UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

GEOVANNY MAGALHÃES NOVAIS

Machine Learning e a detecção de violência contra a mulher em tweets

Aplicação de ontologia para avaliar a eliminação de ruído

Goiânia
2023

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [ ] Tese [ ] Outro*:__monografia_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

GEOVANNY MAGALHÃES NOVAIS

3. Título do trabalho

***Machine Learning* e a detecção de violência contra a mulher em *tweets* - Aplicação de ontologia para avaliar a eliminação de ruído**

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 24/08/2023, às 16:57, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Geovanny Magalhães Novais**, **Discente**, em 28/08/2023, às 15:22, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3976658** e o código CRC **8BDEE4FE**.

GEOVANNY MAGALHÃES NOVAIS

Machine Learning e a detecção de violência contra a mulher em tweets

Aplicação de ontologia para avaliar a eliminação de ruído

Trabalho de Conclusão apresentado à Coordenação do Curso de Computação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Bacharel em Computação.

Área de concentração: Otimização.

Orientadora: Profa. Deborah Silva Alves Fernandes

Goiânia
2023

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Novais, Geovanny Magalhães
Machine Learning e a detecção de violência contra a mulher em tweets [manuscrito] : Aplicação de ontologia para avaliar a eliminação de ruído / Geovanny Magalhães Novais. - 2023.
LXII, 62 f.: il.

Orientador: Profa. Deborah Silva Alves Fernandes.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Informática (INF), Ciência da Computação, Goiânia, 2023.
Bibliografia.
Inclui tabelas, lista de figuras, lista de tabelas.

1. Ontologia. 2. Machine Learning. 3. Processamento de Linguagem Natural. 4. Violência contra a mulher. 5. Twitter. I. Fernandes, Deborah Silva Alves, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Ao(s) **vinte e quatro** dia(s) do mês de **agosto** do ano de **2023** iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado "**Machine Learning e a detecção de violência contra a mulher em tweets - Aplicação de ontologia para avaliar a eliminação de ruído**", de autoria de GEOVANNY MAGALHÃES NOVAIS, do curso de **Ciência da Computação**, do(a) **Instituto de Informática** da UFG. Os trabalhos foram instalados pelo(a) **Profa. Dra. Deborah Silva Alves Fernandes** com a participação dos demais membros da Banca Examinadora: **Profa. Dra. Luciana de Oliveira Berretta (INF/UFG).** Após a apresentação, a banca examinadora realizou a arguição do(a) estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de **8,5** , tendo sido o TCC considerado **aprovado**.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 24/08/2023, às 16:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luciana De Oliveira Berretta**, **Professora do Magistério Superior**, em 24/08/2023, às 17:29, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3976659** e o código CRC **58C3B5C8**.

Resumo

Novais, Geovanny Magalhães. ***Machine Learning* e a detecção de violência contra a mulher em tweets**. Goiânia, 2023. 62p. Relatório de Graduação. Instituto de Informática, Universidade Federal de Goiás.

A monografia que se segue irá produzir uma ontologia com objetivo de eliminar ruído em textos colhidos na rede social Twitter. Ao realizar o filtro por meio da ontologia, se avaliará o impacto produzido perante a base de dados, quanto à detecção automatizada de violência contra a mulher. Os termos que a ontologia descreve estão no idioma Português Brasileiro e a validação de sua efetividade será realizada através de técnicas de aprendizado de máquina. Conforme a decisão, se optou pela aplicação de quatro técnicas, o *Logistic Regression*, *Multinomial Naïve Bayes*, *Random Forest*, *Support Vector Machine*, todas utilizadas por meio de Aprendizado Supervisionado, onde seus dados de treino e teste foram obtidos por catalogação manual.

Palavras–chave

Ontologia, *Machine Learning*, Processamento de Linguagem Natural, Violência contra a mulher, Twitter

Geovanny Magalhães Novais

Graduou–se em Ciências da Comuptação na UFG - Universidade Federal de Goiás. Durante sua graduação, foi um hábil aluno, membro de múltiplas eletivas. Atualmente desenvolve soluções para área de farmácia.

Dedico esse a minha prima Andressa e a minha avó Laura, mulheres que merecem todo o amor do mundo.

Agradecimentos

Agradeço aos meus pais por me permitirem investir tempo em evoluir como pessoa e pesquisador, aos meus amigos, em especial a Isabela e Victor pessoas tão especiais e essenciais em minha vida, que apoiaram em todo o tempo de pesquisa. Agradeço ao colega de pesquisa Heinrych pelas conversas que permitiram o avanço deste. Por fim, sou profundamente grato a minha orientadora Prof Drª Deborah Fernandes que traçou a rota para tornar este trabalho em realidade. A todos que me prestaram auxílio durante o meu processo estudantil, meu muito obrigado.

Ninguém é mais arrogante em relação às mulheres do que um homem que duvida da sua virilidade

Simone de Beauvoir,
O segundo sexo, 1949.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Segundo a filósofo Karl Marx em [26] *"não é a consciência dos homens que determina o seu ser, mas, ao contrário, é o seu ser social que determina sua consciência"*. A natureza do ser humano é se relacionar, se comunicar com o outro. Com o advento da internet a troca de informações em tempo real com pessoas ao redor do globo, se tornou algo comum. Logo, utilizar das Redes Sociais para expressar suas ideias e sentimentos é corriqueiro para a maioria das pessoas. Desse modo a afirmação da revista Forbes em [25], de que no ano de 2018, se produziram 2.5 quintilhões de bytes de dados por dia, é corroborada.

A expressividade das Redes Sociais, permite que uma variedade de pautas circulem entre os usuários. Essas pautas circulam temas simplórios como discussões sobre gosto musical dos usuários ou em temas delicados como as vivências pessoais de violência contra a mulher. Para o segundo caso, os comentários deixados em Redes Sociais podem ser tanto um alento, se forem gentis, como um outro caso de violência se não forem. Para poder discernir o que cada comentário é, alguma maneira de detecção de violência precisa ser utilizada.

A população brasileira é composta por mais de 200 milhões de brasileiros, segundo levantamento do IBGE em [6] sendo 51,1% destes, mulheres desta nação. Ainda assim, segundo o Portal da Justiça do Trabalho em [23], um terço das mulheres brasileiras já sofreram violência física ou sexual ao menos uma vez na vida. Mas o que é a violência contra a mulher?

Segundo [1] 'qualquer ação ou conduta, baseada no gênero, que cause morte, dano ou sofrimento físico, sexual ou psicológico à mulher, tanto no âmbito público como privado' é considerada violência contra a mulher. A violência de gênero realizada por meio virtual, o discurso de ódio a mulher, difundido por meio de Redes Sociais é um estorvo para a sociedade, como relata o Jornal O Dia em [7], na matéria de 2017, na qual apresenta o gasto consequente de violência contra a mulher como sendo de 1 bilhão de reais. Visto que existe uma dificuldade na detecção correta de ofensas em texto, o que atrasa a devida punição aos responsáveis. Diante disto, o dilema confrontado neste é a violência digital sofridas por mulheres, em específico em mulheres que participam da

rede social Twitter.

O objetivo desse trabalho é remover o ruído contido em textos da Rede Social *Twitter* no tema de violência contra a mulher. O tópico é um desafio pois o idioma português brasileiro conta com expressões ofensivas que podem ou não indicar violência, variando com o contexto em que são encontradas aplicadas. Ademais, o informalismo presente no diálogo pode elevar a dificuldade da tarefa. Juntamente a isto, a tarefa de processar texto, em virtude de que alguns modelos de predição foram projetados para lidar apenas com números, tem de ser considerada. Como já foi citado há um alto volume de dados existentes, sem embargo com nenhuma classificação prévia.

