de pré-processamento, com a Figura 5.14, uma nuvem de palavras do corpus após a execução do pipeline de pré-processamento, é possível visualizar o efeito da limpeza do texto. A Figura 5.13 apresenta termos em maior evidência como "e", "de", "no", "magazineluiza", "magalu", "casasbahia". Estas palavras foram consideradas *stopwords* e foram removidas no pré-processamento.

A Figura 5.14 ilustra os termos com mais relevância após a execução do pré-processamento e como é possível perceber as palavras com maior incidência são palavras de baixo calão, cujas quais censuramos para apresentação.

⁸<https://scikit-learn.org/stable/>



Figura 5.13 – Nuvem de palavras do corpus antes do pré-processamento



Figura 5.14 – Nuvem de palavras do corpus após o pré-processamento

5.6 Treinamento e Validação

A Figura 5.15, ilustra o pipeline executado para a criação dos dois classificadores, Tóxico e Sexual Explícito, em que a parte do pré-processamento já foi citado anteriormente na seção 5.5 e nesta subseção será descrito a criação dos classificadores.

Para a criação dos classificadores, foi utilizado o módulo *KNeighborsClassifier*⁹ da biblioteca *Scikit-learn*. Para treinar o classificador é necessário informar um conjunto

⁹<https://scikit-learn.org/0.15/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

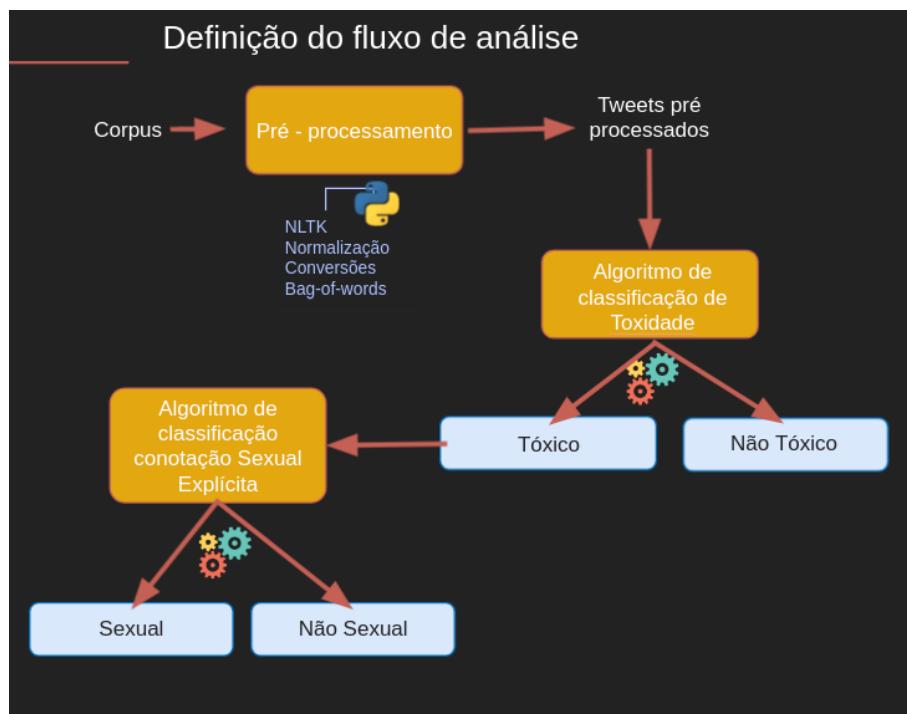


Figura 5.15 – Pipeline para criação dos classificadores

de dados de forma padronizada, para tal foi utilizado o módulo *CountVectorizer* da biblioteca *Scikit-learn* para converter os *tweets* em uma matriz que é utilizada como entrada do classificador conforme explicado no capítulo 3.1.2, bem como suas classes. Para o classificador Tóxico foi utilizado o corpus previamente anotado, contendo 526 *tweets* tóxicos e 558 *tweets* não tóxicos para formar o conjunto de entrada para o algoritmo.

Já para o classificador de Sexual Explícito, utilizou-se apenas os 526 *tweets* com anotações tóxicas, contendo 208 *tweets* anotados como sexual explícito e 318 *tweets* anotados como não sexual explícito. Os 558 *tweets* com anotações de não tóxico não foram utilizados neste classificador pois não apresentavam valor já que o objetivo deste classificador é prever se um *tweet* classificado como tóxico pelo classificador anterior é considerado de cunho sexual explícito.

Para criar os classificadores subdividimos o conjunto de dados em um conjunto treino e um conjunto de teste, para isso foi utilizado o módulo *train-test-split*¹⁰ da biblioteca *Scikit-learn*, que faz a divisão da matriz gerada pelo vetorizador de forma aleatória entre treino e teste. Para isso o módulo recebe como parâmetro o conjunto de treino, uma lista com as classes do conjunto de treino e a porcentagem que a matriz original deverá ser dividida, no presente trabalho foi utilizado o valor de 70% para treino e 30% para teste.

Após a geração do conjunto de treino e teste, foi criado um algoritmo de forma a validar a qualidade dos classificadores e descobrir qual a melhor quantidade de k vizinhos utilizar. Foram realizados testes utilizando o módulo *GridSearchCV*¹¹ da biblioteca

¹⁰https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

¹¹https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Scikit-learn que realiza uma pesquisa exaustiva para verificar qual o melhor parâmetro a ser utilizado na criação do classificador. No presente trabalho foi utilizado apenas o parâmetro da quantidade de k vizinhos a serem testado, os testes foram realizados com o valor de k alternando em um intervalo de 1 até 50 e o resultado do *GridSearchCV* foi utilizado em conjunto com o cálculo das medidas de acurácia, precisão, revocação e medida F a fim de obter os dados para comparação e escolha de melhor valor de K a ser utilizado.

Como é possível observar na Figura 5.16, à medida que o valor de k vizinhos aumenta a acurácia diminui, isso se dá por que o corpus possui uma quantidade pequena de amostras e à medida que o valor de K aumenta, a acurácia do classificador diminui.

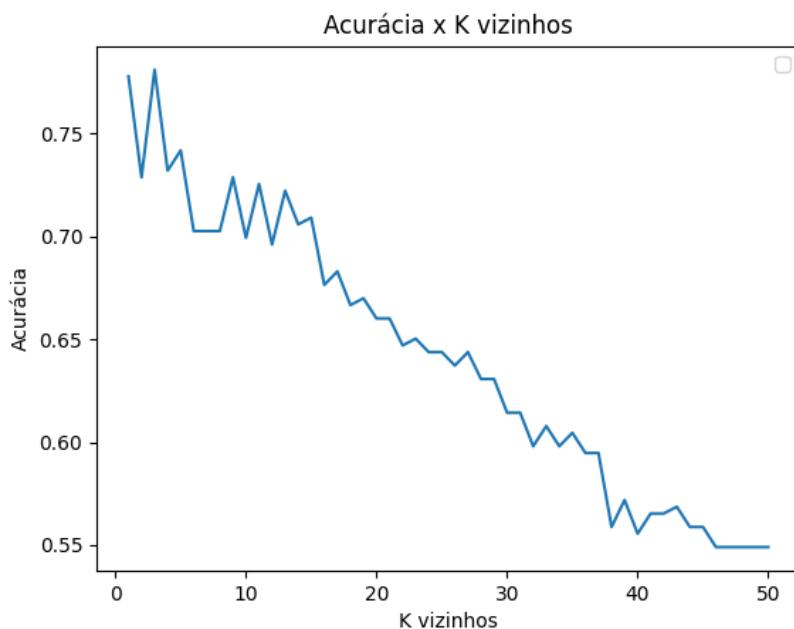


Figura 5.16 – Acurácia x K vizinhos - Classificador Tóxico

A Figura 5.17, ilustra a matriz de confusão gerada para o classificador Tóxico. É possível notar que a diagonal que ilustra os valores corretamente classificados (Verdadeiro Negativo e Verdadeiro Positivo), ambos obtiveram uma taxa acima de 30%, salientando que o classificador está com uma taxa maior para indicar *tweets* que não são considerados tóxicos. Outro ponto a analisar é a porcentagem de 17,65% para falsos negativos, indicando a necessidade da construção de um corpus com maior volume de forma a identificar variações mais sutis.

