PUC Minas

pista: periódico interdisciplinar

[sociedade tecnologia ambiente]

RECONHECIMENTO AUTOMÁTICO DE NARRATIVAS VIOLENTAS CONTRA A MULHER UTILIZANDO CLASSIFICADORES NEURAIS

AUTOMATIC RECOGNITION OF VIOLENT NARRATIVES AGAINST WOMEN USING NEURAL CLASSIFERS

Submetido em: 03/082021
Aprovado em: 06/09/2021

Lucas Amancio Mantini[1]
Yanna Paula de Araújo da Silva[2]
Magali Rezende Gouvêa Meireles[3]

Este trabalho foi vencedor do II Prêmio de Trabalhos Científicos de Graduação da Pontifícia Universidade Católica de Minas Gerais (PUC Minas), unidade São Gabriel, na categoria Tecnologia.

RESUMO

Identificar traços de violência contra as mulheres de forma automática em um ambiente de rede social é de extrema importância, tendo em vista o espaço ocupado por elas nas relações humanas modernas. Este artigo tem como objetivo implementar um classificador neural capaz de identificar narrativas violentas em *tweets*. A criação do classificador abrange as etapas de criação de um dicionário de palavras relativas ao tema; a criação de uma base de dados, por meio da coleta de postagens do *Twitter*; a utilização de técnicas de pré-processamento de textos e processamento de linguagem natural e o treinamento de uma Rede Neural Artificial. Os resultados do processo de classificação serão apresentados baseando-se nas métricas de acurácia, precisão, *F1 score* e revocação. Espera-se que o classificador possa ser utilizado para combater a prática de violência contra mulheres de forma eficiente e rápida, além de possibilitar que as redes sociais desenvolvam métodos automáticos para bloquear e/ou banir postagens dessa natureza.

Palavras-chave: narrativas de violência contra a mulher; processamento de linguagem natural; redes neurais artificiais; *Twitter*.

[1] Graduado em Sistemas de Informação pela Pontifícia Universidade Católica de Minas Gerais. Contato: lucas.mantini@sga.pucminas.br
[2] Graduada em Sistemas de Informação pela Pontifícia Universidade Católica de Minas Gerais. Contato: yanna.silva@sga.pucminas.br
[3] Doutora em Ciência da Informação pela Universidade Federal de Minas Gerais. Professora do curso de Sistemas de Informação da PUC Minas. Contato: magali@pucminas.br

ABSTRACT

Automatically identifying traces of violence against women in social media is of utmost importance, given their status in modern human relationships. This article aims to implement a neural classifier capable of identifying abusive speeches in tweets. The creation of a classifier covers the steps of creating a dictionary of words related to the theme; the creation of a database through the collection of Twitter posts; the use of text pre-processing techniques and natural language processing and the training of an Artificial Neural Network. The results of the classification process will be presented based on the metrics accuracy, precision, F1 score and recall. It is expected that the classifier will be able to be used in order to combat the practice of violence against women in a efficient and fast way, in addition to allowing social networks to develop automatic methods to block and/or ban posts of this nature.

Keywords: *narratives of violence against women; natural language processing; artificial neural Networks; Twitter.*

1 INTRODUÇÃO

A sociedade atual está relacionada a um contexto tecnológico que lhe permite uma grande variedade de facilidades e comodidades. O advento, a popularização e a constante melhoria da rede mundial de computadores permitiram o surgimento de novas formas de comunicação representadas, fortemente, pelas redes sociais. A favor da tecnologia, elas demonstram a sua utilização por meio da interação e propagação de informação. Porém, podem servir tanto a bons quanto a maus propósitos.

A impressão de anonimato, mesmo que falso, dá espaço para que diversos eventos de cunho negativo ocorram. Episódios de violência, discriminação, preconceito, danos morais, exposição indevida, dentre outros, ocorrem diariamente e é de extrema importância a sua identificação para a tomada de providências junto às leis e aos órgãos responsáveis.

Em um levantamento realizado pela Folha de S. Paulo (2020), com dados de 2019, revelou-se um aumento de 7,2% de casos de feminicídio no Brasil com crescimento expressivo em vários estados. Quando associada aos crimes virtuais, a violência contra a mulher ganha novas formas. Em ambientes virtuais, mesmo que transmitida por meio de texto, insultos, humilhações, ameaças e ofensa sexual podem ser disseminados. Segundo o Globo (2019), a prática via Internet desse tipo de violência aumentou de 1,2% em 2017 para 8,2% em 2018, de acordo com 1092 brasileiras entrevistadas, revelando uma ampliação do uso dessas ferramentas.

Novas definições para práticas abusivas como o termo "estupro virtual" surgem e casos são registrados, alertando sobre o abuso psicológico e suas consequências que transpõem as barreiras virtuais, se assemelhando às de um estupro físico (UOL, 2020). Considerando-se essas

taxas e fatos, faz-se necessária uma intervenção a fim de criar e reforçar ações para a identificação desses atos.

Este trabalho possui como objetivo geral implementar um classificador, utilizando Redes Neurais Artificiais, capaz de identificar traços de narrativas consideradas violentas direcionadas às mulheres. Será criado um dicionário de palavras relacionadas ao tema central, que servirá de referência para a coleta e a identificação dos textos contendo traços violentos nas redes sociais. A base de dados será construída a partir da coleta dos textos. E passará por etapas de pré-processamento aliadas ao processamento de linguagem natural, utilizadas como recursos para o refinamento dos dados. E, por fim, o processo de classificação será por meio de uma Rede Neural Artificial.

Na próxima seção, o referencial teórico é apresentado, contendo os conceitos de redes sociais, narrativas de violência, pré-processamento de textos, Processamento de Linguagem Natural, Aprendizado de Máquina e Redes Neurais Artificiais. A Seção 3 refere-se aos trabalhos relacionados que dão suporte ao desenvolvimento e entendimento das etapas que compõem os processos de construção do classificador. A Seção 4 apresenta a metodologia aplicada à coleta e à classificação dos textos.

2 REFERENCIAL TEÓRICO

Esta seção apresenta os principais conceitos relacionados a este trabalho. Na subseção 2.1 é discorrido sobre as redes sociais. A violência contra a mulher é retratada pela subseção 2.2. A subseção 2.3 aborda o pré-processamento de textos. O processamento de linguagem natural é apresentado na subseção 2.4. A subseção 2.5 trata do aprendizado de máquina, suas tarefas e métodos de aprendizagem. Por fim, a subseção 2.6 aborda o conceito de Redes Neurais Artificiais.

2.1 Redes Sociais

Com o advento da Internet, novas formas de comunicação surgiram abrangendo um grande número de pessoas ao redor do mundo. Toda essa interação juntamente com a ligação em rede se consolidou na formação das redes sociais.

Uma rede social é um conjunto de atores e suas conexões que possuem como elementos a interação, relação e laços sociais (RECUERO, 2011). Para Silva (2018), nesse novo cenário, "[..] a comunicação é estabelecida por meio dos discursos, os quais transmitem significados e

ações, que, na Internet, são propagados seja em forma textual ou por meio de áudios, imagens, vídeos, gifs e memes". Existe uma grande variedade de redes sociais para diversas finalidades, uma delas é o *Twitter*.