Para que seja possível remover o ruído no diálogo inerente das Redes Sociais, emerge a proposta da produção de uma ontologia que coopere neste sentido, ao eliminar dados que não se tratam a violência contra a mulher. Seguido da comparação dos resultados obtidos por meio de aprendizado de máquina, sem a ontologia e com a ontologia. Com o intuito de progredir o estado da arte em rumo a uma detecção confiável de violência digital contra as mulheres.

A meta desejada é que a ontologia construída seja eficiente durante a remoção do ruído presente em uma base de dados que abrigue traços de violência contra a mulher, e que os testes e validação utilizados para a comparação confirmem esta hipótese.

Em apoio a este processo, a monografia possui 5 capítulos, sendo este o primeiro deles, o segundo capítulo refere-se a Revisão Bibliográfica e indica o estado da arte quanto a pesquisa em Inteligência Artificial, o terceiro capítulo segue ao longo da bibliografia ao explorar as pesquisas vigentes quanto à produção e uso de Ontologias. O quarto capítulo observa o experimento realizado durante a pesquisa, o quinto e último capítulo pontua os resultados obtidos pela experimentação e os analisa.

CAPÍTULO 2

Revisão Bibliográfica

2.1 Trabalhos relacionados

Para identificar tweets misóginos no idioma Turco, os autores de [37] recorreram a algoritmos de *Machine Learning* (ML) dentre eles: *Decision Tree* (DT), *Random Tree* (RT), *Naïve Bayes* (NB) e *Support Vector Machine* (SVM). A pesquisa teve como etapa de pré-processamento a transformação das palavras em minúsculas. Para compor sua base de dados utilizaram de quatro anotadores humanos. Se três concordassem na análise de que existiu assédio o tweet pertenceria à classe 'possui assédio', caso contrário 'não possui assédio'. Após a etapa de classificação, os autores optaram por gerir uma base de dados perfeitamente balanceada entre as classes supracitadas. Na etapa seguinte a base foi dividida em 75% para treino e os 25% restantes para teste. Os autores decidiram por meio de teste, qual seria o mais eficaz nesse contexto, dentre: Agrupamento de palavras, de caracteres, de sentenças ou de sílabas. Para cada caso, se testou também o melhor tamanho para esses agrupamentos. Em seguida, utilizaram a técnica estatística TF-IDF para mensurar a importância de uma palavra em relação ao corpo do texto em que se localiza. Após sucessivos testes, uma SVM com as 50 palavras com os maiores TF-IDF teve o melhor resultado do estudo com precisão de 0.97, porém com um recall de 0.32, considerado ruim pelos autores. Nota-se como dilema da pesquisa a necessidade de analisar os dados manualmente para utilizar o aprendizado de máquina supervisionado.

O estudo dos pesquisadores brasileiros exposto em [39] objetivou verificar a qualidade da transferência de conhecimento entre domínios distintos. Para amparar seus anseios, fez-se uso de cinco *datasets* categorizados com dados colhidos da plataforma Twitter no idioma Português Brasileiro. Para tal, realizaram uma etapa de pré-processamento contendo remoção de *stop-words*, de caracteres especiais e acentos e conversão em letras minúsculas. Subsegue-se que por meio de um SVM-linear e parâmetros como distância de Jaccard, Euclidiana e de Cosseno, a tentativa de detectar os sentimentos dos eleitores. Ao sondar por maneiras de melhorar os resultados obtidos, os autores optaram por realizar a combinação de *dataset*s e avaliar seu impacto. As combinações escolhidas são: dois *datasets* mais similares, depois os dois mais díspares, seguido pela união dos três mais

próximos, assistido pelos quatro mais próximos, por fim com todos os *datasets*. Notou-se resultados aproximados mesmo com *datasets* dissimilares e que os melhores resultados da pesquisa, aconteceram ao unir os dois *datasets* mais correlatos. A união dos dois *dataset*s similares, alcançou o *F1-score* de 0.73, contra 0.67 e 0.66 das outras combinações. Além disso, uma melhor detecção ocorre quando os valores da distância de Jaccard e Euclides são afins. As dificuldades notadas na pesquisa foram o esforço humano em categorizar dados suficientes em tempo hábil (durante a corrida eleitoral), a volatilidade de termos, além da própria dissipação do conhecimento, causado pela variação do contexto local e temporal ao qual o conhecimento é aplicado

Ao realizar uma busca ampla sobre discurso de ódio no idioma indonésio, temos dois trabalhos na área, o trabalho dos pesquisadores de [20] com enfoque em *Machine Learning* clássico (ML) com classificação para múltiplas classes e o dos pesquisadores de [24] que compara *Machine Learning* e *Deep Learning* (DL). Em ambas etapas de pré-processamento houve a conversão de texto para minúsculas e a remoção de caracteres especiais, ressalta-se a distinção que os investigantes de [24], excepcionalmente não removeram o caractere especial "#", e mediram o impacto da remoção ou não de *stop-words*. Por sua vez realizou-se *stemming*, normalização de vocabulário e a remoção de todas as *stop-words* pelos cientistas que criaram [20]. Ambos recorreram a *datasets* com aproximados 13 mil tweets, no caso da investigação contida em [24] os dados foram desbalanceados. Ambos quando processaram seus *datasets* aplicaram *Support Vector Machine* (SVM) e *Random Forest Decision Tree* (RFDT), porém os pesquisadores de [20] aplicaram juntamente *Naïve Bayes Multinomial* (MNB), sobre outro prisma os cientistas de [24] aplicaram ademais *Bidirectional Recurrent Neural Network* (Bi-GRU) e uma Bi-GRU treinada com IndoBERT (um conjunto de dados de referências específico do idioma indonésio). Na observação criada em [24] os tweets foram classificados quanto à presença ou ausência de ódio, enquanto na visão contida em [20] os tweets eram classificados em: Há ódio, há abuso, há ódio a indivíduo, há ódio a um grupo, qual grupo recebe ódio, há ofensa, qual o grau de ofensa. No processo de desígnio multiclasses, utilizar o modelo SVM e separar os tweets palavra a palavra (unigrama de palavra), gerou a maior acurácia encontrada 68.43% pelos cientistas de [20]. Por sua vez, a escolha de manter as *stop-words* e utilizar a versão de Bi-GRU treinada com IndoBERT gerou a maior acurácia 84.77% para os investigadores de [24].

Com foco nas eleições quenianas, os pesquisadores de [33] desejavam gerar um *framework* para anotadores humanos que auxiliasse na detecção de discurso de ódio político. Engajados nesse propósito, construíram uma banca de 9 avaliadores, comprometidos a avaliar 4031 tweets que poderiam conter ódio. Sistematizaram além de que grupos contendo 3 avaliadores ponderavam sobre os três componentes de ódio segundo a literatura. Sendo eles, o afastamento do objeto de ódio, a paixão por algo que supostamente o objeto

de ódio pode lhe tomar e o compromisso de agir contra esse objeto de ódio. Onde existiu consenso entre os 3 anotadores, o tweet foi classificado como "possui ódio". Os dados analisados contavam com mais de 390 mil tweets, coletados entre 2012 e 2017 em Suáli e Inglês e por meio da classificação manual tornaram-se casos de teste confiáveis, para o aprendizado de máquina supervisionado. O processo exprimiu que dentre a amostra observada houve 18% de discurso de ódio, 11% de discurso ofensivo, 7% fora inconclusivo e os demais 64% não demonstraram ódio. Sob o discurso de ódio conclui-se também que 41% era distante, 30% era apaixonado, 23% era compromissado e 6% era de outro modo.