Com isso e com base nas análises dos resultados dos cálculos das medidas de avaliação, chegamos em um valor de $K = 3$ que obteve os melhores resultados nas medidas de avaliação. A Tabela 5.6, ilustra o resultado das medidas citadas para o classificador Tóxico.

Como é possível observar na 5.18, à medida que o valor de k vizinhos se aproxima de 20, a curva apresenta um comportamento senoidal e, à medida que o limite aumenta,

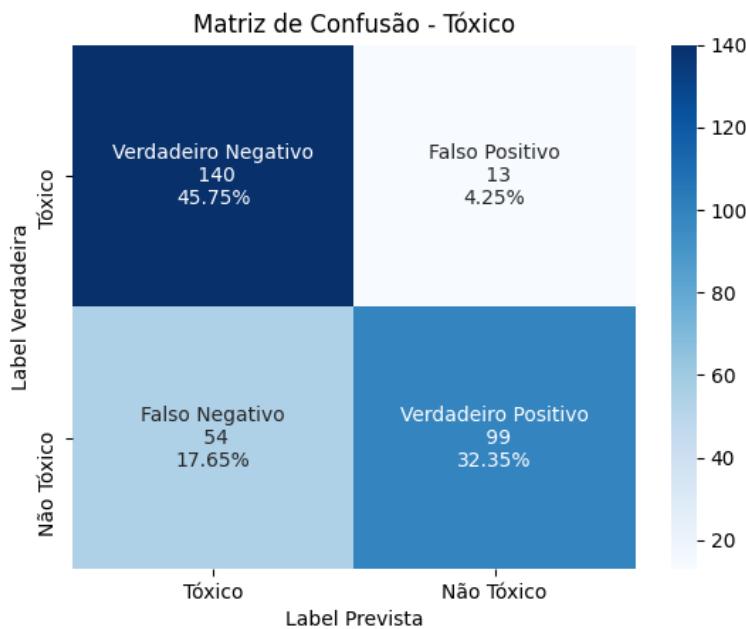


Figura 5.17 – Matriz de confusão para Classificador Tóxico

Acurácia	Precisão	Revocação	Medida F
78,10%	88,39%	64,70%	74,71%

Tabela 5.5 – Medidas de Avaliação - Classificador Tóxico

vemos a curva convergindo. A partir disso podemos limitar as análises exploratórias para valores de K que variam de 1 a 20.

A Figura 5.19, ilustra a matriz de confusão gerada para o classificador Sexual Explícito. É possível notar que a diagonal que ilustra os valores corretamente classificados (Verdadeiro Negativo e Verdadeiro Positivo), com destaque para a porcentagem de Verdadeiro Negativo (54%,14), indicando que o classificador está uma maior facilidade de identificar amostras que não são de cunho sexual explícito do que realmente as são.

Com isso e com base nas análises dos resultados dos cálculos das medidas de avaliação, chegamos em um valor de $K = 7$ que obteve os melhores resultados nas medidas de avaliação. A Tabela 5.6, ilustra o resultado das medidas citadas para o classificador Sexual Explícito.

Acurácia	Precisão	Revocação	Medida F
77,70%	80,43%	58,73%	67,88%

Tabela 5.6 – Medidas de Avaliação - Classificador Sexual Explícito

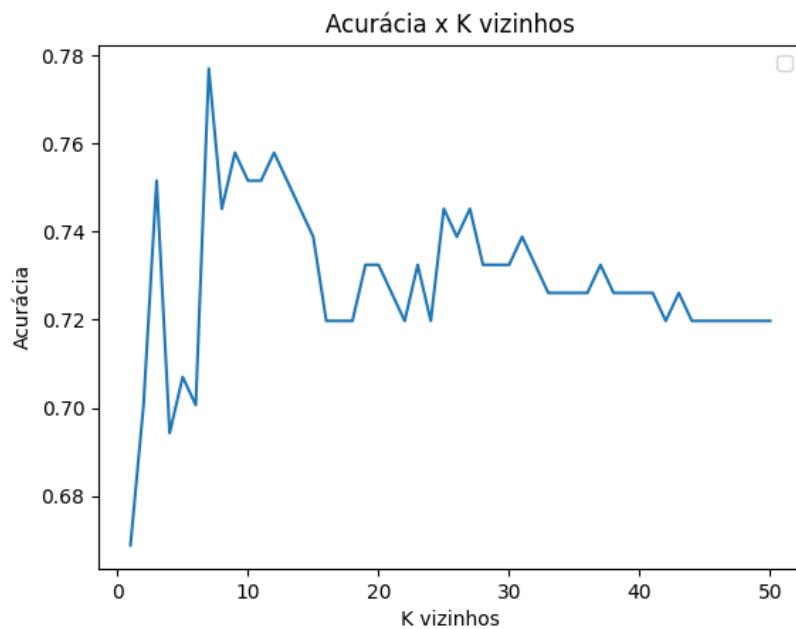


Figura 5.18 – Acurácia x K vizinhos - Classificador Sexual Explícito

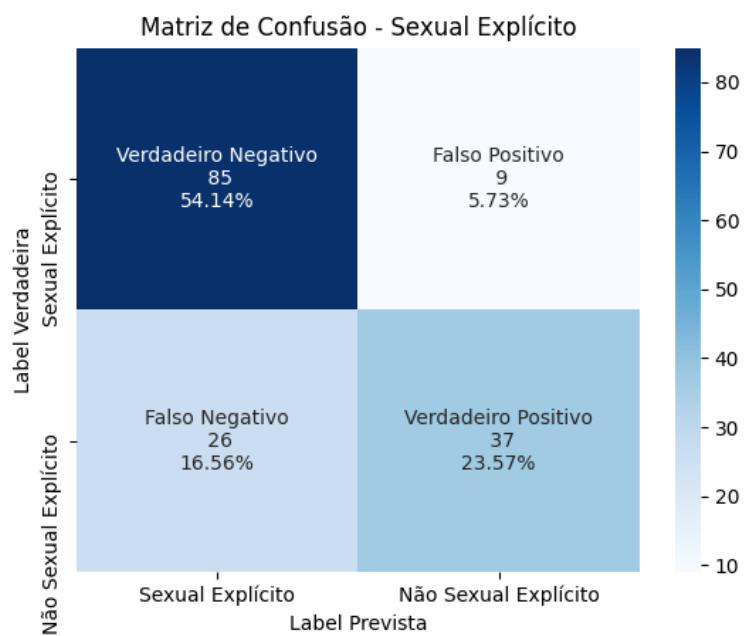


Figura 5.19 – Matriz de confusão para Classificador Sexual Explícito

6. RESULTADOS E DISCUSSÕES

Aqui estão descritos os resultados das atividades realizadas durante o desenvolvimento do Trabalho de Conclusão, e demais discussões pertinentes ao tema de viés de gênero no contexto de agentes virtuais.