Surgido em 2006 e ainda em atividade no Brasil, o *Twitter* se estabeleceu por meio da ideia de possibilitar aos seus usuários o compartilhamento do que estão fazendo, comentando ou discutindo por meio de textos curtos restringidos a um determinado número de caracteres no formato de microblog (COUTINHO, 2019). Essa é uma rede que facilita e possibilita a extração de informações para a geração de conhecimento, um dos pontos fundamentais a serem explorados no presente trabalho.

2.2 Violência Contra a Mulher

Restringir, reprimir e ofender, física e moralmente, uma pessoa ou um grupo delas pode ser compreendido como violência. Esses atos estão presentes dentro da sociedade sendo presenciados e noticiados frequentemente impedindo "[...] a outra pessoa de manifestar seu desejo e sua vontade, sob pena de viver gravemente ameaçada, até mesmo ser espancada, lesionada ou morta"(TELES; MELO, 2017).

Dentre os diferentes grupos atingidos por manifestações violentas estão as mulheres. Teles e Melo (2017), ainda, afirmam que a expressão "violência contra a mulher", trazida à tona pelo movimento feminista dos anos 1970, é a grande referência quando tratada a violência de gênero que é a representação das diferenças socioculturais existentes entre os sexos masculino e feminino.

O combate a esse tipo de violência é de extrema importância e, no Brasil, a Lei nº 11.340/2006 (BRASIL, 2006), também conhecida como Lei Maria da Penha, encabeça essas ações. Visando a coibição da violência doméstica e familiar, essa lei cria mecanismos para punir e erradicar a violência contra as mulheres e eliminar a discriminação que elas sofrem.

Como formas de violência, ainda segundo a Lei Maria da Penha, artigo 7º, destacam-se:

- Violência física que ofende a integridade ou saúde corporal;
- Violência psicológica, podendo ser qualquer conduta que cause dano emocional e diminuição da autoestima, também contemplando a degradação ou controle de ações, comportamentos, crenças e decisões mediante qualquer tipo de ameaça, humilhação, manipulação, entre outros;
- Violência sexual, abrangendo a presença ou a participação em relação sexual não desejada sendo, para tanto, a vítima intimidada, ameaçada, coagida ou

levada por meio do uso de força;

- Violência patrimonial, que aborda a conduta de retenção, subtração, destruição parcial ou total de qualquer bem como instrumentos de trabalho, documentos pessoais, valores ou recursos econômicos destinados a satisfação de necessidades da vítima;
- Violência moral, cuja conduta se configura como calúnia, difamação ou injúria.

O *Twitter* já considera, em suas regras e políticas, a violência sexual tratando do comportamento abusivo que se dá quando há "tentativas de assediar, intimidar ou silenciar a voz de outra pessoa"(TWITTER, 2020) por meio de perseguição envolvendo ameaças violentas, xingamentos e apelidos, a redução de uma pessoa a uma condição degradante e a incitação ao medo vinculada por um perfil e discurso expressado.

Ainda segundo as políticas do Twitter, textos (*tweets*) com teor abusivo são aqueles que envolvem:

- Expressão do desejo de que uma pessoa sofra algum tipo de lesão;
- Discussão de cunho sexual sobre o corpo de alguém;
- Solicitação de atos sexuais;
- Uso de insultos agressivos com caráter de intimidação e assédio;
- Incentivo ou incitação ao assédio.

2.3 Pré-processamento de textos

Uma forma comum de tramitação de dados são os textos. Eles contêm informações valiosas, porém de difícil obtenção em sua forma bruta. O pré-processamento correto de textos é um fator fundamental para que técnicas de análise e aprendizagem possam ser aplicadas com sucesso na extração de informações úteis.

Documentos textuais podem ser classificados tanto na categoria não estruturada, quanto na categoria semi-estruturada de dados (MELO, 2018). Segundo Vijayarani et al. (2015), dados não estruturados geralmente não são armazenados em bases de dados do tipo linha-coluna e podem ser considerados opostos aos dados estruturados.

Vijayarani et al.(2015) discutem passos chave do pré-processamento de textos, sendo eles a extração, a remoção de *stop words*, o *stemming* e a aplicação de algoritmos que implementam métodos TF/IDF. Ainda segundo os autores, a técnica de extração consiste em

tokenizar todo o conteúdo do texto em palavras separadas; a técnica de remoção de *stop words* é utilizada para remover palavras que sejam de menor importância para a análise, reduzindo o volume de dados. *Stemming* é a técnica que busca a relação entre as palavras e sua origem, assim eliminando sufixos, segundo Ramasubramanian e Ramya (2013). Por fim, os métodos baseados em TF/IDF extraem somente os termos mais relevantes do texto, eliminando assim os termos mais comuns e menos relevantes, segundo Moon e Raju (2013).

2.4 Processamento de Linguagem Natural

A forma de comunicação dos seres humanos, fala e escrita, apesar de simples para os praticantes, é de difícil compreensão do ponto de vista computacional. Uma linguagem humana possui nuances que, em sua maioria, a máquina não compreende. O Processamento de Linguagem Natural (PLN) surgiu com o intuito de propor soluções para problemas desse tipo.

Segundo Barbosa (2018), O PLN consiste na junção da Computação, Matemática, Linguística, Psicologia e Ciência da Informação para estudar a representação e interpretação de informações faladas e/ou escritas por seres humanos.

Para Chowdhury (2003), PLN é um ramo de pesquisa que busca compreender como os computadores podem ser utilizados para entender e manipular a linguagem natural de falas e textos para gerar algo útil. Ele ainda aborda que algumas áreas de estudo que utilizam PLN incluem a inteligência artificial, processamento de textos de linguagem natural, interfaces para usuários, reconhecimento de discursos, dentre outras.

2.5 Aprendizado de Máquina

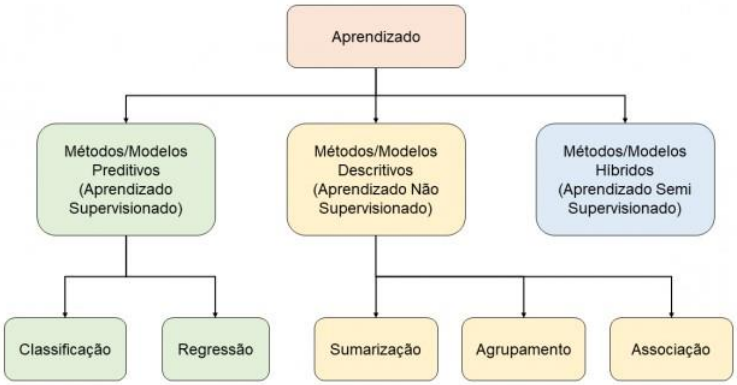
(GAMA et al., 2011) Um ramo da área computacional que vem ganhando destaque e força nos últimos anos é a Inteligência Artificial (IA). São inúmeras as possibilidades e aplicações que a IA pode proporcionar em termos de inovação e tecnologia para o ser humano. Uma ramificação muito promissora da IA é o Aprendizado de Máquina (AM). De acordo com Gama et al. (2011), a máquina é capaz de aprender por meio de experiências passadas a ela, empregando a indução, um princípio de inferência. Desse modo, os algoritmos de AM podem induzir hipóteses ou funções para a resolução de problemas que utilizam dados, representando instâncias do real problema inicialmente proposto. Ainda conforme os autores, o AM, em geral,

não utiliza somente conceitos computacionais, mas de diversas áreas de conhecimento diferentes, como a Estatística e a Probabilidade, Teoria da Informação, Neurociência, entre diversas outras.