Na tentativa de minimizar ambiguidades em detecção de ódio geral, por meio de ontologias e lógica *fuzzy* temos o trabalho dos investigadores de [32]. Ao usar três *datasets* construídos e testados por outrem, os autores tiveram bases sólidas para poder se concentrar na vinculação da ontologia com a lógica *fuzzy*. Dessa forma, propuseram a criação de uma ontologia onde se limpava os dados de entrada, se removia o corpo do texto tags, menções e URLs. Sem demora, se avaliava a existência das palavras na ontologia, e caso não existissem, adicionavam-as, mensura ademais o ódio de cada palavra prévia a sua armazenagem. Posteriormente, durante a etapa de fuzzificação, por meio de função triangular, *left-shoulder* e *right-shoulder*, gerava membros. Por seguinte, a etapa de defuzzificação usava de centróide por prover uma defuzificação mais completa e um melhor retorno entre estados. Após a avaliação do processo, notou-se uma melhoria na detecção da classe de ódio, de ofensa e no recall da pesquisa, com valores de 0.4514, 0.9805 e 0.677 respectivamente, contra os obtidos em um trabalho anterior dos autores, que possuíam valores respectivamente de 0.44, 0.91 e 0.61. Denota-se o acréscimo proporcionado pela adesão das ontologias e lógica *fuzzy*.

Ao questionar a origem do antinacionalismo na população, os cientistas de [28], avaliaram o comportamento de uma parcela da sociedade, por meio da linguagem R. Para iniciar a pesquisa, os dados obtidos foram limpos, tokenizados e passaram por *stemming*. Seguinte fizeram uso de Matriz de Termos do Documento (DTM como forma de relacionar termos e documentos associados a eles), juntamente de TF-IDF (Term Frequency-Inverse Document Frequency, uma técnica para ponderar a recuperação de dados por mineração de texto), *Bag of Words* (uma forma de tornar o texto o multiconjunto de suas palavras e facilitar sua interpretação com base nisto), e N-grama de palavras (agrupar N palavras por vez, para melhorar seu entendimento) sendo todas técnicas para obter melhora na extração de conhecimento. Assistida pelo aprendizado de máquina supervisionado com *Naïve Bayes*. Ulteriormente mediante validação 5 vezes cruzada e avaliação de Área Abaixo da Curva (AUC), o algoritmo autoral proposto atingiu 74.14% de acurácia, o que ultrapassou os 70% atingidos pelos aprendizados de máquina clássicos. Ao considerar o feito, os autores levantam a hipótese de que o ódio cresce juntamente da quantidade de usuários nas redes sociais.

Pelo anseio por compreender o estado da arte na identificação automatizada de discurso de ódio, com ênfase no discurso racista e sexista, o trabalho dos pesquisadores de [2] emerge. Os cientistas realizaram uma revisão sistemática sobre o tema na língua inglesa, com pesquisas em dados obtidos pela rede social Twitter, com posterior exibição dos dados levantados. Desta maneira informam que o *dataset* mais utilizado fora 'Wasem and Hovy', *dataset* com 16914 tweets, dos quais há 3383 tweets sexistas, 1972 tweets racistas e 10645 que não se enquadram em nenhum dos casos. Os autores exibem as técnicas de processamento de texto mais utilizadas, citando primeiro as mais frequentes, sendo essas a TF-IDF, N-grama de caractere, N-grama de palavra, Unigrama, Vetor de *Bag of Words*, *Global Vector* (GloVe), *Random Embedding* (agrupamento aleatório). Seguem ao informar que as técnicas de *Machine Learning* (ML) mais usadas foram *Support Vector Machine* (SVM) e *Gradient Boosting Decision Tree* (GBDT). Enquanto que as *Deep Learning* (DL) mais empregues foram *Convolutional Neural Network* (CNN) e *Long Short Term Memory* (LSTM). Ao sumarizar os percalços encontrados constataram que a falta de dados de treino, a necessidade do pré-processamento para utilizar dos dados, acrescidos de um balanceamento irreal em tempo de coleta, e um custo material, digital e legal de capturar mais dados, são desafios recorrentes para pesquisadores da área.

O trabalho dos instigadores que compõem [42] teve enfoque em observar como os cidadãos se sentiam com a criação da lei que criminaliza nove tipos de violência sexual. Os autores buscaram em tweets no idioma indonésio discurso de ódio com tema sexual, coletaram então 1015 tweets sobre o assunto. Para analisar corretamente os dados obtidos, pré-processaram os tweets ao tokenizar e remover *stop-words*. Por meio de três técnicas de aprendizado de máquina supervisionado, *Decision Tree* (DT), *Naïve Bayes* (NB) e *Random Forest* (RFDT) classificaram os textos em 'positivo', 'neutro', 'negativo' quanto aos sentimentos dos usuários perante a lei. Nesse contexto a aplicação de *Naïve Bayes* obteve melhor resultado ao atingir acurácia de 83.94% e *F1-score* de 59.41, seguido por *Random Forest* com 75.72% e 28.72 respectivamente, por fim, fica a *Decision Tree* com a acurácia de 75.31% e *F1-score* de 28.70. A pesquisa ilustrou o apoio da população à lei. Como dilema de pesquisa, ressalta-se a sensibilidade do tema.

O desígnio de apontar tendências de técnicas durante a mineração de texto, para detecção de ódio em indonésio e em inglês, foi o que motivou a pesquisa dos investigadores de [35]. Os autores decidiram analisar 38 trabalhos na área, pesquisados com as palavras-chave *"hate speech abusive ujaran kebencian"* em bases de dados como ACM, IEEE, ScienceDirect, Springer e Garuda. Como resultado apontaram que a coleta de dados para mineração de texto provém das mais diversas plataformas, sendo as maiores fontes as redes sociais Twitter, Facebook e YouTube. Inferiram que o uso de um bom *dataset* afeta positivamente os resultados provenientes da mineração textual. Constataram de mesma forma que o uso de técnicas N-grama de palavras, N-grama de caracteres,

word *embedding*, análise de sentimentos, análise de sintática e análise de semântica, lexicon e *Bag of Words* são escolhas frequentes para a área. Notaram ademais o uso do processamento automatizado por aprendizado de máquina por meio de *Support Vector Machine*, *Linear Regression*, *Naïve Bayes*, *Random Forest*, *Long Short Term Memory*, *Decision Tree*, e aprendizado profundo por meio de *Convolutional Neural Network*, Recurrent Neural Network e Biological Neural Network, como as maneiras mais aceitas para extração automática de informação. Por fim, acentuaram a preferência por classificar o discurso em duas frentes, 'há ódio' e 'não há ódio'.

Com o intuito de utilizar os conceitos sociológicos e a tecnologia da informação para percorrer tweets em espanhol e identificar ódio contra a mulher, como forma de auxílio para a prevenção e proteção da mulher pelas forças legais competentes, a pesquisa dos cientistas de [3] se fez. Como base de sua pesquisa, os autores consideram os conceitos de ódio contra a mulher definidos por Bourdieu, Richardson-Self e a lei de defesa contra a violência digital contra as mulheres no México. Uma decisão de pesquisa foi, ao encontrar termos enquadrados como 'ódio estrutural' considerar-lhes termos de discurso de ódio. Para mapear a violência contra a mulher aplicaram algoritmos baseados em representação, clustering e semântica latente. Para tal emergiu a necessidade de uma etapa de pré-processamento, onde os dados passaram por uma CNN multitarefa e pela biblioteca spacy da linguagem Python, ambas treinadas em língua espanhola. Quando os autores compararam os *datasets* Zenodo, SemEval 1, SemEval 2 e SemEval 3, por meio de técnicas de classificação simplista obtiveram, 78.3%, 64%, 68.8%, 78.9% de acurácia respectiva, abaixo do considerado útil para auxiliar as forças legais (90% estipula-se), mesmo com o uso da técnica de mitigar viés de Claudia Volpetti em [46]. Fator que corroborou com a motivação da pesquisa, mas indica uma demanda de refinamento para com seus resultados.