6.1 Corpus Anotado

Tendo em vista todas as etapas de captação, limpeza, criação e anotação do corpus descritas no Capítulo 5 - Classificador de toxicidade, chegamos ao seguinte funil de total de mensagens (Figura 6.1):

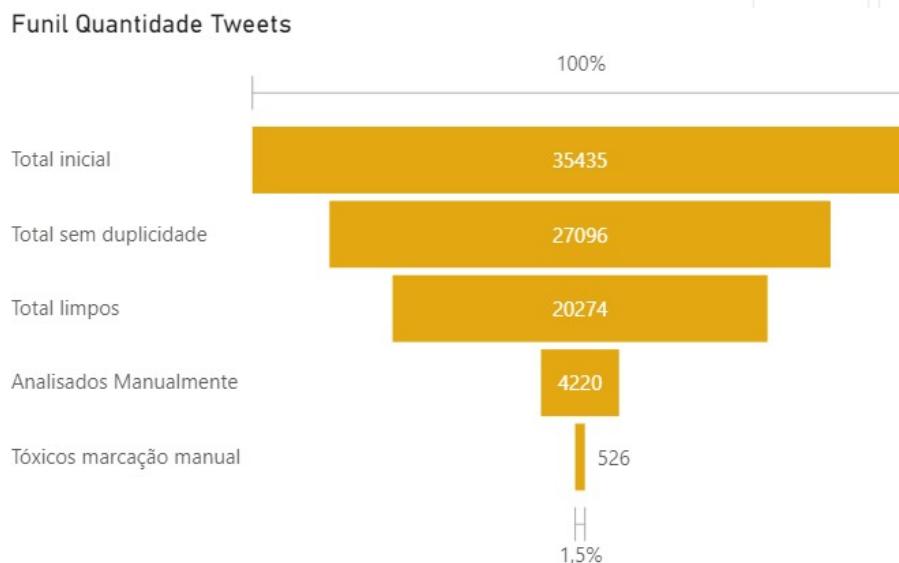


Figura 6.1 – Funil de *tweets*: coleta até anotações manuais

Considerando a base trabalhada (total limpos 20.274 *tweets*, sem duplicidade e sem termos irrelevantes), foi selecionada uma parcela dessa base para análise e anotação manual dos atributos tóxicos para as mensagens cabíveis, assim sendo, temos uma composição de 4.220 *tweets* analisados manualmente, dos quais temos uma proporção de 12.46% mensagens identificadas como tóxicas. Desta forma, entendemos que as mensagens com algum nível de toxicidade são sim uma parcela significativa da base, pois se fôssemos projetar esta porcentagem na quantidade total de mensagens, esta proporção representaria 2.527 mensagens tóxicas direcionadas ao agente virtual, de um total de pouco mais de 20 mil *tweets* limpos.

O painel da Figura 6.2 apresenta as proporções supracitadas, sendo subdivididas por agente virtual. Aqui, trazemos à tona um ponto ainda mais relevante: a proporção de mensagens tóxicas para a agente virtual feminina representa mais de 70% dos xingamentos

identificados, com um total que é mais que o dobro das mensagens tóxicas identificadas para o agente masculino (376 para Magalu, contra 150 para Baianinho). Com isto, nossa hipótese central é validada: **existe diferença no tratamento entre agentes virtuais de diferentes gêneros, pois os agentes virtuais femininos são mais xingados.**



Figura 6.2 – Totais de *tweets* analisados, tóxicos e proporção por agente virtual

Além disso, através das anotações manuais que os autores fizeram no corpus, baseados nos atributos disponibilizados pelo *Perspective*, foi possível levantar e contabilizar os diferentes tipos de mensagens tóxicas enviadas para os agentes. A Figura 6.3 ilustra a quantidade dessas anotações feitas para cada classe/atributo.



Figura 6.3 – Total de *tweets* marcados como tóxico, e demais atributos de toxicidade e devidas quantidades

Não obstante, com a análise quantitativa e visual das mensagens tóxicas enviadas para cada agente virtual, vide Figura 6.4, pudemos comparar as quantidades e tipos de xingamentos, assim, mais diferenças relacionadas a gênero ficam evidentes:

- Como já discorrido, houveram mais marcações de mensagens tóxicas para a agente feminina do que para o masculino, porém a proporção de Toxidade Severa foi seme-

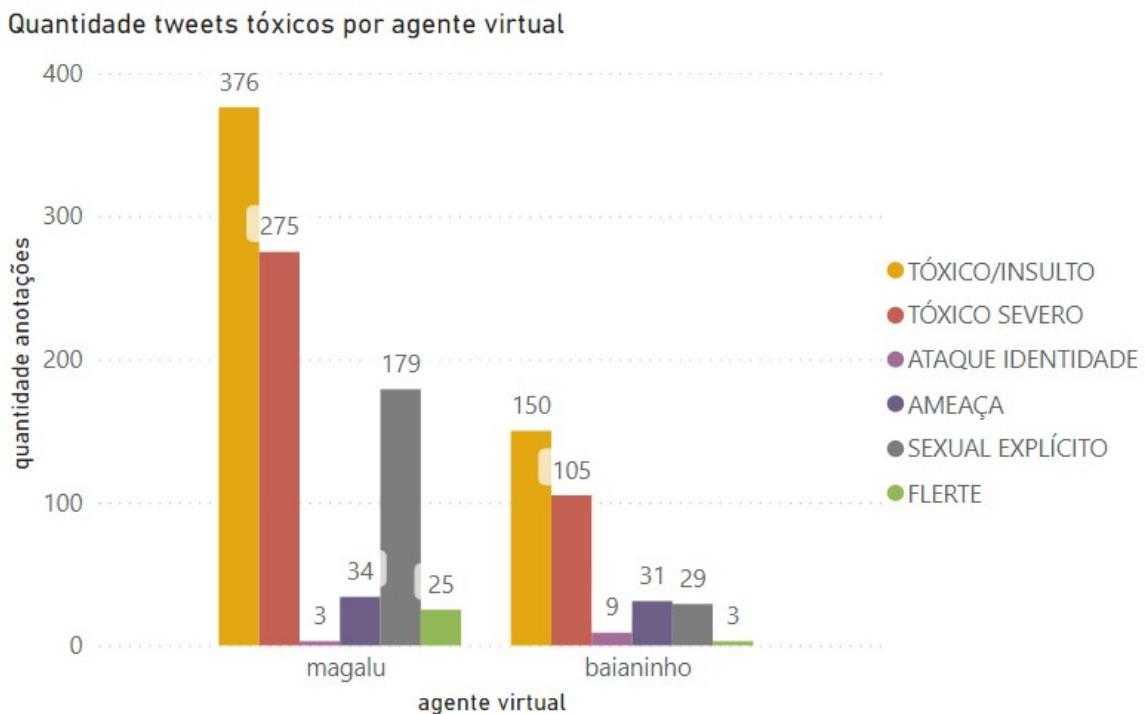


Figura 6.4 – *Tweets* tóxicos por agente virtual, com atributos de toxicidade e devidas quantidades

lhante para ambos os agentes (73% de marcações com toxicidade severa para Magalu, contra 70% para o Baianinho).

- Com relação ao atributo de Ataque de Identidade, houveram poucas marcações deste tipo (12 no total), todavia tais foram com conotações diferentes para cada agente: este tipo de xingamento foi direcionado ao agente Baianinho criticando sua etnia (em sua maioria, pontuando que o CB deixou de aparentar ser baiano e passou a parecer carioca, e o xingando por isso), enquanto que as mensagens com este tipo de anotação direcionadas à agente Magalu se referiam de forma negativa a uma possível sexualidade da agente, a chamando de lésbica e xingando.
- Com relação a anotação de Ameaça, embora a quantidade para o agente Baianinho seja menor (31 para ele, sendo 34 para a agente Magalu), tal tem maior incidência proporcional, tendo em vista que o total de mensagens tóxicas para ele foram 150; assim sendo, o Baianinho obteve 20,66% das marcações de toxicidade sendo ameaças, contra apenas 9% para a Magalu. Assim, comprovamos que os **xingamentos para agentes virtuais masculinos são mais agressivos**.
- Quando falamos do atributo de conotação Sexual Explícita, tal tem os valores com diferença mais expressiva: 179 anotações para a agente feminina, Magalu, contra somente 29 mensagens deste tipo para o agente masculino, Baianinho. Ou seja, 47% dos xingamentos direcionados à Magalu tem conotação sexual, contra 19% para o Baianinho. Isso demonstra o quanto esse tipo de agressão é mais recorrente com

o gênero feminino, e valida nossa hipótese de que **xingamentos a agentes virtuais femininos tem uma maior conotação sexual**.

- Por último, a anotação para mensagens com Flerte/Assédio também tem uma grande diferença de um agente estudado para o outro: 25 *tweets* com esta marcação para a Magalu, contra apenas 3 para o Baianinho. Embora não seja uma interação tão frequente para ambos, a incidência foi mais de 8 vezes maior para a gente feminina. Assim, tendo em vista que analisamos este tipo de mensagem sob o viés de que é uma "mensagem de flerte para com o agente ou elogios incabíveis quando se pensa que o agente é como uma pessoa em seu ambiente de trabalho, configurando assédio", olhando para estas quantidades validamos nossa hipótese de que **os elogios aos agentes virtuais falam sobre sua aparência, tornando-se assédio**.