2.5.1 Tarefas de Aprendizado

As tarefas delegadas ao AM podem ser classificadas em dois grupos: descritivas ou preditivas. As tarefas do tipo descritivas não utilizam de um atributo da saída, sendo assim visam explorar e/ou descrever um conjunto de dados qualquer, já as do tipo preditivas utilizam atributos de entrada e saída, pois a meta deste tipo é encontrar um modelo (função ou hipótese) que possa ser utilizado para prever um valor ou rótulo com base nos dados utilizados durante a fase de treinamento (GAMA et al., 2011). Na Figura 1 é possível observar a hierarquia de aprendizado seguida pelos algoritmos de Aprendizado de Máquina.

Figura 1 – Hierarquia de aprendizado dos algoritmos de AM.



Fonte: Embarcados (2019).

2.5.2 Métodos de Aprendizagem

Os métodos utilizados para realizar o aprendizado em máquinas são diversificados e possuem características específicas. Segundo Corcovia (2019), o aprendizado supervisionado possui uma classe para cada amostra utilizada na etapa de treinamento. Deste modo, a máquina possui acesso aos dados e a qual classe os dados trabalhados pertencem. Assim, é possível medir os níveis de acerto e erro nas classificações. Os modelos preditivos utilizam algoritmos que aplicam técnicas de aprendizado supervisionado.

Por sua vez, no aprendizado não-supervisionado a classe de cada amostra utilizada no treinamento não é conhecida, assim como o número total de classes pode ser desconhecido em

um primeiro momento (CORCOVIA; ALVES, 2019). Os modelos descritivos usam algoritmos que põem em prática técnicas de aprendizado não-supervisionado.

O terceiro tipo de aprendizado utilizado pelo AM é denominado Aprendizado por Reforço (AR). O AR utiliza uma política em que um estado é mapeado, dentro de um mapa de estados, e uma possível ação para este estado (MENDONÇA; SOUZA, 2018). Esta técnica é utilizada pelo mesmo para verificar o quão bom ou útil é a combinação estado-ação mapeada por meio de um esquema de recompensas. Desse modo, o desafio de AR é encontrar pares de estado-ação que maximizem a recompensa que a máquina irá obter. A maximização da recompensa pode ser feita explorando novos pares de estado-ação ou revisitando aqueles já descobertos.

Os métodos de aprendizagem podem ser utilizados para a resolução de diversos tipos de problemas. Escolher o método mais adequado para o problema proposto é importante para obter sucesso em sua resolução. Graficamente, as aplicações dos métodos de AM são mostradas pelo diagrama presente na Figura 2. É válido ressaltar que as aplicações dos métodos não são rígidas, como sugerido. Elas podem ser utilizadas em conjunto para um mesmo método.

Figura 2 – Aplicações dos Métodos de Aprendizado de Máquina.

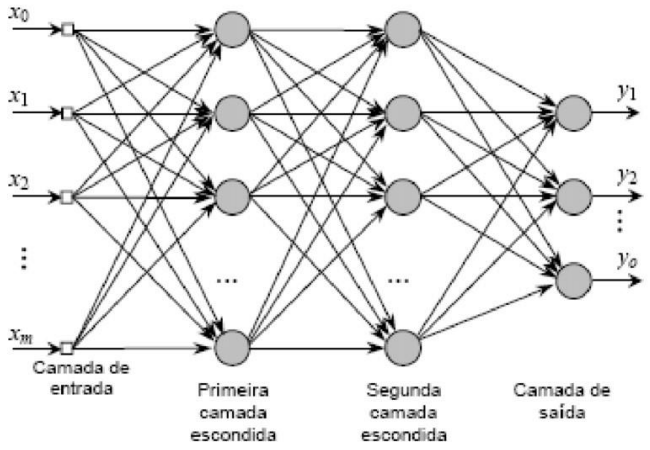
Fonte: Adaptado de Cognub (2020).

2.6 Redes Neurais Artificiais

As Redes Neurais Artificiais são um dos tipos de algoritmos que integram o AM. Haykin (2011) as define como processadores paralela e maciçamente distribuídos com propensão natural para armazenar conhecimento experimental, tornando-o disponível para uso posterior,

além de constituídas de pequenas unidades de processamento de informação denominadas neurônios. Os neurônios são baseados nos presentes no sistema nervoso central humano e são interligados entre si, formando assim uma rede. A obtenção de conhecimento em uma RNA se dá pelo reajuste dos pesos sinápticos, valores presentes nas ligações entre os neurônios. E elas podem ser constituídas por uma única camada de neurônios ou por diversas, conforme ilustra a Figura 3, tornando-se assim uma *Multilayer Perceptron* (MLP).

Figura 3 – Rede Neural Multicamada (MLP).



Fonte: Oliveira et al. (2010).

3 TRABALHOS RELACIONADOS

Com o objetivo de entender e explorar as etapas que compõem a mineração de textos, e o uso de Redes Neurais Artificiais, foram elencados e considerados alguns trabalhos. O primeiro trabalho é de Teixeira (2019), que aborda a análise de sentimento de usuários do *Twitter* relacionada a situação política brasileira. No segundo, Sanga (2017) analisa a mineração de textos para o tratamento automático em sistemas de atendimento ao usuário. No terceiro trabalho selecionado, Santos (2017) faz um estudo de caso sobre a análise de opiniões em redes sociais, com foco no *Twitter*. Por fim, Jesus (2018) evidencia técnicas de identificação e classificação da misoginia em redes sociais. Todas essas perspectivas facilitam o entendimento das diversas abordagens que permeiam a área em enfoque, principalmente, por sua aplicação relacionada ao *Twitter*, além de contribuir para a implementação deste estudo.

Teixeira (2019) em seu trabalho, estabelece como objetivo a pesquisa e a aplicação da análise de sentimentos juntamente com a mineração de dados. Por meio de textos obtidos no *Twitter* (*tweets*), com o uso de parâmetros envolvendo palavras-chave de cunho político e a mineração de textos, foram gerados gráficos por meio de nuvens de palavras para destacar e interpretar a frequência das mesmas, e, assim, obter resultados satisfatórios que indicaram o

sentimento positivo, predominante entre a maioria das pessoas, quanto situação política do Brasil. É relevante destacar, também, que aspectos isolados puderam ser observados indicando sentimento negativo quanto ao corte da verba na educação e positivo quanto à reforma da previdência na época.