Por intervenção de técnicas de aprendizado não supervisionado, os pesquisadores em [14] anseiam por checar a possível presença de discurso de ódio contra as mulheres, nos arredores da cidade de Monterrey. Para iniciar a pesquisa, os autores coletaram dados da plataforma Twitter usando sua API nativa e o sistema de geolocalização ativado para receber tweets apenas na região de interesse, a cidade de Monterrey, capital do estado de 'Nuevo Léon' no México. Em decorrência desse processo, cerca de dois milhões de tweets foram coletados, no período de 20 de setembro a 20 de novembro. Como maneira de elucidar as informações contidas nesses dados, os autores decidiram usar das técnicas de k-média (classificar os dados nos chamados grupos ou clusters, no caso da pesquisa foram oito clusters, a seleção de alocação de um valor para um cluster específico é decidida pela média de seus valores), *Bag of Words* (forma de associar o texto as palavras associadas dentro dele, independente da ordem em que elas apareçam), PCA (Principal Component Analysis, maneira de detectar dados úteis ou não, em um conjunto, dada sua variância) e

elbow (técnica para encontrar a quantidade ideal de clusters k). Logo após o treino fora realizado com o *dataset* Hateval, e testado primeiramente com o *dataset* disponibilizado pelo Kaggle e posteriormente, na base de dados coletada, para classificar o tweet como "apresenta ódio"ou "não apresenta ódio". À vista disso a validação dos dados ocorre com o recurso de acurácia pura. Afinal a CNN nativa da biblioteca Python Scikit-learn, obteve 75% de precisão e encontrou dentre os tweets no escopo de Monterrey 19.8% de discurso de ódio contra a mulher.

O cerne da pesquisa dos investigadores de [22] foi a identificação da presença ou ausência de misoginia em textos da plataforma de dicionário informal Urban Dictionary. Motivados pelo baixo uso de *Deep Learning* (aprendizado profundo) na área de busca por ódio, os autores escolheram utilizar Bi-LSTM e Bi-GRU e testar sua eficácia quando comparados a métodos de inteligência artificial clássicos como *Linear Regression* (LR), *Naïve Bayes* (NB) e *Random Forest* (RFDT). As redes escolhidas para tal foram RNN (Recurrent Neural Network) um tipo de rede neural com atualização dinâmica. Para pré-processar os dados, geraram um vetor de números (representações de palavras) ordenado, encriptaram esse vetor como uma lista de inteiros, recorreram ao padding para manter todos os vetores de mesmo tamanho (igual ao tamanho da maior palavra). Em seguida, Bi-LSTM foi configurado com 3 camadas com 50 células LSTM em cada camada e camada de rejeição valendo 30%, para prevenir over-fit nos dados. Enquanto o Bi-GRU teve 2 camadas de 50 células GRU e uma camada de rejeição de 15%. Ambos foram processados com sigmóide duro e tangente hiperbólico. Dentre os 2285 textos coletados do Urban Dictionary, foram utilizados 80% para treino (2056), e os demais 20% (229) para teste, um *dataset* levemente desbalanceado, onde 1034 textos continham misoginia e 1251 não a continham. Por meio da função de Mean Square Loss (perda quadrada de média) e validação 10 vezes cruzada, se observou que polir dados para processar é custoso, notaram também que sarcasmo e humor são difíceis de detectar. Os resultados mostram que nesse contexto, o Bi-GRU obteve a maior acurácia 93.10%, maior sensitividade 92.08% e segunda melhor especificidade com 93.96%, seguido logo de perto por Bi-LSTM com 90.12%, 87.53% e 92.41% respectivamente, e em terceiro classificou-se o *Random Forest* com 89.50%, 81.64% e 96% respectivamente, possuindo RF a melhor performance quanto a especificidade.

Motivados pelo âmago de entender ódio entre falantes de línguas diferentes, em específico Inglês, Hindi e o dialeto que mistura os dois o 'Hinglish', com um pouco de Bengali, os pesquisadores de [4] propuseram seu trabalho. No qual coletaram 3189 tweets em textos nos idiomas e dialetos estudados, para compreender esses dados, precisaram de uma etapa de pré-processamento. Nessa etapa removeram *stop-words*, links, arrobas e transformaram o texto para minúsculas, além de usarem a biblioteca Python fastext para comprimir as palavras para manipularem em *Machine Learning* (ML) e *Deep Learning*

(DL) posteriormente. Na etapa seguinte, 80% dos tweets (2538) foram usados para treino e os demais 20% (651) para teste. Por meio do Keras, uma biblioteca para implementações rápidas de DL, os dados foram transmutados. As decisões de transmutação foram a variação de agrupamento de tweets (N-grama) primariamente com dois, e sob sucessivos testes realizaram até cinco agrupamentos, acrescidos do uso de 20 épocas, além de tamanho da batch com o valor 16, ademais o uso de Softmax na última camada (função de ativação para normalizar os dados de saída preditos), nas demais Relu (função de ativação para permitir comparação mais veloz de grandes volumes de dados). A pesquisa avaliou os impactos de aprendizados de máquina clássicos sozinhos e quando agrupados (ensemble), onde sua validação ocorre com acurácia e *F1-score*. Os melhores resultados obtidos para ML nos seguintes métodos Logistic Regression, *Random Forest*, *Support Vector Machine* e Multinomial *Naïve Bayes*, foram em *F1-score* respectivamente, 0.827, 0.803, 0.821 e 0.757. Aliás, ao usar modelos de DL treinados com cem, duzentas ou trezentas camadas, como melhores resultados obtiveram 0.834 para uma CNN de 200 dimensões e 0.86 para uma Bi-LSTM de 300 dimensões. Sem delonga, ao realizarem o agrupamento (ensemble) de modelos de ML que usavam Count Vectorizer obtiveram 0.847, o agrupamento de modelos de ML que usavam de TF-IDF Vectorizer resultou em 0.737, enquanto o agrupamento empilhado (stacked) de modelos de DL resultou em *F1-score* de 0.873. Isto posto ilustra o melhor resultado no contexto da pesquisa. Ainda assim, os autores sentiram a falta de uma base maior de treino e assim como, sentiram dificuldade de transliterar frases entre as regras pouco claras do dialeto Hinglish. Em suma, indica que métodos de DL agrupados são melhores que ML agrupados, que por sua vez são melhores que métodos de DL isolados, sendo esses ainda assim, melhores que métodos de ML isolados.

Pelo anelo de decifrar o comportamento nas redes sociais Twitter e Facebook como se comportam os falantes de Afan Oromo, um dialeto índigena etiópe, quanto a étnica. política e religião, os cientistas de [11] se motivaram. O *dataset* utilizado na pesquisa possuia 13600 tweets, sendo este desbalanceado onde 6468 tweets possuiam ódio e 7138 não possuiam. A seguir os autores pré-processaram e padronizaram os tweets quanto à escrita, onde colocaram seu texto em minúsculas, e removeram *stop-words* e quaisquer pontuação exceto "hudhaa". Segue que para compreender esses dados os autores optaram por implementar algoritmos variados de *Machine Learning* (ML), sendo esses, Linear Support Vector Classifier (LSVC), *Support Vector Machine* (SVM), *Decision Tree* (DT), *Linear Regression* (LR), *Random Forest* (RFDT) e um *Multinomial Naïve Bayes* (MNB), e depois implementaram a sua versão *"tuning"* para encontrar hiperparâmetros ótimos em essência (formas de encontrar os melhores resultados possíveis, sem gerar over-fit) para os algoritmos de propósito geral selecionados. Sem demora na etapa de treino, separaram 67% do *dataset* para treino e os demais 33% para teste. Em seguida como

medida de validação, fizeram uso de *F1-score* e acurácia para comparação. Ademais os resultados no caso demonstram uma superioridade dos métodos com o *"tuning"* em relação aos seus equivalentes clássicos. Dado reforçado pelos valores encontrados para os clássicos: LSVC, SVM, DT, LR, RF, MNB que atingiram respectivamente: 66%, 66%, 59%, 65%, 64%, 66%. Por outro lado, após o processo se obtiveram respectivamente: 91.4%, 91.85%, 87.78%, 91.40%, 91%, não se possuia dados sobre o novo valor de MNB. Ainda assim, salientam que sim, existe ódio étnico, político e religioso no *dataset*, e que o processo de *"tuning"* ajuda a identificá-lo mais facilmente. Por fim se destaca como dificuldade do estudo a sensibilidade das técnicas a um grande volume de dados.