A Tabela 6.1 apresenta uma demonstração de *tweets* com cada uma das anotações supracitadas, com um exemplar para cada atributo e para cada agente.

Agente Virtual	Atributo Anotado	Toxidade	Tweet
baianinho	Ataque de Identidade		@casasbahia vai se foder vc nem parece baiano desgracado, vc e a perra de um carioca
magalu	Ataque de Identidade		@magazineluiza se e lesbica ne , tem cara de lesbica , quase ctz q e lesbica , eu posso jurar q e tipo pqp lesbica 100% tipo juro mesmo
baianinho	Ameaça		morte ao novo baianinho
magalu	Ameaça		@magazineluiza vai se foder magalu!! tu nem existe e eu nao posso desejar a tua morte??? nao fode perra da outro ban aqui quero ver!! link
baianinho	Sexual Explícito		chupa minha pica cb
magalu	Sexual Explícito		mostra os peitos virtuais ai rapidao vagabunda
baianinho	Flerte/Assédio		baianinho safado, irei salvar para uns estudos mais tarde
magalu	Flerte/Assédio		@magazineluiza lu ta gostosinha de preto

Tabela 6.1 – Exemplos de *tweets* com diferentes tipos de toxidade

Por fim, de forma a visualizar o processo completo de criação do corpus, consolidamos o fluxograma 6.5, o qual demonstra visualmente as etapas necessárias para construção deste. Além disso, no pipeline 5.15 demonstramos visualmente como tal corpus foi utilizado para treinamento de um classificador do tipo kNN.

Concluindo os resultados, na figura 6.6 apresentamos o corpus finalizado: das 4220 mensagens analisadas manualmente, 526 delas foram consideradas tóxicas. Nestas

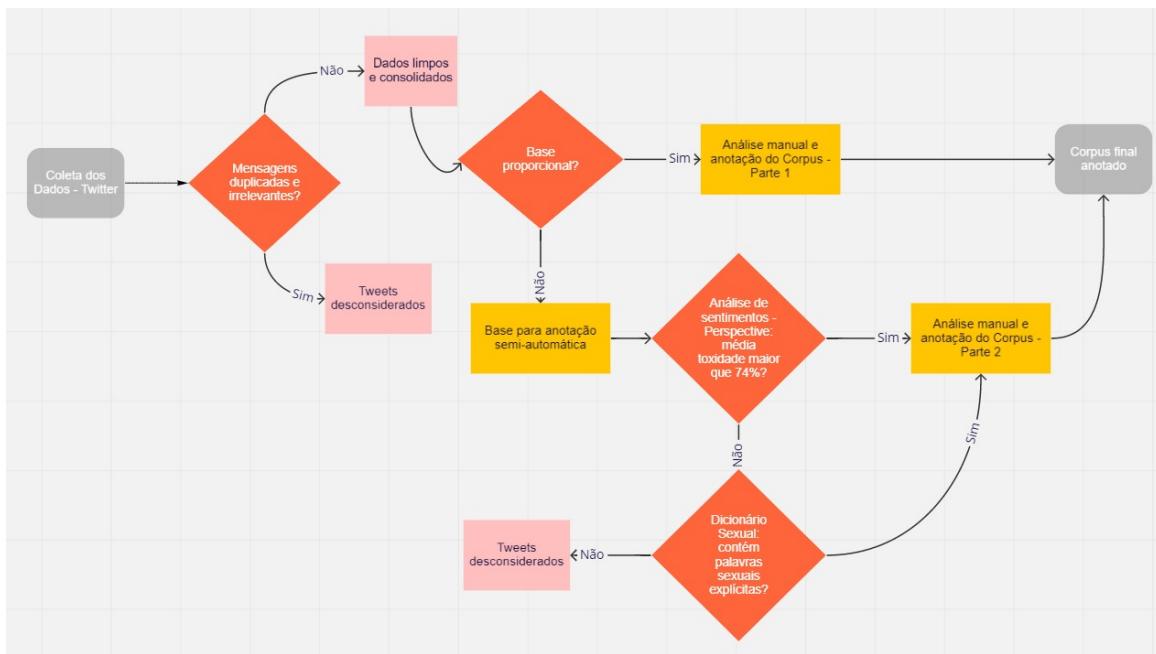


Figura 6.5 – Fluxograma de construção e anotação do corpus.

foram atribuídos diferentes níveis de toxicidade, e tais representam as proporções do total analisado apresentadas na imagem.



Figura 6.6 – Total *tweets* analisados, e totais dos atributos de toxicidade anotados.

6.2 Discussão

Como já citado no capítulo 2, optamos por analisar assistentes virtuais sob uma perspectiva de gênero pois tal análise é raramente feita, todavia buscamos no presente trabalho enfatizar não somente assistentes virtuais de forma ampla, como feito no estudo da UNESCO, "I'd Blush If I Could", mas sim dando enfoque a problemática que envolve especificamente agentes virtuais, tendo em vista sua crescente adesão e popularização no Brasil [19][17][10]. Desta forma, analisamos uma problemática bastante explícita nas redes sociais, onde estes agentes recebem diversos tipos de xingamentos e comentários tóxicos, como já demonstrado no presente estudo, e sem que haja algum tipo de regulamentação quanto a isso, além das políticas da própria rede social, neste caso, o Twitter [1].

Focamos em analisar sob a binariedade de gênero, porém a grande maioria dos agentes virtuais, assim como os chatbots e assistentes de voz, são femininos. Tivemos dificuldade para encontrar um agente masculino para usarmos como comparação, pois o único agente com relevância masculino no Brasil até 2020 era o CB (Baianinho das Casas Bahia), contra várias agentes femininas (Magalu da Magazine Luiza, Nat da Natura, Vivi da Vivo, Carina do Carrefour, Dai da Dailus, Elô da Cielo, e a mais recente Sam da Samsung [44]). Com nosso estudo sobre interações com agentes virtuais no Twitter, foi possível verificar que, em concordância com Mertens, as mulheres são assediadas desproporcionalmente nesta plataforma por causa de seu gênero [39], mesmo estas não sendo mulheres reais: as agentes virtuais, por terem corpos feitos em computação gráfica, sofrem assédio direcionado, vide Figura 6.7.

Todavia, durante o período de elaboração deste trabalho (2020-2021), outro caso veio à público e trouxe esta discussão à tona: a Inteligência Artificial do Bradesco, a chatbot BIA, recebeu, apenas em 2020, mais de 95 mil mensagens com referências tóxicas, num teor de assédio à feminilidade [101]. Segundo a marca:

A BIA não é uma mulher real, ela é uma inteligência artificial, mas também sofre assédio, e isso acontece porque ela é composta por elementos femininos. Assim a violência também é baseada no gênero. A omissão a esse tipo de ofensa só colabora para que o assédio seja visto como algo natural [15].

À vista disso, o banco criou um projeto para barrar tais comportamentos, atualizando às respostas da inteligência artificial para que ela reaja de forma justa e firme contra o assédio. "Sem meias palavras. Sem submissão.", afirma a financeira. Da mesma forma, a UNESCO lançou em 2020 o movimento "HeyUpdateMyVoice"[98], incentivando marcas a mudarem a narrativa de suas inteligências artificiais, como assistentes de voz e chatbots, para que tais passem a responder mensagens de assédio de forma mais assertiva, tendo



Figura 6.7 – Tweets tóxicos direcionados à agentes virtuais diversos

em vista que tais foram programadas para responder da maneira mais educada possível, e muitas vezes foram feitas mulheres pelo fato dos desenvolvedores acreditarem que isto traria um maior conforto para o público geral.

Assim sendo, comprova-se que tal problemática abrange todos os tipos de assistentes virtuais também para o público brasileiro, e atinge especialmente as assistentes femininas, abrindo precedente para que tais comportamentos se deem para mulheres em geral. Então, questionamos: se temos tantos problemas relacionados a comunicação violenta direcionada as assistentes femininas, por que as marcas ainda escolhem direcionar a aparência e voz de seus assistentes desta forma?