Sanga (2017) propôs, em sua tese, o estudo da mineração de textos no contexto de empresas de telecomunicações que utilizam sistemas de atendimento ao cliente. Ela utilizou de duas bases distintas, a primeira contendo somente dados do ambiente de CRM de uma empresa. A segunda pode ser considerada uma extensão da primeira, pois foram adicionados dados, resultado da mineração de textos, de uma base de reclamações da ANATEL. Feito esse processo de formação do conjunto de dados para os experimentos, a autora utilizou os algoritmos de classificação: árvores de decisão, classificadores *Naive Bayes*, *K-Nearest Neighbors* (K-NN), *support vector machines* (SVM) e Redes Neurais Artificiais. Os resultados obtidos mostram que a mineração de textos interfere positivamente em algoritmos de classificação, melhorando consideravelmente seu desempenho. E que os algoritmos de árvore de decisão e SVM apresentaram o melhor desempenho após a conclusão dos experimentos.

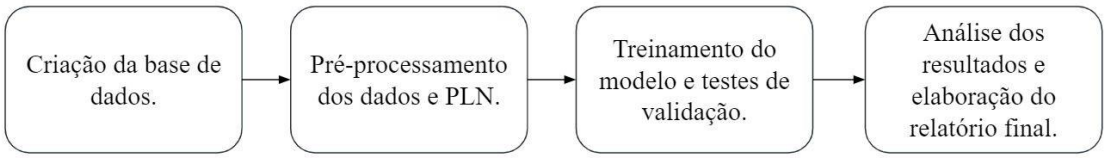
Em seu trabalho, Santos (2017) visa realizar uma análise de opiniões em redes sociais sobre o momento político vivenciado pelo Brasil no ano de 2017. Para a montagem da base de dados, ela utilizou uma interface de programação de aplicações (API) do próprio *Twitter*. Essa API obtém as postagens em tempo real via *streaming*. Feita a coleta e o tratamento dos dados obtidos, a autora utilizou os algoritmos classificadores *Naive Bayes*, K-NN e SVM, para classificar os tweets. Também foram utilizadas algumas técnicas de visualização de dados, como a nuvem de palavras. Os resultados obtidos ao final dos testes demonstram que todos os algoritmos tiveram resultados satisfatórios quanto à classificação dos *tweets*, porém o K-NN se destacou. Além disso, demonstram o sentimento, majoritariamente, negativo da população brasileira em relação ao momento político estudado.

Finalmente, Jesus (2018) também aborda o *Twitter* como fonte de obtenção de comentários para o desenvolvimento de um algoritmo capaz de detectar e identificar o comportamento misógino nessas publicações. Realizando um comparativo entre mapas auto organizáveis (SOM), classificação (MLP) e clusterização (*K-Means*), o autor busca a melhor alternativa considerando a precisão, sensibilidade e *F1-Score* ao aplicá-los a análise das categorias de misoginia geradas por meio da nuvem de palavras. Os resultados encontrados indicaram que a MLP obteve os melhores desempenhos se mostrando apenas mais difícil de ser implementada. O *K-Means* e a rede SOM mostraram desempenhos similares e insatisfatórios quanto ao objetivo desejado.

4 METODOLOGIA

A metodologia deste trabalho é estruturada em 4 etapas: a criação da base de dados, o pré-processamento da base de dados, o treinamento do classificador e a execução dos testes de validação do modelo obtido. A Figura 4 representa as quatro etapas descritas.

Figura 4 – Fluxograma geral da metodologia.



Fonte: Elaborado pelos autores.

4.1 Criação da base de dados

O ponto chave desta etapa e de todo o presente trabalho é a criação de um dicionário contendo as palavras mais pertinentes ao tema abordado. Seu principal objetivo é auxiliar e guiar a coleta de dados no *Twitter*. Para tanto, fizemos uma busca de palavras-chave em reportagens que retratem as histórias de violência contra mulheres. Desse modo, espera-se garantir que as palavras utilizadas sejam as mais coerentes possíveis com a realidade. É de suma importância que esse dicionário seja elaborado antes da captação dos dados, pois ele serve como parâmetro alimentador da busca.

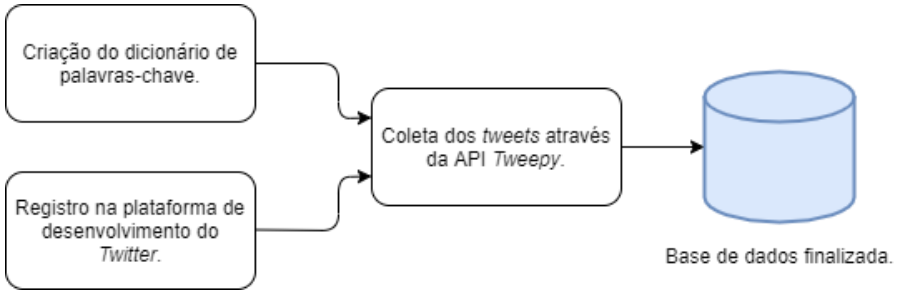
Após a sua conclusão, a coleta dos *tweets* é feita. Para tal, é utilizada uma API, chamada *Tweepy*, construída na linguagem *Python*, capaz de se comunicar diretamente com o *Twitter*.

Para que seja estabelecida a comunicação com a rede social, um aplicativo deve ser registrado em sua plataforma de desenvolvimento. Esse procedimento se faz necessário, pois assim informações necessárias para a autenticação da API são fornecidas.

O funcionamento da *Tweepy* é baseado na transmissão de dados via streaming. Dessa maneira a API fica "ouvindo" e quando um *tweet* contendo as palavras-chave buscadas é postado, ela é capaz de interagir com o *tweet*. Pode-se obter, assim, o nome do usuário, autor da postagem, o conteúdo do texto e, ainda, interagir com a postagem favoritando-a.

A Figura 5 ilustra, utilizando um fluxograma, como é feita a coleta dos *tweets* que compõem a base de dados.

Figura 5 – Fluxograma da coleta dos *tweets*.



Fonte: Elaborado pelos autores.

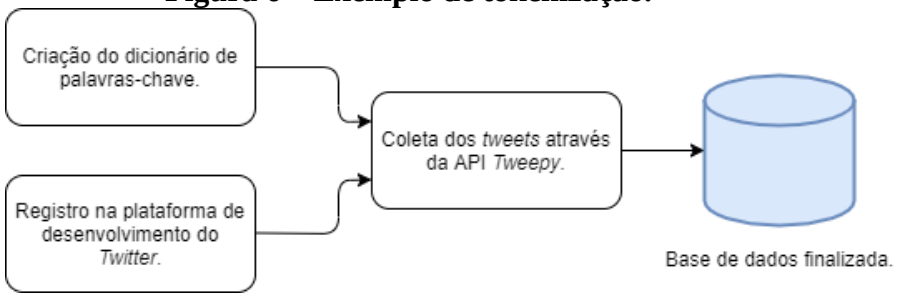
Ao final desta etapa, a base de dados é formada por textos que tratam de assuntos pertinentes ao tema, contendo as palavras-chaves buscadas, e os não pertinentes, que tratam de assuntos diversos.

4.2 Pré-processamento dos dados e processamento de linguagem natural

Após formada a base, os textos estão em seu estado puro. Portanto, é necessário que eles sejam pré-processados, porém não em sua totalidade, pois, uma porcentagem dela precisa ser preservada em seu estado original afim da realização dos testes de validação. Portanto, aproximadamente 80% dos dados são pré-processados, enquanto o restante é mantido sem alterações.