A abordagem contida neste foi a de utilizar algumas das técnicas de classificação conduzidas pelos pesquisadores de [11] e [4], sendo elas *Linear Regression*, *Multinomial Naïve Bayes*, *Random Forest*, porém sem a proposta de *"tuning"* de [11] ou a abordagem de outros idiomas de [4]. De maneira similar aos trabalhos de [37], [33], [28] e [42] que lidam com vários tipos de discurso de ódio, foi utilizado Aprendizado de Máquina Supervisionado para o treino dos modelos. A distinção principal aqui é a aplicação de ontologia como uma maneira de remover ruído do discurso, que exista nos *tweets* da base de dados, onde se almeja manter somente os que possuam traços de violência contra a mulher e o aprendizado de máquina foi uma das formas para mensurar o ganho obtido.

Tabela 2.1: Tabela sumária dos trabalhos relacionados

Nome do estudo	**Objetivo**	**Idioma**	**Técnicas**	**Resultado**
Detection of Hate Speech towards women on twitter	Identificar misoginia	Turco	DT, RT, NB, SVM	Acurácia de 0.97, recall de 0.32
Combining labeled datasets for sentiment analysis from different domains based on dataset similarity to predict electors sentiment	Identificar ódio político	Português Brasileiro	SVM-linear	F1-score de 0.73
Annotation framework from hate speech identification in tweets: Case study during Kenyan elections	Identificar ódio político	Sualí e Inglês	Framework próprio para anotadores humanos relatarem sobre discurso de ódio	Krippendorff' alpha de 0.507
HateSense: Tackling ambiguity in hate speech detection	Minimizar ambiguidades com ontologia e lógica fuzzy	Não especifica	Ontologia, lógica fuzzy	Acurácia de 0.882
Hierarchical multilabel classification to identify hate speech and abusive language on Indonesian tweet	Identificar ódio em textos indonésios	Indonésio	SVM, RFDT, MNB	Acurácia de 0.6843
Crowdsourcing of hate speech for detecting abusive behavior in social media	Identificar discurso de ódio de cunho antinacionalista	Não especifica	NB com validação cruzada	Acurácia de 0.7414
Sentimental analysis of social media users using Naïve Bayes, Decision Tree, Random Forest algorithm: A case study of draw law on the elimination of sexual violence	Identificar sentimentos dos cidadãos sobre a nova lei	Indonêsio	NB, DT, RF	Acurácia de 0.8394 F1-score de 0.5941
A comparison of Machine Learning approaches for detecting misogynistic speech in urban dictionary	Identificar discurso de ódio de cunho misógino	Não especifica	Bi-LSTM, Bi-GRU, LR, NB, RF	Acurácia de 0.931 Sensitividade de 0.9208 Especificidade de 0.9396
Systematic literature review of hate detection with text mining	Reunir conhecimento sobre discurso de ódio	Indonésio e Inglês	Revisão sistemática	SVM, LR, NB, RF, CNN, LSTM, DT, RNN, BNN foram as IAs mais utilizadas
Understanding violence against women in digital space from a data science perspective	Identificar discurso de ódio de cunho misógino	Espanhol	CNN	Acurácia de 0.789
A sentiment analysis and unsupervised learning approach to digital violence against women: Monterrey Case	Identificar discurso de ódio de cunho misógino nos arredores de Monterrey	Espanhol	CNN	Acurácia de 0.75
Ensemble bases Hinglish hate speech detection	Identificar discurso de ódio Por meio do dialeto Higlish	Hinglish	LR, RF, SVM, MNB, Bi-LSTM, CNN, ensemble das técnicas supracitadas	F1-score 0.873
Hate speech detection Indonesian Twitter texts using Bidirectional Gated Recurrent Unit	Identificar discurso de ódio	Indonésio	RF, SVM, Bi-GRU, Bi-GRU com IndoBERT	Acurácia de 0.8477
Racist and sexist hate speech detection: literature review	Reunir conhecimento sobre discurso de ódio racista e machista	Inglês	Revisão sistemática	SVM, SBDT, CNN, LSTM foram as IAs mais acertivas
Tuning Hyperparameters of Machine Learning Methods for Afan Oromo Hate Speech Text Detection for Social Media	Identificar discurso de ódio e verificar o nível de melhoria ao se dar 'tuning' nos parâmetros	Afan Oromo	LSVC, SVM, DT, LR, RF, MNB	Acurácia de 0.9175

Elaborado pelo autor

CAPÍTULO 3

Base teórica

Durante a realização do experimento será necessário produzir uma ontologia, ser capaz de aplicá-la fronte a torrente de dados disponíveis e avaliar seus efeitos nos dados. Portanto, a teoria contida no capítulo é uma ferramenta para uma execução correta da aplicação prática do capítulo subsequente. Neste capítulo há a conceituação do que é a violência contra a mulher, sendo o combate a violência a maior motivação para a existência deste, seguido pelos processos necessários para se compreender, criar e gerenciar ontologias, além das opções existentes para o processamento automatizado de texto.

3.1 Discurso de violência contra a mulher no meio cibernético

De acordo com o Relatório da ONU contido em [13] até os anos 90, em países latino americanos 50% de todas as mulheres já relataram algum tipo de violência sexual ou física por seus parceiros, além disto em outro estudo citado por [13] em 2019, quando perguntaram a mulheres brasileiras acima dos 16 anos, se em algum momento dos últimos 12 meses, elas haviam sofrido de violência, 28% delas responderam que sim e quase metade delas, relataram ter mantido o ocorrido em segredo. Pode-se notar um progresso para a diminuição da violência mas os valores altos ainda surpreendem negativamente.
A definição de violência segundo as pesquisadoras de [1] é “qualquer ação ou conduta, baseada no gênero, que cause morte, dano ou sofrimento físico, sexual ou psicológico à mulher, tanto no âmbito público como privado". A violência psicológica citada aqui ocorre de muitas maneiras, via agressão verbal, hostilização pública e por meio de linchamento virtual, o *cyberbullying*.
Em [38] o *cyberbullying* é descrito como a versão a distância do *bullying* mantendo as características básicas de sua variante, como a presença de agressores, vítimas e espectadores, além da inteção de promover dor e sofrimento ao alvo. A pesquisa de [38] ainda reforça que o potencial de dano causada pela agressão virtual é maior, visto a rápida

divulgação de conteúdos *online* e a dificuldade para a remoção completa de tais atos da web.
Como forma de combate ao ato nefasto da violência de gênero, o governo do Brasil criou leis que lidam com esse tipo de violência, descritas em [19] e [5] a 'Lei Maria da Penha' de 2006 com atuação no combate a violência doméstica, a 'Lei Carolina Dieckmann' de 2012 que criminaliza crimes virtuais, de roubo e exposição indevida de dados privados, a 'Lei do Feminicídio' de 2015, que combate as mortes por ódio as mulheres e a 'Lei 14188' de 2021, responsável por criminalizar violência de genêro psicológica, fatores relevantes para a preservação da saúde pública e o bem-estar geral das cidadãs em geral.