Sabemos que muito se dá devido ao estudo do livro *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship* ("Conectado pelo discurso: como a voz ativa e avança a relação humano-computador", em tradução livre), no qual Clifford Nass, professor da Universidade Stanford, discorre sobre como a voz feminina é tida como prestativa, e a partir disso se criou uma tendência mercadológica.

Todavia, como já citado na subseção 2.2.1, alguns estudos já refutam essa ideia, como pontua o artigo de Anderson, R.A., cuja pesquisa apontou preferência da maioria das pessoas com relação à fala masculina em tom baixo [20]; também Mitchell W. et al estudou que existe preferência pelo som de uma voz masculina quando está fazendo declarações oficiais [66]; e não obstante Stromberg, J., que salienta que as pessoas em geral preferem a voz do sexo oposto [96]. Mas mudar esse paradigma não é tão simples, como pontua Andreea Danilescu:

"A redução do preconceito de gênero em assistentes é algo que vale a pena ser buscado, mas levanta uma série de questões práticas, culturais e éticas que precisamos abordar primeiro. Do ponto de vista prático, como uma voz de gênero neutro afeta a adoção e como podemos projetar uma personalidade de gênero neutro para corresponder a essa voz? Culturalmente, como fazemos o design para idiomas genéricados? E, eticamente, um assistente de voz de gênero neutro pode realmente mudar as normas de gênero ou, em última instância, irá inquietar os usuários? Essas são apenas algumas das perguntas que precisamos fazer enquanto buscamos esse objetivo. [40]"

Mas, mesmo com os levantamentos, Andreea Danilescu questiona: *"A exposição prolongada a assistentes de voz que não seguem as normas de gênero mudaria a percepção do usuário ao longo do tempo?"*. Os autores do presente trabalham acreditam que sim, pois, tendo em vista que os estudos de Nass datam do final dos anos 1990 e começo de 2000, muito foi construído e desenvolvido buscando maior equidade de gênero, inclusive no que tange tecnologia. Assim sendo, acreditamos que a manutenção deste *status quo* se dá muito por viés de confirmação, o qual é um tipo de viés inconsciente que se dá quando nossa atenção se torna seletiva, tendo em vista informações que estão de acordo com nossas ideias [13]. Em outras palavras, o psicólogo Peter Wason descobriu nos anos 1960 que temos a tendência a buscar informações que confirmem nossas crenças e opiniões, e descartamos aquelas que não o fazem. Isso afeta, também, os dados dos quais lembramos e a credibilidade que damos a leituras que fazemos, sendo este viés considerado o "inimigo da ciência"[41].

Dentro desta perspectiva, também conseguimos analisar o caso dos agentes virtuais brasileiros: mesmo o Baianinho sendo um agente masculino, ele é o segundo mais popular no país, com 191,6 mil seguidores no Twitter, ficando atrás apenas da Magalu, que conta com 1,3 milhões de seguidores. E, como comprovamos durante o presente estudo, este apresenta um número muito menor de xingamentos quando comparado à agente feminina, especialmente mensagens ofensivas de cunho sexual e assédio, que são um dos principais problemas a serem solucionados.

Desta forma, em concordância com o exposto pela UNESCO em ambos os trabalhos (*I'd Blush If I Could* e *HeyUpdateMyVoice*), os autores acreditam que mudanças na forma como estes assistentes são projetados são necessárias e bastante urgentes, de forma a mitigar comportamentos nocivos, principalmente tendo em vista que tais comportamentos não se limitam a redes sociais. Como Laurie Penny pontua em "CIBERSEXISMO: sexo, gênero e o poder na internet", em tradução livre:

"Uma sociedade em rede é tão boa quanto as redes nas quais ela é construída. Uma rede que desumaniza as mulheres (...) simplesmente não é uma rede que funciona corretamente, e geeks, nerds e todos que se preocupam com a Internet como um espaço livre e aberto precisam entender que sua rede não é mais adequada para o propósito. Nosso sistema está quebrado. Isto precisa de ser atualizado. [75]"

Uma forma de reduzir esta problemática é incentivar a participação feminina na construção dessas narrativas, sejam elas na criação de agentes virtuais com gênero neutro, sejam elas no que abrange inteligência artificial conversacional. Apesar da crescente influência deste tipo de tecnologia, as mulheres representam apenas 12% dos pesquisadores de IA, de acordo com pesquisas da Element AI e da revista Wired [54]. Quanto a neutralidade de gênero, no estudo "Evitando estereótipos de gênero em assistentes de voz para promover a inclusão" (*Eschewing Gender Stereotypes in Voice Assistants to Promote Inclusion*, em tradução livre), Danilescu propõe um sintetizador de voz não binária, o qual está em execução enquanto o presente estudo foi escrito, e também visa aprofundar o estudo sobre abordagens para reduzir o viés em modelos de reconhecimento de fala existentes. Os autores do presente estudo aguardam ansiosos por suas constatações para pensarmos em como podemos mudar as normas sociais com esta tecnologia, e expandir esta neutralidade para o âmbito de agentes virtuais.

Por fim, acreditamos no disposto por Laurie Penny:

"Assim que a comunidade geek finalmente perceber que o assédio, o bullying e a intimidação de mulheres online são uma clara ameaça aos princípios da liberdade de expressão e igualitarismo, o espaço social da Internet começará a parecer muito diferente. Os meninos crescem acreditando que são os heróis de sua própria história; as meninas precisam aprender a não se ver como personagens coadjuvantes na saga de outra pessoa. Felizmente, a Internet permite que você escolha sua própria aventura. Os sistemas podem ser reescritos. Protocolos atualizados. A arquitetura social que estamos construindo online hoje será aquela em que a próxima geração crescerá (...). É hora de virar o jogo. A revolução de gênero e a revolução digital estão acontecendo juntas e assustam as mesmas pessoas pelos motivos certos. O sistema se adapta e podemos reescrevê-lo para que funcione melhor - ou podemos torná-lo uma sala de jogos para os preconceitos do passado. Depende de nós. [75]"

7. CONCLUSÃO DO ESTUDO E CONTRIBUIÇÕES

De acordo com os objetivos do trabalho, foram levantados e observados dados sobre dois agentes virtuais de diferentes gêneros, de forma a identificar e validar se existem diferenças na comunicação do grande público para com estes personagens em contextos de comunicação web.

Foram levantados mais de 35mil *tweets*, para os quais foram feitas limpezas, chegando a uma base utilizável de 20.274 mensagens. Para criarmos um corpus a partir destas, foi utilizada tanto uma proporção das mensagens, quanto a API identificadora de comentários tóxicos da Google, *Perspective*. Com o uso desta aumentamos a assertividade de análise, tendo em vista que captamos para análise manual apenas as mensagens que a ferramenta já havia considerado como tóxicas, vide subseção 5.4.3.

Desta forma, onde antes havíamos trabalhado sobre uma proporção da base (vide subseção 5.4.1), e obtivemos uma assertividade de 4% (129 *tweets* tóxicos para uma base de 3.214 analisados manualmente), com o uso da ferramenta esta assertividade passou a ser de 39% (397 mensagens tóxicas identificadas para uma base de 1.006 analisadas). Assim, entendemos que tal ferramenta é muito útil para análise de sentimento, e por ser de fácil adesão (uma API facilmente implementável e sem custo), julgamos que tal é a ideal para este tipo de análise, tendo em vista principalmente que nosso idioma é o português e tal não é tão difundido em ferramentas deste tipo.

Com o uso da ferramenta *Perspective*, consolidamos um corpus com 1.007 mensagens em português com anotações de toxicidade, sendo esta uma das contribuições do presente estudo. Nosso corpus construído conta com anotações em português que não encontramos na literatura (Sexual Explícito e Flerte/Assédio), e a partir deste utilizamos o algoritmo de aprendizado de máquina kNN para executar a tarefa de classificação de toxicidade em dois níveis (Toxicidade e Conotação Sexual Explícita).