O primeiro passo desta etapa consiste na tokenização das sentenças. Segundo Palmer (2010), a tokenização ou segmentação de palavras consiste em identificar o início e o término de cada palavra de uma sentença, separando-a. A Figura 6 ilustra um exemplo de tokenização de uma sentença.

Figura 6 – Exemplo de tokenização.

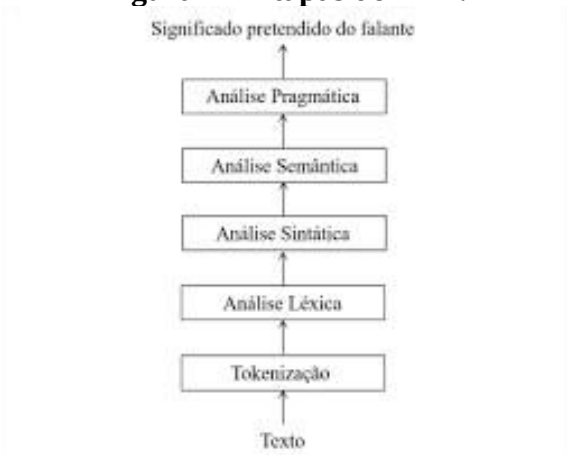


Fonte: Bóson Treinamentos em Tecnologia (2019).

A segunda parte dessa etapa é realizar o PLN. Esse processo é necessário pois os textos da base são escritos de forma muito próxima à forma falada da língua portuguesa. Bulegon et al. (2010) afirmam que o PLN possui quatro etapas em que aspectos morfológicos, sintáticos, semânticos e pragmáticos, nesta ordem, são analisados. Eles ainda definem que a análise

morfológica define artigos, substantivos, verbos e adjetivos, e os armazena em um dicionário. A análise sintática utiliza o dicionário construído para mostrar o relacionamento entre as palavras, assim como verificar outros aspectos da língua. Por sua vez, a análise semântica visa encontrar o sentido real de uma sentença ou palavra, baseada no encontro de termos ambíguos, afixos e sufixos. Por fim, a análise pragmática é responsável por unificar todo o mecanismo e mostrar visualmente o seu resultado. Unir a tokenização ao PLN resulta no processo de engenharia de software desenvolvido por Dale (2010), apresentado na Figura 7.

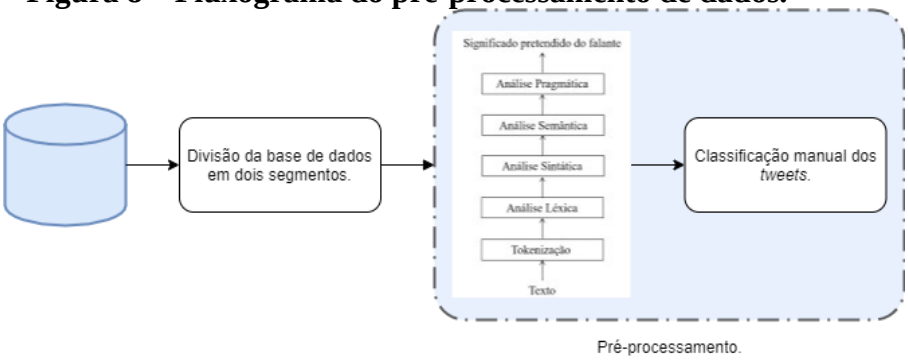
Figura 7 – Etapas do PLN.



Fonte: Dale (2010).

A terceira e última parte consiste em adicionar uma classe aos textos manualmente. Esse passo é importante pois os *tweets* precisam ser previamente classificados como "violentos" ou "não-violentos". Dessa maneira, os requisitos para treinar um modelo classificador serão cumpridos. O fluxograma mostrado na Figura 8 refere-se a esta etapa da metodologia.

Figura 8 – Fluxograma do pré-processamento de dados.



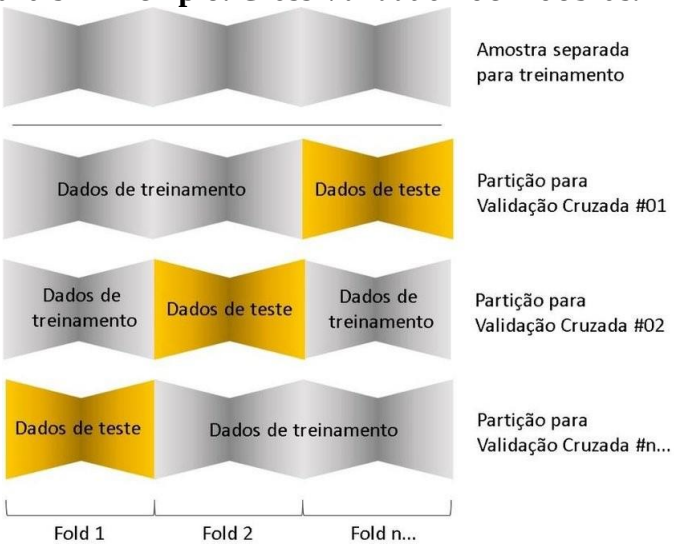
Fonte: Elaborado pelos autores.

4.3 Treinamento do classificador e a realização de testes de validação

A partir da conclusão do pré-processamento dos textos do segmento destinado ao treinamento do classificador, inicia-se esta etapa. O primeiro passo é a montagem da RNA. Neste trabalho, é utilizada uma rede do tipo *Multilayer Perceptron* juntamente com o algoritmo de treinamento *error backpropagation*. Então, os dados são fornecidos para que o treinamento efetivamente se inicie.

Para validar cada etapa do treinamento e buscar garantir que não haja nenhum tipo de vício, será utilizado o método de *cross validation* de n-dobras, exemplificado pela Figura 9. A avaliação do desempenho do modelo é feita utilizando as métricas de acurácia, precisão, revocação e *f1 score*. A matriz de confusão é utilizada a fim de prover suporte para a avaliação das métricas.

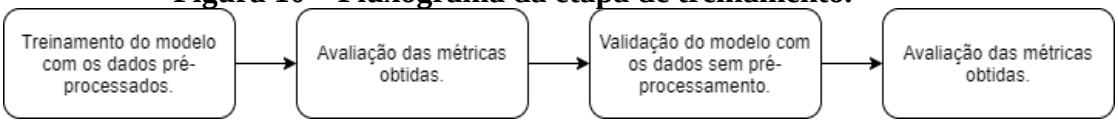
Figura 9 – Exemplo: *Cross Validation* de n-dobras.



Fonte: Versage (2015).

Concluído o treinamento, é utilizado o segmento da base de dados que não recebeu tratamento para a validação do modelo. O texto, em seu estado original, será fornecido para o modelo classificá-lo. Novas métricas são obtidas, baseadas nos números de novos dados classificados correta e incorretamente. Os resultados são analisados a fim de verificar o desempenho do modelo e, ao mesmo tempo, validá-lo. A presente etapa está representada no fluxograma da Figura 10.