3.2 Ontologias

Ontologias são ferramentas de representação formal de conceitos e seus relacionamentos semânticos, sendo comumente criadas por especialistas sobre algum tema para melhorar a precisão de suas definições. Elas informam os principais aspectos sobre um conceito e não todos os fatores que o compõem. A palavra provém da junção dos termos gregos *"ontos"* e *"logos"* que significam, respectivamente ser e palavra. É livremente traduzida como "a palavra que define o ser".

As ontologias são um campo de estudo em várias áreas, tendo maior destaque em três delas: a Filosofia, a Ciência da Informação e a Ciência da Computação. O foco do estudo ontológico perante a filosofia é o processo de categorização do ser, quanto a sua existência e características, tais como as visões de categorização distintas de Aristóteles, com sua predicação, Kant e seu julgamento, de Husserl. Por outro lado, o foco ontológico para a Ciência da Informação está em definir informalmente o conceito de ontologia, para compreender um domínio e classificá-lo quanto a termos, somada ao seu uso como forma de gerar vocabulários consistentes para a reuso da informação, como os catálogos e tesauros. Por fim, na Ciência da Computação, ontologias são tratadas como um aspecto teórico que permite a modelagem de um domínio, além de ser um artefato de software responsável por representar inferências e sistemas de maneira padronizada, como as linguagens OML, DAML, dentre outras.

As definições de ontologias variam com o contexto em que se encontram, sendo para a Ciência da Computação tanto "uma hierarquia de conceitos do domínio, suas relações e leis que apresentam"em [31], mas também "uma especificação explícita e formal de uma conceptualização partilhada"em [17].

Segundo os autores de [9] ontologia é em termos sucintos, "(um) catálogo de tipos de coisas", porém para o pesquisador de [15] ontologia é:

> "uma especificação explícita de uma conceitualização [...] Em tal ontologia, definições associam nomes de entidades no universo do discurso (por exemplo, classes, relações, funções etc.). Com textos que descrevem o que os nomes significam e os axiomas formais que restringem a interpretação e o uso desses termos)[...]".

Já o cientista de [17] define ontologia como:

> "[...] se refere a um artefato constituído por um vocabulário usado para descrever uma certa realidade, mais um conjunto de fatos explícitos e aceitos que dizem respeito ao sentido pretendido para as palavras do vocabulário. Este conjunto de dados tem a forma da teoria da lógica de primeira ordem, onde as palavras do vocabulário aparecem como predicados unários ou binários".

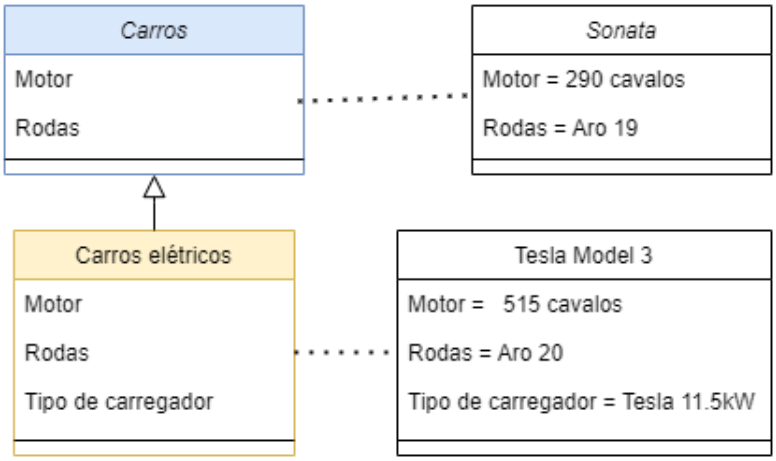
Formalmente ontologia é um par $<D, W>$ tal que D é o domínio a ser estudado e W são os conceitos presentes nesse domínio. Por intervenção da compreensão do espaço de domínio, produz-se uma conceitualização $C = <D, W, R>$ pela qual adiciona-se ao domínio e conceitos, as relações existentes R entre estes conceitos. Para uma correta representação do mundo, se estabelece a representação almejada S e uma linguagem L que é a conversora da conceitualização em tal representação, desta forma $L(C) \subseteq S$.

Para compor uma ontologia existem termos necessários, e estes são aceitos universalmente, sendo as Classes, as Subclasses, as Superclasses, os Atributos, as Relações e as Instâncias. As definições seguintes são uma adaptação das explicações originalmente descritas por Morais & Ambrósio em [29].

Primeiramente temos as classes, que são uma categorização, um agrupamento de objetos ou seres similares mediante um domínio, um exemplo disto é a classe 'Carros' que reúne todos os veículos automotivos movidos por rodas. Em seguida temos os Atributos que são os componentes, são as características intrínsecas aos membros de determinada classe, todo carro possui 'rodas' com uma determinada espessura e um 'motor' com uma quantidade de cavalos de potência.

O conceito posterior é o de Relações que é a forma de associar duas classes. No nosso exemplo, podemos citar então a classe 'Carros elétricos' que reúne os veículos automotores de quatro rodas, com tração por motor elétrico. Uma maneira para se relacionar 'Carros' e 'Carros elétricos' é dizer por exemplo que todo carro elétrico 'é um' carro. Deste modo, no conjunto de todos os 'Carros' existentes, existe uma parte menor destes que são os 'Carros elétricos'.

Em razão desta relação se pode observar outros dois conceitos, as subclasses e as superclasses. Por ser a classe mais importante nesta análise, 'Carros' é a classe principal, a também chamada 'superclasse', por 'Carros elétricos' serem uma variação da anterior, 'Carros elétricos' são a 'subclasse'.



Elaborado pelo autor

Figura 3.1: Exemplo sobre termos presentes em uma ontologia

Por mérito de existir uma subclasse, é necessário discutir o conceito de 'Herança' entre classes, no qual é definido que todos os atributos contidos na superclasse também existirão na subclasse, portanto as 'rodas' e 'motor' atributos de 'Carros' também compõem 'Carros elétricos'. Nada obstante de subclasses possuírem atributos exclusivos que apenas tem valor em seu próprio contexto, para 'Carros elétricos' o atributo 'tipo de carregador' é relevante enquanto que para os demais 'Carros' essa característica não é.
O termo final para compor uma ontologia é a Instância, onde se faz menção a elementos específicos, no qual se atribui valores nos atributos da classe escolhida e observa se a representação obtida pela ontologia cumpre seus propósitos ou necessita de melhorias. No exemplo disto seria a representação de modelos de carros reais, na qual se fez uso do Tesla Modelo 3 e do Sonata, para o propósito educacional do exemplo, o processo foi um sucesso.

A conjectura sobre veículos levantada aqui pode ser visualizada pela figura 3.1, nesta representação há caixas com duas divisões, na divisão superior fica o nome da classe que está sob análise, na divisão inferior, cada linha representa um atributo presentes nesta classe. Para exprimir a presença de uma subclasse, se faz uso da seta com um triângulo que liga o bloco 'Carros' ao bloco 'Carros elétricos'. Enquanto que uma reta pontilhada entre dois blocos diz que o bloco à esquerda é uma classe e o bloco a direita é uma instância daquela classe.

Considerando a pluralidade de propósitos e usos para as ontologias, nota-se a necessidade de agrupá-las. Por meio de abordagens dissemelhantes, elevam-se classificações de ontologias, quanto ao: Formalismo definido pelos autores de [44], quanto a aplicação definido pelos cientistas em [45], quanto a função definido pelos pesquisadores de [27] e quanto ao conteúdo definido pelos idealizadores de [12]:

- Ao formalismo: Que pode ser altamente informal, usando apenas de linguagem

natural; semi-informal, usando linguagem natural estruturada; semi-formal, usando de uma linguagem artificial com definições formais; além do rigorosamente formal, que possui semântica formal, teoremas e provas em sua estruturação.