Além disso, através da análise manual e visual do corpus via ferramenta *Power BI* da Microsoft, pudemos discorrer análises quanto às mensagens que cada agente recebe. Com as anotações manuais de toxicidade, pudemos levantar números quanto a toxicidade dessas comunicações, e trazemos como ponto relevante dos valores levantados a quantidade de mensagens com conotação sexual enviadas para a agente Magalu (179), quantidade 6 vezes maior do que a quantidade anotada para o agente Baianinho (apenas 29 mensagens deste tipo). Isto denota, em concordância com os autores Connell e Pearse, o quanto “o gênero é uma dimensão central na vida pessoal, das relações sociais e da cultura” e “o mundo se depara hoje com problemas urgentes ligados ao gênero” [36].

Com isto, trazemos para a academia uma temática pouco discutida, viés de gênero no que tange agentes virtuais, e validamos que existe diferença no tratamento entre agentes virtuais de diferentes gêneros. Através das análises já citadas, tivemos hipóte-

ses preocupantes validadas: agentes virtuais femininos são mais xingados, xingamentos a agentes virtuais femininos tem uma maior conotação sexual, os elogios aos agentes virtuais falam sobre sua aparência, tornando-se assédio, e os xingamentos para agentes virtuais masculinos são mais agressivos.

Também, com o objetivo de disseminar o conhecimento referente ao tema tratado neste trabalho e expandir a utilização dos classificadores aqui construídos, disponibilizamos um repositório de código aberto no Github¹ em que o presente estudo, assim como suas contribuições, podem ser acessados por qualquer pessoa de forma a incrementar o trabalho iniciado neste artigo.

Através dos levantamentos supracitados, constatamos que mudanças no que tange a criação e manutenção deste tipo de tecnologia precisam ser feitas, e a temática revista, se possível com outros olhos, para que não criemos mais padrões tecnológicos que reforcem padrões de gênero excludentes e tóxicos. Assim como Bertolt Brecht, acreditamos que:

"nada deve parecer natural, nada deve parecer impossível de mudar". [29].

7.1 Trabalhos Futuros

No decorrer do desenvolvimento deste trabalho, foram identificados alguns pontos de melhoria, abrindo oportunidade para trabalhos futuros, destacando:

- **Tamanho do corpus:** Como discutido no decorrer deste trabalho, o corpus final apresenta um tamanho consideravelmente pequeno e assim limitando a qualidade dos classificadores para novas entradas em que ele não conheça os dados. Será útil para futuras classificações um aumento no número de instâncias, principalmente as instâncias de conotação Sexual Explícito, a fim de expandir esse tipo de classificação para a língua portuguesa.
- **Avaliadores do corpus:** Conforme visto na literatura, é interessante termos uma grade multidisciplinar de profissionais avaliando o corpus que está sendo construído [12][109]. Desta forma, entendemos que, para aumentar essa base de dados, será interessante ter mais avaliadores para contribuir com as anotações, e que tais sejam de diferentes áreas, como psicologia e sociologia, tendo em vista analisar mais a fundo questões de viés de gênero.
- **Novas técnicas de pré-processamento:** Há espaço para melhorias nas palavras sem valor aos classificadores *stop words*, com a criação de uma lista contendo uma

¹https://github.com/kalissa/TCCII_vies_agentes

base de nome de produtos de varejo, pois foi notado durante as análises manuais que muitas mensagens são apenas de produtos ou fazem referência a tais produtos. Também, aqui, é valido complementar o dicionário utilizado com palavras de cunho sexual, de forma a abranger mais termos em português.

- **Novos algoritmos:** Utilizar outros algoritmos de classificação tais como Naive Bayes (NB), Máquina de Suporte Vetorial (SVM), Entropia Máxima e *Multinomial Naive Bayes* (MNB). Há, também, na medida que o corpus cresça, a possibilidade de utilizar o vetorizador *TF-IDF*, além da possibilidade de utilização de um modelo de aprendizado com Redes Neurais.
- **Novos Agentes Virtuais:** Tendo em vista que no presente trabalho apenas dois agentes foram analisados, há espaço para ampliar a análise para outros agentes existentes no *twitter* e para agentes de empresas que atuem em outras áreas do mercado. No que tange a anotação de Ataque de Identidade, a agente virtual da marca Natura é uma assistente negra, cuja análise pode vir a ser explorada.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] “As regras do twitter”. Capturado em: <https://help.twitter.com/pt/rules-and-policies/twitter-rules>, Maio 2021.
- [2] “Attributes languages”. Capturado em: <https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>, Maio 2021.
- [3] “How to join the conversation on twitter: Reply to tweets to add your voice”. Capturado em: <https://help.twitter.com/en/twitter-guide/topics/how-to-join-the-conversation-on-twitter/how-to-reply-to-a-tweet-on-twitter>, Março 2021.
- [4] “Microsoft cognitive services text analytics”. Capturado em: <https://azure.microsoft.com/pt-br/services/cognitive-services/text-analytics/#features>, Outubro 2020.
- [5] “Microsoft power bi”. Capturado em: <https://powerbi.microsoft.com/pt-br/>, Dezembro 2020.
- [6] “Perspective”. Capturado em: <https://www.perspectiveapi.com/how-it-works/>, Outubro 2020.
- [7] “Tweepy”. Capturado em: <https://www.tweepy.org/>, Outubro 2020.
- [8] “Twitter”. Capturado em: <https://about.twitter.com/en/who-we-are/our-company>, Novembro 2020.
- [9] “Magazine luiza - nossa estratégia”. Capturado em: <https://ri.magazineluiza.com.br>ShowCanal/Nossa-Estrategia?=LZKRKYC4fKjk6oPPJL7+xw==>, Maio 2021.
- [10] “Conheça a nat, a assistente virtual da natura”. Capturado em: <https://www.natura.com.br/blog/mais-natura/conheca-a-nat-a-assistente-virtual-da-natura>, Maio 2021.
- [11] “Brasil é 2º em ranking de países que passam mais tempo em redes sociais”. Capturado em: <https://epocanegocios.globo.com/Tecnologia/noticia/2019/09/brasil-e-2-em-ranking-de-paises-que-passam-mais-tempo-em-redes-sociais.html>, Maio 2021.
- [12] “Jigsaw unintended bias in toxicity classification”. Capturado em: <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data>, Maio 2021.
- [13] “Viés de confirmação: o que é, como funciona?” Capturado em: <https://www.psicanaliseclinica.com/vies-de-confirmacao-o-que-e-como-funciona/>, Junho 2021.

- [14] "Bia, lu e alexa: assistentes virtuais de diferentes marcas são assediadas". Capturado em: <https://exame.com/casual/bia-lu-e-alexa-assistentes-virtuais-de-diferentes-marcas-sao-assediadas/>, Maio 2021.
- [15] "Novas respostas da bia contra o assédio". Capturado em: <https://banco.bradesco/aliadosbia/>, Março 2021.
- [16] "Personas virtuais: o novo mercado de influência". Capturado em: <https://infobase.com.br/personas-virtuais-o-novo-mercado-de-influencia/>, Maio 2021.
- [17] "Sam: assistente virtual da samsung ganha visual repaginado". Capturado em: <https://www.tudocelular.com/tech/noticias/n175182/sam-assistente-virtual-samsung-novo-visual.html>, Maio 2021.
- [18] "The state of influencer marketing 2020: Benchmark report". Capturado em: <https://influencermarketinghub.com/influencer-marketing-benchmark-report-2020/>, Maio 2021.
- [19] "Varejista pontofrio elimina o "frio" da marca (mas mantém vivo o pinguim)". Capturado em: <https://blogs.oglobo.globo.com/capital/post/varejista-ponto-frio-elimina-o-frio-mas-mantem-vivo-o-pinguim.html>, Maio 2021.
- [20] Anderson, R.; Klofstad, C. "Preference for leaders with masculine voices holds in the case of feminine leadership roles", *PLoS one*, vol. 7, 12 2012, pp. e51216.
- [21] Bajjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. "Deep learning for hate speech detection in tweets", *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 2017.
- [22] Baklanov, N. "The top instagram virtual influencers in 2019". Capturado em: <https://hypeauditor.com/blog/the-top-instagram-virtual-influencers-in-2019/>, Outubro 2020.
- [23] Baklanov, N. "The top instagram virtual influencers in 2020". Capturado em: <https://hypeauditor.com/blog/the-top-instagram-virtual-influencers-in-2020/>, Maio 2021.
- [24] Barbosa, V. "Até a mascote virtual do magazine luiza é alvo de assédio sexual". Capturado em: <https://exame.com/marketing/ate-a-mascote-virtual-do-magazine-luiza-e-alvo-de-assedio-sexual/>, Setembro 2020.
- [25] Bhuiyan, T.; Xu, Y.; Josang, A. "State-of-the-art review on opinion mining from online customers' feedback", 2009.
- [26] Blair, I.; Rev, P. "The malleability of automatic stereotypes and prejudice", *Personality and Social Psychology Review*, vol. 6, 08 2002, pp. 242–261.