Figura 10 – Fluxograma da etapa de treinamento.



Fonte: Elaborado pelos autores.

APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Os experimentos iniciais foram realizados visando obter um dicionário de palavras relacionadas ao tema central deste trabalho. Para obtê-lo, primeiro foi utilizada a técnica TF/IDF, aplicada a 15 reportagens com diferentes abordagens que englobam narrativas abusivas contra as mulheres no período de 2018 a 2020. O objetivo foi reduzir o contingente inicial de palavras relacionadas a serem analisadas. A escolha do algoritmo TF/IDF foi baseada na sua simplicidade e eficiência na ponderação de termos chave. Além disso, foi utilizado, como referência, o trabalho de Reginaldo et al. (2017), em que alguns algoritmos de extração de palavras-chave são comparados.

Para fins de otimização de resultados e execução do algoritmo, a técnica de remoção de sufixos, *stemming*, também foi aplicada. As reportagens pertencem a veículos de grande circulação midiática como o G1, da Rede Globo, o R7, da Rede Record, e afins. Com os termos ponderados, os resultados de todas as execuções foram unificados com o objetivo de alcançar aqueles comuns entre todos os textos.

Após o processo de unificação, apesar da utilização do TF/IDF e da contabilização dos termos comuns, a filtragem das palavras foi necessária para que apenas termos relevantes ao contexto fossem obtidos e aqueles que não se enquadrassem, desconsiderados. A Figura 11 apresenta a nuvem de palavras utilizada para melhor ilustrar o resultado alcançado.

Figura 11 – Nuvem de palavras obtida.

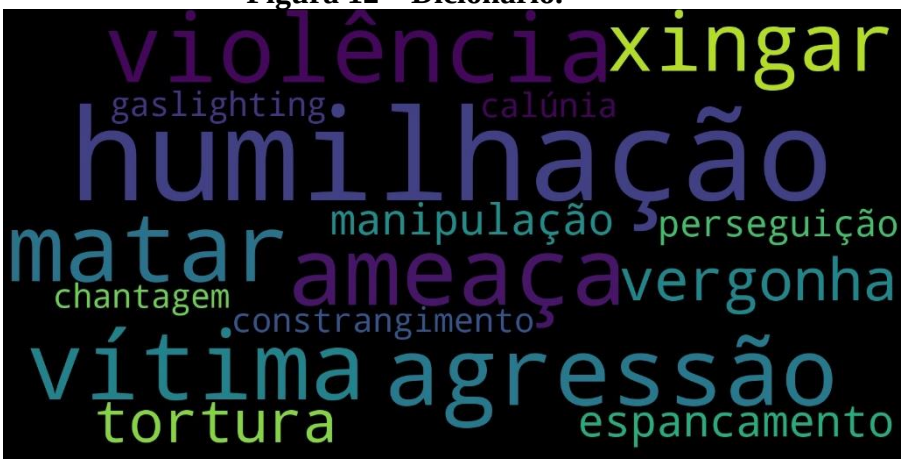
Fonte: Elaborado pelos autores.

Ao todo, foram ponderados 4.436 termos. Dentre estes, 2.789 eram formados por números e repetições e, portanto, foram removidos, totalizando 1.647 palavras. O critério de filtragem utilizado para a análise posterior se estabeleceu na seleção dos vocábulos presentes em 9 ou mais reportagens, acarretando na redução para 88 palavras. Dada a sua relevância, definida pelos autores de acordo com o tema, foram escolhidos os termos: ameaça, agressão,

violência, vítima, matar, xingar e vergonha.

Também foram acrescentados termos obtidos por meio do serviço 180 Play[4] que é uma plataforma de *streaming* que conta com cenas de violência contra a mulher em filmes, novelas e séries, além de textos explicativos, com o objetivo de conscientizar as pessoas sobre o assunto e, principalmente, explicitar a forma como os diferentes tipos de violência ocorrem. Foram considerados, então, os termos: tortura, espancamento, manipulação, perseguição, humilhação, constrangimento, *gaslighting*, chantagem e calúnia. A união dos dois conjuntos representa o dicionário obtido, conforme ilustrado pela Figura 12, composto, agora, por 16 palavras.

Figura 12 – Dicionário.



Fonte: Elaborado pelos autores.

Para a montagem da base de dados, foram utilizadas as palavras do dicionário como parâmetro de busca. Foram coletados cerca de 750 *tweets* para cada uma delas de tal modo que foram flexionadas nos tempos verbais particípio passado e infinitivo, assim como na sua forma substantiva. Por exemplo, para o termo "agressão", 250 *tweets* foram buscados por "agressão mulher", 250 por "mulher agredida" e 250 por "agredir mulher", totalizando 750. Algumas palavras são consideradas como casos especiais, pois as suas flexões e formas derivadas não são interessantes para as buscas. Para elas, não foram feitas alterações, sendo, portanto, mantidas em seu formato original. São elas: vítima, violência, vergonha e *gaslighting*, sendo identificados diretamente 750 *tweets*. Todas essas exceções podem ser vistas no Quadro 1 a seguir.

[4] Projeto de conscientização para mulheres vítimas de violência. Disponível em: https://www.180play.com.br/2/

Quadro 1 - Parâmetros de busca para termos não flexionados

Exceção	Parâmetros
Violência	Violência Mulher
Vítima	Mulher Vítima
Matar	Matar Mulher, Mulher Morta
Xingar	Xingar Mulher, Mulher Xingada
Vergonha	Vergonha Mulher, Mulher Envergonhada
gaslighting	*gaslighting*

Fonte: Elaborado pelos autores.

Os dados da base, inicialmente, foram manualmente classificados em "Sim", para dados que representam narrativas de violência, e "Não", para aqueles que não apresentam traços violentos. Após esse momento, foram removidos os símbolos visuais (popularmente conhecidos como "emoticons") que os usuários utilizam, assim como os sinais de pontuação. Além disso, as palavras abreviadas passaram por uma correção a fim de completá-las, também de forma manual. Então, as técnicas de *stemming* e tokenização são aplicadas. Todos esses procedimentos compõem o pré-processamento da base de dados, também sendo parte do processo de PLN. O Quadro 2 demonstra as diferenças entre uma mensagem antes e depois do pré-processamento.

Quadro 2 - Exemplo de pré-processamento

Tweet não processado	*Tweet* processado

Sim, mas isso é passado. Entendi que o problema não estava em mim, não tinha culpa por ter sido passada para trás, por ter sido humilhada. Entendi que o problema era com ele, que ele não tinha nenhum caráter. E hoje sou uma mulher segura, e não permito mais que façam isso comigo.	Sim pass entend problem mim culp ter sid pass trá ter sid humilh entend problem nenhum carát hoj mulh segur permit faç comig

Fonte: Elaborado pelos autores.

Com uma base parcialmente definida e fazendo uso do Weka[5], mais especificamente, suas funcionalidades voltadas para a classificação de instâncias e a RNA *MultilayerPerceptron*. Para o treinamento da RNA, foi utilizada a técnica de *cross-validation*, em 10 dobras, que é o padrão sugerido pela ferramenta. A taxa de aprendizagem foi estabelecida em 0.3. Foram utilizadas 500 épocas de treinamento, que são, também, padrões sugeridos pelo Weka.