- A aplicação: que pode ser de acesso comum, para igualar entre as partes o conhecimento sobre dado conceito; de autoria neutra, para ser amplamente reutilizável em outros contextos; e de especificação, para auxiliar na documentação e manutenção do software.
- A Função: pode ser de tarefa, onde exprime as ações necessárias para solucionar um certo dilema; de domínio, para definir um micro-mundo particular e seus desafios; genérica, para explicar fenômenos naturais, físicos, químicos, temporais e espaciais.
- Ao conteúdo pode ser de conhecimento, para criar conceitos bem formados; de domínio, onde descreve o domínio do conhecimento somado às restrições na sua estrutura e no seu conteúdo; genérico, para conceituar termos em alto-nível, de maneira a serem reutilizáveis em múltiplas áreas de conhecimento; de aplicação, para explicar minuciosamente termos de um contexto pequeno e específico. Além das divisões em representação, para auxiliar na expressão formal de termos; terminológica, para modelar termos exclusivos de certo domínio do discurso; de informação, para organizar de forma eficiente e estruturada dados; e modelagem de conhecimento, para mapear termos específicos do contexto aplicado, por meio de uma semântica rica.

Neste projeto final de curso deseja-se avaliar os impactos do uso de ontologia como ferramenta reveladora de conceitos e guia de terminologia, na presença de ruídos constituintes da mineração de texto, provinda de *tweets*. E dessa forma, exemplificar o uso de ontologias para melhoria na identificação de violência contra a mulher no idioma Português Brasileiro. Para a aplicação do conceito de ontologia na realização deste, fez-se uso de uma ontologia semi-formal, de acesso comum, terminológica e de domínio que será melhor descrita a seguir.

3.2.1 Metodologias, ferramentas e linguagens para Ontologias

O estabelecimento de um padrão de produção é uma característica presente em conteúdos bem definidos perante a sociedade. A divisão de uma atividade complexa em várias atividades menores mais simples, é um padrão de projeto conhecido como 'dividir para conquistar' em computação, e possui explicação em [40]. As descrições logo após identificam visões díspares dentre autores para tal atividade, mesmo com etapas distintas entre si, o resultado final de todos é o mesmo, a geração de uma nova ontologia.

- **Enterprise** vide [43]: Propõem-se identificar o propósito da ontologia, mapear seu conceito-chave, o domínio de estudo e os relacionamentos existentes nesse domní-

nio. Seguido da codificação da ontologia, sua verficação de validade e documentação da ontologia criada.

- **Gruninger & Fox** vide [16]: Propõem-se descrever a motivação do estudo, gerar questões de competências inicias, para iniciar a formalização da ontologia mediante a formalização das questões de competência e de seus axiomas. A etapa final do processo é checar a validade e completude da ontologia produzida.
- **Menthology** vide [21]: Propõem-se a executar cinco etapas, a Especificação onde se planeja tarefas, quanto ao tempo e recurso exigirão. A Conceitualização onde se executa e checa a qualidade das tarefas propostas, ao explicitar como, por quê e quem age na ontologia. Em sequência há a etapa de Formalização onde conceituam os signos a nível de conhecimento, com a transformação para um modelo formal, semi-computável. A próxima etapa é a de Implementação da ontologia onde se traduz este modelo formal para linguagem para ontologias. A quarta etapa é então a Manutenção que serve como uma etapa de verificação, observa-se a ontologia para avaliar a funcionalidade da ontologia, ela atende o papel proposto durante sua criação, se sim, se para a próxima etapa; se não realizam-se alterações na ontologia, para que ela se torne útil conforme o proposto. A etapa final diz respeito a documentar os passos realizados nas demais etapas.
- **On-To-Knowledge** vide [41]: Propõem-se realizar o 'pontapé inicial' que é a etapa de projeção de uma ontologia simplificada, baseada somente o conhecimento comum e o observado em outras ontologias. Segue-se para o refino e avaliação desta criação, identificando itens a melhorar por meio da gestão do conhecimento. A posteriori fica a etapa de gerenciamento e manutenção da ontologia produzida.
- **Ontology Methodology 101** vide [31] Propõem-se que se determine o Domínio e escopo da ontologia na primeira etapa, ao gerar questões de competência. Em seguida, considera-se ontologias existentes e a enumeração de termos importantes. Após tem-se a definição de Classes e sua Hierarquia, além das Propriedades de Classes e das Propriedades dos Atributos, ao fim são criadas as instâncias da ontologia.

A produção de uma ontologia é um processo cíclico de produção e avaliação do produto gerado, onde existe uma relação entre termos mais simples e mais complexos, classes com suas instâncias, relações entre classes e atributos de classes, graças a quantidade de informação salva em cada item, o suporte visual é uma vantagem pela qual se permite uma criação consistente de ontologias mais rapidamente. Como solução para este dilema, emerge a necessidade de uma ferramenta que satisfaça estas necessidades. As ferramentas que são usadas para o desenvolvimento das ontologias são descritas a seguir:

- **Protégé 2000:** Criada pela Universidade de Stanford [1] é uma plataforma gratuita de criação, integração e modificação de ontologias. Possui interface gráfica e exporta arquivos nos formatos RDF, OWL, OBO, Turtle, JSON e LaTeX.
- **OilEd:** Criada pela Universidade de Manchester [2] é um editor de ontologias focado na lianguagem DAML+OIL. Possui menos funcionalidades que os demais, sendo essencialmente um bloco de notas para desenvolver ontologias. Entretanto, não é mais suportado ou mantido.
- **WedODE:** Criada pela Universidade de Madrid [3] é um conjunto extensível de engenharia de ontologia. Possui interface gráfica e exporta arquivos nos formatos XML, RDF, OIL, DAML+OIL, OWL, CARIN, FLogic, Jess, Prolog. Edita axiomas por meio de WAB (WebODE Axiom Builder). Entretanto, não sofre suporte ou manutenção desde o ano de 2006.

Durante o processo de produção de ontologias, pode-se recorrer a linguagens específicas para este tópico, estas contam com ferramentas que facilitam a descrição do conteúdo desejado, ao acelerar o desenvolvimento portanto, um exemplo de linguagem para tal é a DAML+OIL[4] uma linguagem de marcação focada em desenvolvimento Web, caracterizada pela semântica limpa e bem definida, e utiliza de conceitos de XML. Outro exemplo reconhecido é a OWL[5] uma linguagem que objetiva gerar ontologias não ambíguas para Web, sua linguagem computacional é baseada em lógica, assim como DAML+OIL, a OWL utiliza de conceitos de XML, todavia juntamente de conceitos RDF. Uma amostra adicional de linguagem é a Turtle, que se assemelha a OML ao usar conceitos de XML e de RDF, e com a DAML+OIL pelo propósito Web, porém tem seu diferencial pela capacidade de criação de grafos RDF, sua compatibilidade com o padrão existente para triplas e o suporte para N-Triplas.

3.3 Ruídos

Para o dicionário *online* Dicio em [8] o ruído é 'Som indistinto, sem harmonia; entrondo', assim como 'Som de várias vozes ao mesmo tempo; gritaria, tumulto', estas definições não compreendem por completo o ruído digital a qual este trabalho tenta combater. Ainda segundo [8], a definição de ruído na eletrônica é a 'designação genérica de todos os defeitos que perturbam ou impedem uma transmissão radiofônica ou telefônica'.

[1] htto://www.stanford.edu/
[2] httos://www.manchester.ac.uk/
[3] htto://www.fi.upm.es/
[4] httos://www.w3.org/TR/daml+oil-reference
[5] httos://www.w3.org/2001/sw/#owl

Porém, com relação a este trabalho, o problema está em lidar com o conceito de ruído contextualizado quanto ao aprendizado de máquina. Neste sentido, as modificações correntes nos dados, possuem vasta origem podendo ser oriundas de erros durante a etapa de coleta de dados, tanto por limitações físicas, ambientais ou ferramentais, quanto por impossibilidade de armazenamento adequado aos dados coletados. Há também a possibilidade de uma transferência ineficiente sobre estes dados, modificando-os erroneamente no processo. Independente da motivação inicial para sua existência, comumente há presença de ruído durante a análise dos dados supracitados.