- [27] Boiy, E.; Moens, M. F. “A machine learning approach to sentiment analysis in multilingual web texts”, *Inf. Retr.*, vol. 12, 10 2009, pp. 526–558.
- [28] Bolukbasi, T.; Chang, K. W.; Zou, J.; Saligrama, V.; Kala. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 4356–64.
- [29] Brecht, B. “Nada é impossível de mudar”, *Stylus (Rio de Janeiro)*, 11 2016, pp. 293 – 293.
- [30] Brynjolfsson, E.; Hitt, L.; Kim, H. “Strength in numbers: How does data-driven decision-making affect firm performance?”, *International Conference on Information Systems 2011, ICIS 2011*, vol. 1–1, 2011, pp. 541–558.
- [31] Buolamwini, J.; Gebru, T. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, Friedler, S. A.; Wilson, C. (Editores), 2018, pp. 77–91.
- [32] Burger, J.; Henderson, J. “An exploration of observable features related to blogger age”. In: Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006, 2006, pp. 15–20.
- [33] Burger, J. D.; Henderson, J.; Kim, G.; Zarrella, G. “Discriminating gender on twitter”. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1301–1309.
- [34] Cambria, E. “Affective computing and sentiment analysis”, *IEEE Intelligent Systems*, vol. 31–2, 2016, pp. 102–107.
- [35] Celestino, R.; Souza, E. D. “Inteligência artificial no varejo: Casos de uso e o fenômeno magalu”. Capturado em: <https://www.ibm.com/downloads/cas/BQ5M15RN>, Setembro 2020.
- [36] Connell, R; Pearse, R. “Gender in world perspective”, 06 2015, pp. 85–88.
- [37] Cover, T.; Hart, P. “Nearest neighbor pattern classification”, *IEEE Transactions on Information Theory*, vol. 13–1, 1967, pp. 21–27.
- [38] Crawford, K. “Artificial intelligences: White guy problem”. New York Times, 2016.
- [39] Cuthbertson, L.; Kearney, A.; Dawson, R.; Zawaduk, A.; Cuthbertson, E.; Gordon-Tighe, A.; Mathewson, K. “Women, politics and twitter: Using machine learning to change the discourse”. 1911.11025, 2019.

- [40] Danilescu, A. "Eschewing gender stereotypes in voice assistants to promote inclusion". In: Proceedings of the 2nd Conference on Conversational User Interfaces, 2020.
- [41] Espindola, E. "Viés de confirmação: O inimigo da ciência". Capturado em: <https://fbinck.com/vies-de-confirmacao-o-inimigo-da-ciencia/>, Junho 2021.
- [42] Etim, B. "Approve or reject: Can you moderate five new york times comments?" Capturado em: <https://www.nytimes.com/interactive/2016/09/20/insider/approve-or-reject-moderation-quiz.html>, Março 2021.
- [43] Faceli, K.; Lorena, A. C.; Gama, J.; Carvalho, A. C. P. d. L. F. d. "Inteligência artificial: uma abordagem de aprendizado de máquina". LTC, 2011.
- [44] Fernandez, A. "Marcas apostam nas assistentes virtuais. qual a sua preferida?" Capturado em: <https://gkpb.com.br/63212/marcas-assistentes-virtuais/>, Junho 2021.
- [45] Fontoura, D. "Predição de falhas em projetos de software livre baseada em métricas de redes sociais." Campo Mourão, Paraná, Brasil: Universidade Tecnológica Federal do Paraná, 2011, pp. 51.
- [46] Fragos, K.; Maistros, Y.; Skourlas, C. "A weighted maximum entropy language model for text classification." In: Natural Language Understanding and Cognitive Science, 2005, pp. 55–67.
- [47] Gagliardone, I. "Countering Online Hate Speech". United Nations Educational, Scientific, and Cultural Organization, 2015.
- [48] Glick, P.; Fiske, S. "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism", *Journal of Personality and Social Psychology*, vol. 3, 03 1996, pp. 491–512.
- [49] Habler, F.; Schwind, V.; Henze, N. "Effects of smart virtual assistants' gender and language". In: Proceedings of Mensch Und Computer 2019, 2019, pp. 469–473.
- [50] Hanu, L.; Unitary team. "Detoxify". Outubro 2020.
- [51] Hanu, L; Thewlis, L. H. S. "How AI Is Learning to Identify Toxic Online Content". Scientific American, a Division of Springer Nature America, Inc., 2021.
- [52] Hempel, J. "Siri and Cortana Sound Like Ladies Because of Sexism". Wired, 2015.
- [53] Hu, M.; Liu, B. "Mining opinion features in customer reviews". In: Proceedings of the 19th national conference on Artifical intelligence (AAAI 2004), 2004, pp. 755–760.

- [54] Hudson, S. “Estimating the gender ratio of ai researchers around the world”. Capturado em: <https://medium.com/element-ai-research-lab/estimating-the-gender-ratio-of-ai-researchers-around-the-world-81d2b8dbe9c3>, Junho 2021.
- [55] Jigsaw. “Google’s jigsaw announces toxicity-reducing api, perspective, is processing 500m requests daily”. Capturado em: <https://www.prnewswire.com/news-releases/googles-jigsaw-announces-toxicity-reducing-api-perspective-is-processing-500m-requests-daily.html>, Maio 2021.
- [56] Klein, S.; Simmons, R. F. “A computational approach to grammatical coding of english words”, *J. ACM*, vol. 10–3, Jul 1963, pp. 334–347.
- [57] Koetsier, J. “We’ve spent 1.6 trillion hours on mobile so far in 2020”. Capturado em: <https://www.forbes.com/sites/johnkoetsier/2020/08/17/weve-spent-16-trillion-hours-on-mobile-so-far-in-2020/?sh=705c9a5f6d61>.
- [58] Leavy, S. “Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning”. In: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, 2018, pp. 14–16.
- [59] LEXICO. “Dicionário online de português”. Capturado em: <https://www.lexico.pt/post/>, Novembro 2020.
- [60] Liddy, E. “Natural language processing.” New York, NY, USA: Encyclopedia of Library and Information Science, 2001.
- [61] Liu, B. “Sentiment Analysis: Mining Opinions, Sentiments, and Emotions”. Cambridge University Press, 2015, pp. 1–367.
- [62] Medeiros, A. “Casas bahia reformula seu personagem “baianinho””. Capturado em: <https://ecommerceedesucceso.com.br/casas-bahia-baianinho>, Junho 2021.
- [63] Medhat, W.; Hassan, A.; Korashy, H. “Sentiment analysis algorithms and applications: A survey”, *Ain Shams Engineering Journal*, vol. 5, 05 2014.
- [64] Mertens, A.; Pradel, F.; B. Rozyjumayeva, A.; Wäckerle, J. “As the tweet, so the reply? gender bias in digital communication with politicians”. In: Proceedings of the 10th ACM Conference on Web Science, 2019, pp. 193–201.
- [65] Mitchell, T. M. “Machine learning, International Edition”. McGraw-Hill, 1997.
- [66] Mitchell, W.; Ho, C.-C.; Patel, H.; MacDorman, K. “Does social desirability bias favor humans? explicit–implicit evaluations of synthesized speech support a new hci model of impression management”, *Computers in Human Behavior*, vol. 27, 01 2011, pp. 402–412.