A RNA continha 3 camadas, a primeira definida como de entrada, a segunda como oculta e a última de saída. A camada de entrada contém o total de neurônios igual ao número de instâncias da base de dados utilizada. A camada oculta tem o número de neurônios igual ao total de instâncias somadas ao total de classes e dividido por dois, padrão sugerido pelo Weka, totalizando cerca de 93 neurônios. É válido ressaltar que a ferramenta ajusta a quantidade de neurônios automaticamente, utilizando as quantidades de instâncias trabalhadas como base de cálculo. Por fim, a camada de saída tem somente 2 neurônios que representam as duas classes utilizadas na base. Dessa maneira, tornou-se possível o alcance dos resultados obtidos. A base utilizada possui 184 instâncias, sendo metade para cada classe ("Sim"e "Não"), e contém apenas dois atributos: o *tweet* e a classe.

Observando os valores retornados pela matriz de confusão, gerada ao final de cada teste, os parâmetros citados previamente, foram obtidos os resultados apresentados no Quadro 3. Os cálculos consideraram os valores de verdadeiro e falso positivo (46 e 42), além de verdadeiro e falso negativo (50 e 46), obtidos por meio da matriz.

Quadro 3 - Resultados Parciais

Métrica	Resultado
Acurácia	52,17%
Precisão	52,27%
Revocação	50%
F1 Score	51,11%

[5] *Waikato Environment for Knowledge Analysis* software de aprendizado de máquina de código aberto amplamente usado para aplicações de ensino, pesquisa e industriais, contendo várias ferramentas integradas. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/

Fonte: Elaborado pelos autores.

Pode-se notar, analisando a acurácia, que o valor em termos de instâncias classificadas corretamente foi de 52,17%. É válido ressaltar que os demais valores calculados indicam altas taxas de falsos positivos e falsos negativos.

A fim de melhorar os resultados iniciais, foram coletados novos *tweets*, alcançando uma base que contém 600 instâncias, sendo, 172 classificadas como "Sim" e 428 como "Não". Novos testes foram realizados, sem alterações dos parâmetros da RNA (com exceção daqueles ajustados automaticamente), mas os valores se mantiveram em torno de 50% de acurácia mesmo com esse aumento significativo do número de *tweets*.

Com as métricas em níveis ainda insatisfatórios, o próximo passo adotado foi o balanceamento da base de dados. O Weka possibilita a utilização de filtros correspondentes a técnicas de balanceamento, sendo uma delas o *undersampling*, que se caracteriza pela diminuição da classe majoritária, que no caso, é a das instâncias não violentas. Foram testados dois desses filtros de *undersampling*, o *Resample* e o *SpreadSubsample* cuja diferença se estabelece na produção da subamostra do conjunto de dados. O *Resample* pode fazer inserções de dados se necessário e o *SpreadSubsample* apenas equipara a quantidade de dados aleatoriamente. Com ambos testados, o filtro *Resample* foi o escolhido devido aos resultados alcançados, onde a base atingiu 300 instâncias, 50% para cada classe, tendo 232 mensagens classificadas corretamente e as métricas em torno de 70% conforme apresentado no Quadro 4.

Quadro 4 - Resultados Finais

Métrica	Resultado
Acurácia	77,3%
Precisão	77,4%
Revocação	77,3%
F1 Score	77,3%

Fonte: Elaborado pelos autores.

Apesar da redução no tamanho da base avaliada, percebe-se que houve uma melhora significativa na classificação realizada pela RNA. Houve um aumento dos valores de verdadeiro negativo e positivo e uma queda dos erros para 22,67%, correspondente a 68 instâncias. Devido à limitação de tempo para melhoria dos dados, ampliação da base e treinamento da rede, esse percentual é considerável e válido para um primeiro experimento.

CONSIDERAÇÕES FINAIS

As RNA são algoritmos cuja amplitude de aplicações pode trazer benefícios valiosos para a sociedade por meio da sua capacidade de aprendizado. Áreas como a da saúde vêm sendo impactadas pelo uso da inteligência artificial e do aprendizado de máquina e diferenças sutis na fala, por exemplo, vêm sendo estudadas com o objetivo de tornar possível a identificação de doenças como o Alzheimer, Parkinson, depressão e autismo, a partir de diagnósticos vocais. Ainda nesse contexto, esforços estão sendo realizados para que, futuramente, a Covid-19 possa ser detectada e a pessoa infectada tenha um diagnóstico preliminar por meio dos alto-falantes embutidos em *smartphones* e assistentes virtuais (Alexa, Siri, entre outros) de acordo com a maneira com que ela se expressa vocalmente (VIDALE, 2020). Estes experimentos confirmam que as RNA são poderosas ferramentas para a obtenção de padrões a partir de dados que lhes são apresentados. Ao longo do desenvolvimento deste trabalho, foi possível apresentar uma pequena parcela do quão útil uma RNA pode ser no auxílio de problemas corriqueiros.

A base de dados, utilizada neste trabalho, foi construída a partir de um dicionário, também, elaborado pelos autores. A etapa de elaboração da base permitiu um maior aprofundamento no tema central da pesquisa, o que proporcionou um diferente ponto de vista sobre as tratativas de casos de violência contra mulheres na sociedade brasileira. É válido ressaltar, também, a importância de haver palavras-chave coerentes com o tema, pois é por meio delas que a coleta dos dados se tornou viável.

Outro fator crítico identificado no desenvolvimento do trabalho foi o pré-processamento correto dos dados, que se mostrou fundamental para obter os resultados apresentados. É oportuno discutir a importância da etapa de pré-processamento, pois um tratamento refinado no vocabulário pode melhorar significativamente os resultados do processo de classificação por meio da redução do universo de termos, garantindo a presença dos mais relevantes e a exclusão daqueles que são irrelevantes (GUIMARÃES et al., 2019). Além disso, utilizar os parâmetros de ajustes da RNA que melhor modelam o problema associado ao contexto em questão é outro ponto que influencia o sucesso do projeto.

Os resultados obtidos com o experimento demonstram a relevância do estudo realizado. O ajuste da arquitetura da RNA é necessário para "lapidar" o modelo e garantir sua eficácia. A análise do impacto da remoção das *stop words* para os *tweets* presentes na base de dados não foi realizada.

Como sugestões de trabalhos futuros, primeiramente, é proposta a melhoria do dicionário, acrescentando mais palavras, para expandir o horizonte de buscas, e realizando uma colaboração com especialistas no assunto, se possível. Além disso, propõe-se a criação de um robô, dotado da rede, para analisar diretamente os *tweets* em tempo real e verificar se os padrões de acerto e erro se mantêm para textos não processados previamente. Outra proposta seria a ampliação da base experimental, agregando novas instâncias que poderão ser melhor trabalhadas e analisadas para a obtenção de informações relevantes, contribuindo, assim com a

detecção e com a denúncia de casos de violência contra a mulher.

Além desses aspectos, a utilização de outros algoritmos de classificação e de diferentes tipos de RNA podem ser sugeridos como meios de expandir o experimento proposto, podendo contribuir para a melhoria dos resultados encontrados.

REFERÊNCIAS

GUIMARÃES, Lucas Marques Sathler AND Meireles, Magali Rezende Gouvêa AND Almeida, Paulo Eduardo Maciel de.

BARBOSA, Júlio César. **Mineração de texto: uso de técnicas de processamento de linguagem natural para suporte à geração de projeções baseadas em opiniões do consumidor. 2018.** Tese (Doutorado) — Mestrado em Sistemas de Informação e Gestão do Conhecimento.

BÓSON TREINAMENTOS EM TECNOLOGIA. **Como tokenizar strings em Java com o método split.** 2019. Disponível em: <http://www.bosontreinamentos.com.br/java/ como-tokenizar-strings-em-java-com-o-metodo-split/>. Acesso em: 12 mai. 2020.

BRASIL. Lei n^{o} 11.340, de 07 de agosto de 2006. 2006. Cria mecanismos para coibir a violência doméstica e familiar contra a mulher, nos termos do § 8º do art. 226 da Constituição Federal, da Convenção sobre a Eliminação de Todas as Formas de Discriminação contra as Mulheres e da Convenção Interamericana para Prevenir, Punir e Erradicar a Violência contra a Mulher; dispõe sobre a criação dos Juizados de Violência Doméstica e Familiar contra a Mulher; altera o Código de Processo Penal, o Código Penal e a Lei de Execução Penal; e dá outras providências.

BULEGON, Hugo; MORO, Claudia Maria Cabral. Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. **Journal of Health Informatics**, v. 2, n. 2, 2010.

CHOWDHURY, Gobinda G. Natural language processing**. Annual review of information science and technology**, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.

COGNUB. **Cognitive Computing and Machine Learning**. 2020. Disponível em: <http: //www.cognub.com/index.php/cognitive-platform/>. Acesso em: 04 dez. 2020.

CORCOVIA, Lucas Oukus; ALVES, Renato Santos. Aprendizagem de máquina e mineração de dados. **Revista Interface Tecnológica**, v. 16, n. 1, p. 90–101, 2019.

COUTINHO, Vinícius Matheus de Medeiros Silva. Classificação de insultos em mensagens lgbtqfóbicas no twitter. **Trabalho de conclusão de curso apresentado ao curso de bacharelado em sistemas de informação.**, Universidade Federal da Paraíba, 2019.

DALE, Robert. Classical approaches to natural language processing. In: **Handbook of natural language processing, second edition.** [S.l.]: CRC Press, Taylor & Francis Group, 2010. p. 3–7.

EMBARCADOS. **Classificação Multirrótulo Hierárquica: Introdução. 2019.** Disponível em: <https://www.embarcados.com.br/classificacao-multirrotulo-hierarquica-intro/>. Acesso em: 11 abr. 2020.

FOLHA DE S. PAULO. **Feminicídio cresce no Brasil e explode em alguns estados.** 2020. Disponível em: <https://www1.folha.uol.com.br/cotidiano/2020/02/ feminicidio-cresce-no-brasil-e-explode-em-alguns-estados.shtml>. Acesso em: 12 mai. 2020.

GAMA, J. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina.** Rio de Janeiro: LTC, 2011. 396 p.

GUIMARÃES, Lucas Marques Sathler; MEIRELES, Magali Rezende Gouvêa; ALMEIDA, Paulo Eduardo Maciel de. Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação. **Perspectivas em Ciência da Informação**, scielo, v. 24, p. 169 – 190, 03 2019. ISSN 1413-9936. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362019000100169&nrm=iso>.

HAYKIN, Simon. **Redes neurais: princípios e prática.** [S.l.]: Bookman Editora, 2011.

INDURKHYA, Nitin; DAMERAU, Fred J; PALMER, David D. Text preprocessing. **Handbook of Natural Language Processing, Second Edition. Chapman et Hall/CRC**, p. 9, 2010.

JESUS, Fabrício Velôso de. Identificação e classificação automática de misoginia em redes sociais. 2018.

MELO, Fabrício Silva. Extração de relações a partir de dados não estruturados baseada em deep learning e supervisão distante. Pós-Graduação em Ciência da Computação, 2018.

MENDONÇA, Pedro Ginnari Barcelos Souza de; SOUZA, Vinícius Santos de. Uma implementação de um agente autônomo para o jogo bomberman com aprendizado por reforço. Niterói, 2018.

MOON, Ashish; RAJU, T. A survey on document clustering with similarity measures. **International Journal of Advanced Research in Computer Science and Software Engineering**, v. 3, n. 11, p. 599–601, 2013.

O GLOBO. **Perseguição. Pornografia de vingança. Ofensa sexual. A violência contra a mulher cresce nas redes.** 2019. Disponível em: <https://cutt.ly/Rg6f3tw>. Acesso em: 12 mai. 2020.

OLIVEIRA, Anderson Castro Soares de et al. Aplicação de redes neurais artificiais na previsão da produção de álcool. **Ciencia E Agrotecnologia - CIENC AGROTEC**, v. 34, 04 2010.

RAMASUBRAMANIAN, C; RAMYA, R. Effective pre-processing activities in text mining using improved porter's stemming algorithm. **International Journal of Advanced Research in Computer and Communication Engineering**, v. 2, n. 12, p. 4536–4538, 2013.

RECUERO, Raquel. **Redes sociais na internet**. 2. ed. Porto Alegre: Sulina, 2011. 206 p.

REGINALDO, Thiago et al. A comparison of algorithms for the extraction of keywords in a patent database. In: . [S.l.: s.n.], 2017.

SANGA, Dione Aparecido de Oliveira et al. **Mineração de textos para o tratamento automático em sistemas de atendimento ao usuário.** 2017. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná.

SANTOS, Tatiane Gomes dos. Análise de opiniões utilizando técnicas de mineração de dados em redes sociais. estudo de caso: Twitter. 2017.

SILVA, Micael Machado da. Análise da violência discursiva contra a mulher no twitter: estudo dos casos de anitta, dilma rousseff e ludmilla. 2018.

TEIXEIRA, Carolina Barcelos. Análise de sentimento dos usuários do twitter em relação à atual situação política do brasil. 2019.
TELES, Maria Amélia de Almeida; MELO, Mônica. **O que é a violência contra a mulher.** São Paulo: Brasiliense, 2017.

TWITTER. **Comportamento Abusivo.** 2020. Disponível em: <https://help.twitter.com/pt/rules-and-policies/abusive-behavior>. Acesso em: 26 abr. 2020.

UOL. **Abuso pela internet, estupro virtual entra na mira da polícia no Brasil.** 2020. Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2020/01/25/ abuso-pela-internet-estupro-virtual-entra-na-mira-da-policia-no-brasil.htm>. Acesso em: 12 mai. 2020.

VERSAGE, Rogerio. **Metamodelo para estimar a carga térmica de edificações condicionadas artificialmente.** 05 2015. Tese (Doutorado).

VIDALE, Giulia. A cura pela voz. **Veja**, Editora Abril, v. 2712, n. 46, p. 66 – 67, nov. 2020.