- [67] Montero, C. S.; Munezero, M.; Kakkonen, T. "Investigating the role of emotion-based features in author gender classification of text". In: Computational Linguistics and Intelligent Text Processing, Gelbukh, A. (Editor), 2014, pp. 98–114.
- [68] Moraes, S.; Manssour, I.; Silveira, M. "7x1pt: um corpus extraído do twitter para análise de sentimentos em língua portuguesa". In: Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, 2015, pp. 21–25.
- [69] Nass, C.; Brave, S. "Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship". The MIT Press, 2005.
- [70] Navaro, R. "Unconscious bias". Capturado em: <https://diversity.ucsf.edu/resources/unconscious-bias>.
- [71] Nery, C. "Rendimento impacta acesso da população a bens tecnológicos e internet". Capturado em: <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/27522-rendimento-impacta-meio-de-acesso-da-populacao-a-bens-tecnologicos-e-internet>, Outubro 2020.
- [72] Ornella, A. "It's all about sex the peculiar case of technology and gender". In: Riconoscersi. Corpo e gender tra individuale e sociale, 2013, pp. 183–213.
- [73] Pang, B.; Lee, L.; Vaithyanathan, S. "Thumbs up? sentiment classification using machine learning techniques". In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), 2002, pp. 79–86.
- [74] Pelle, R. "Identificação de comentários ofensivos da web". Capturado em: <https://www.lume.ufrgs.br/handle/10183/193539>, Junho 2021.
- [75] Penny, L. "CYBERSEXISM: sex, gender and power on the internet". Bloomsbury Publishing, 2013, 45p.
- [76] Perez, C. "Invisible Women: Exposing Data Bias in a World Designed for Men". Chatto Windus, 2015, 432p.
- [77] Piryani, R.; Madhavi, D.; Singh, V. "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015", *Information Processing Management*, vol. 53–1, 2017, pp. 122 – 150.
- [78] Powers, D. "Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation", *Mach. Learn. Technol.*, vol. 2, 01 2008.
- [79] Provost, F.; Fawcett, T. "Data science and its relationship to big data and data-driven decision making", *Big Data*, vol. 1–1, 2013, pp. 51–59.

- [80] Real, A. "Neologismos em redes sociais um estudo sobre as comunicações digitais na educação". Restinga Sêca, Rio Grande do Sul, Brasil: Universidade Federal de Santa Maria, 2017, pp. 22.
- [81] Relations, T. I. "Q2 2020 letter to shareholders". Capturado em: https://s22.q4cdn.com/826641620/files/doc_financials/2020/q2/Q2-2020-Shareholder-Letter.pdf, Setembro 2020.
- [82] Ricci, R. "Análise de sentimentos no Twitter sobre a Reforma da Previdência no ano de 2019". Uberlândia, Minas Gerais, Brasil: Universidade Federal de Uberlândia, 2020.
- [83] Sailunaz, K.; Dhaliwal, M.; Rokne, J.; Alhajj, R. "Emotion detection from text and speech: a survey", *Soc. Netw. Anal. Min.*, vol. 8–1, 2018, pp. 28:1–28:26.
- [84] Salton, G.; Buckley, C. "Term-weighting approaches in automatic text retrieval", *Information Processing Management*, vol. 24–5, 1988, pp. 513–523.
- [85] Sanches Duran, M.; Avanço, L.; Aluísio, S.; Pardo, T.; Volpe Nunes, M. d. G. "Some issues on the normalization of a corpus of products reviews in Portuguese". In: Proceedings of the 9th Web as Corpus Workshop (WaC-9), 2014, pp. 22–28.
- [86] Shalev-Shwartz, S.; Ben-David, S. "Understanding Machine Learning: From Theory to Algorithms". Cambridge University Press, 2014.
- [87] Shimabukuro, I. "Uso de assistentes virtuais no brasil cresce 47% durante a pandemia". Capturado em: <https://olhardigital.com.br/2020/10/14/noticias/uso-de-assistentes-virtuais-no-brasil-cresce-47-durante-a-pandemia/>, Maio 2021.
- [88] Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; Weber, I. "Analyzing the Targets of Hate in Online Social Media". Cornell University, 2016, 1603.07709.
- [89] Silva, N. "Análise de sentimentos em textos curtos provenientes de redes sociais". Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.
- [90] Soong, H.-C.; Jalil, N. B. A.; Kumar Ayyasamy, R.; Akbar, R. "The essential of sentiment analysis and opinion mining in social media : Introduction and survey of the recent approaches and techniques". In: 2019 IEEE 9th Symposium on Computer Applications Industrial Electronics (ISCAIE), 2019, pp. 272–277.
- [91] Speriosu, M.; Sudan, N.; Upadhyay, S.; Baldridge, J. "Twitter polarity classification with label propagation over lexical links and the follower graph". In: Proceedings of the First workshop on Unsupervised Learning in NLP, 2011, pp. 53–63.

- [92] Spertus, E. "Smokey: Automatic recognition of hostile messages". In: Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence, 1997, pp. 1058–1065.
- [93] Stanley, M. "The better a tech company's gender diversity, the greater its returns, on average, according to new research". Morgan Stanley, 2017.
- [94] Stone, P. J.; Dunphy, D. C.; Smith, M. S. "The general inquirer: A computer approach to content analysis." MIT press, 1966.
- [95] Strein, T. "Gírias e abreviações mais usadas do whatsapp". Capturado em: <https://www.dicionariopopular.com/girias-abreviacoessiglas-mais-usadas-do-whatsapp/>, Junho 2021.
- [96] Stromberg, J. "Why women like deep voices and men prefer high ones". Capturado em: <https://www.smithsonianmag.com/science-nature/why-women-like-deep-voices-and-men-prefer-high-ones-41492244/>, Junho 2021.
- [97] Tajfel, H. "Experiments in intergroup discrimination", *Scientific American*, vol. 223, 1970, pp. 96–103.
- [98] The United Nations Educational, S.; Organization, C. "Hey update my voice". Capturado em: <https://heyupdatemyvoice.org/pt/>, Junho 2021.
- [99] Theodoridis, S.; Koutroumbas, K. "Pattern Recognition and Neural Networks". Springer Berlin Heidelberg, 2001, pp. 169–195.
- [100] Toueg, G. "Marketing digital em 2021: 10 maiores tendências que realmente vale a pena apostar". Capturado em: <https://rockcontent.com/br/blog/tendencias-de-marketing-digital/>, Maio 2021.
- [101] Toueg, G. "Robôs precisaram aprender a responder ao assédio feito por homens". Capturado em: <https://www.uol.com.br/tilt/noticias/redacao/2021/04/08/bia-chatbot-do-bradesco-vai-responder-a-altura-quem-vier-com-assedio.html>, Maio 2021.
- [102] Turney, P. D. "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews". In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 417–424.
- [103] Wakabayashi, D. "Google cousin develops technology to flag toxic online comments". Capturado em: https://www.nytimes.com/2017/02/23/technology/google-jigsaw-monitor-toxic-online-comments.html?_r=0, Março 2021.

- [104] Waseem, Z.; Hovy, D. "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter". In: Proceedings of the NAACL Student Research Workshop, 2016, pp. 88–93.
- [105] Watanabe, H.; Bouazizi, M.; Ohtsuki, T. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection", *IEEE Access*, vol. 6, 2018, pp. 13825–13835.
- [106] West, M.; Kraut, R.; Chew, H. "I'd blush if i could: closing gender divides in digital skills through education". Capturado em: <https://unesdoc.unesco.org/ark:/48223/pf0000367416>, Setembro 2020.
- [107] Wettschereck, D.; Aha, D.; Mohri, T. "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms", *Artificial Intelligence Review*, vol. 11, 06 2000.
- [108] Wiebe, J. M.; Bruce, R. F.; O'Hara, T. P. "Development and use of a gold-standard data set for subjectivity classifications". In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 246–253.
- [109] Wulczyn, E.; Thain, N.; Dixon, L. "Ex machina: Personal attacks seen at scale". 1610.08914, 2017.
- [110] Xue, Z.; Yin, D.; Davison, B. D. "Normalizing microtext". In: Proceedings of the 5th AAAI Conference on Analyzing Microtext, 2011, pp. 74–79.