A análise neste contida possui um novo foco, na busca por abordagens distintas para um problema recorrente, a decisão foi de construir uma ontologia que possua palavras tipicamente associadas a violência contra a mulher e utilizar desta ontologia para evidenciar os possíveis *tweets* violentos. Desta forma ser capaz de remover o ruído associado por palavras de baixo calão presentes no idioma, 'puta' e 'safada' são exemplos, palavras no gênero feminino que são consideradas ofensivas em um contexto mais amplo, entretanto mediante o ambiente informal de redes sociais, as mesmas palavras também podem ser usadas com um significado diferente, sendo mais uma representação de indignação sobre um assunto que uma ofensa dirigida a alguém, como em 'meu deus que puta dia ruim' e 'a alça dessa bolsa é safada, não serve para nada'. A utilização da ontologia é uma tentativa de remediar a perda de informações relevantes de contexto quanto a violência, e melhorar a precisão na descrição quanto a sua presença ou não em dado *tweet*.

3.4 Processamento de Linguagem Natural

Esta seção tem o objetivo de apresentar as técnicas de pré-processamento vigentes no domínio da interpretação automática de texto, ou seja a etapa inicial do processamento de texto, no qual se removem detalhes do texto que não são considerados úteis para os modelos de aprendizado de máquina descritos. A seguir, a ambição por uma predição correta, motiva este processo.

3.4.1 Técnicas de Pré-processamento

Existe um vasto conjunto de técnicas de pré-processamento para texto. As técnicas mais simplistas são as de aplicação direta, removendo características não interessantes quanto à pesquisa, como caracteres maiúsculos, a presença de *"@"* ou *"#"*, acentos e demais caracteres especiais, além dos conectivos linguísticos referentes ao idioma de estudo, neste caso o português brasileiro, exemplos de *stopwords* a se remover são 'de', 'a', 'o', 'que', 'e' .

Outra técnica é a tokenização, onde se converte uma frase em uma sequência de palavras aglutinadas, que passam a ser conhecidas como *"tokens"*. Comumente seguida pela normalização, técnica no qual se transforma o texto processado na forma padrão da língua, ao remover gírias e abreviações de palavras.

Outra abordagem possível é a da lematização, que possui três fases, a segmentação, no qual se divide o texto inteiro, nas respectivas frases que o compõem. O agrupamento de palavras com radical semelhante, como 'sorria' e 'sorriso'. Por fim, se segregam palavras polissêmicas, palavras em que seu significado varia com o contexto, como 'dama', palavra que é o feminino de cavaleiro, e 'dama', o jogo de tabuleiro.

3.4.2 Técnicas de classificação

No contexto atual de pesquisa em Inteligência Artificial (IA) existem numerosas maneiras de se produzir uma análise automatizada de dados, geralmente sendo agrupados em três grandes grupos. O primeiro é o de Aprendizado Supervisionado, onde a IA empregada necessita de um volume de dados catalogados corretamente, para criar seus padrões de reconhecimento e poder discernir uma função que computa uma entrada válida em uma saída correspondente válida. O segundo grupo é o de Aprendizado Semi-supervisionado, onde em seu amontoado de exemplos, parte de seus dados não são confiáveis, ou seja, possuem ruídos, e ainda existem dados não catalogados, de toda forma é necessário computar a função que converta uma entrada válida em saída válida. Por fim há o Aprendizado Por Reforço, no qual a IA produzirá uma versão da sua função de conversão de entradas em saídas e um agente externo, humano ou não, irá avaliar as respostas e reforçar as corretas, seja aumentando a recompensa dos acertos, ou aumentando a punição dos erros, após uma sucessão destes passos a função de conversão está apta para uma melhor predição sobre os dados.

A proposta para este trabalho é utilizar de um aprendizado de máquina supervisionado, de classificação booleana. Segue abaixo uma descrição dos modelos utilizados nesta categoria, para este trabalho.

O modelo Linear Regression é definido em [36], como seja 'x_j' um vetor com 'n' elementos, o espaço de hipóteses será o conjunto de funções na forma:

$$h_{sw}(x_j) = w_0 + w_1 x_j, 1 + ... + w_n x_{j,n} = wo + \sum_i wix_{j,i} \tag{3-1}$$

No qual cada termo 'w_i' representa um coeficiente de um valor real a ser descoberto e 'h_{sw}' é a função que comporta estes 'n' dados. Encontrar o melhor valor de 'h_{sw}' é realizar a Regressão Linear. A equação implementada para a obtenção do melhor conjunto

de pesos 'w_i' será:

$$w_i \leftarrow w_i + \alpha \sum_j x_{j,i}(y_j - h_w(x_j)) \tag{3-2}$$

Logo a cada iteração tem-se uma maior aproximação do termo h_{sw} com a distribuição real das classes do problema relatado.

Em paráfrase a [30] e [36] Multinomial Naïve Bayes é uma técnica que faz um ponderação ao priorizar velocidade de treino ao custo de uma performance final menos precisa. Como dito em [36]

> "O modelo é 'ingênuo' porque supõe que os atributos são condicionalmente independentes uns dos outros, dada a classe"

Sua aplicação pode ocorrer em quaisquer dados contínuos. O modelo considera a média de valores de cada recurso para cada uma das classes analisadas. Possui um único parâmetro que rege a complexidade de seu modelo, o alpha, no qual quanto menor seu valor, maior será a complexidade de seu modelo. Sendo uma adaptação do modelo Naïve tradicional, calcula a probabilidade de uma ocorrência 'A' considerando a ocorrência verdadeira de vários elementos 'B'. O cálculo da classe de seus resultados decorre da fórmula:

$$P(A|B_1, ..., B_n) = \alpha * P(A) \prod_i^N P(B_i|A) = \Theta^c * (1 - \Theta)^l \tag{3-3}$$

Em que P(B_i | A) representa a probabilidade de ocorrer B tal que já ocorreu A. Portanto, o cálculo realizado é o de maximizar o valor obtido por $P(A|B_1, ..., B_n)$ o resultado da combinação destas características de modo que, quanto maior o número de ocorrências de pares B_i e A, mais parecidos com alguma das classes o texto será. A classe que tiver a maior semelhança estatisticamente, será a classe predita.

O modelo Random Forest é uma variação do modelo de árvores de decisão booleana. Que por sua vez é definida em [36] como sendo:

> "[...] uma função que toma como entrada um vetor de valores de atributos e retorna uma 'decisão' - um valor de saída único. [...] em que cada exemplo é classificado como verdadeiro (**positivo**) ou falso (**negativo**)".

Por meio de uma sucessão de testes, o modelo decide quanto aos dados da entrada. Cada nó da árvore representa uma pergunta sobre algum valor de um dos atributos da entrada, A_i. Seus nós filhos são as respostas possíveis $A_i = v_ik$. Cada nó folha contém uma classe. Ao percorrer a árvore analisando os atributos da entrada, se decide a classe do texto analisado, como sendo a classe contida no nó folha atingido.

A indução de uma árvore de decisão é um par (A, B), no qual A é um vetor de características para os atributos da entrada e B é uma resposta booleana única. Porém essa abordagem é baseada em tabelas-verdade, onde para uma entrada com n atributos,

a tabela-verdade equivalente teria 2^n linhas, o que tornaria seu modelo pouco prático, entretanto por meio da busca gulosa sua indução se torna viável, ao sempre escolher o teste que tem maior impacto na decisão quanto a uma classe.

A escolha do teste com maior eficiência se baseia em encontrar o maior valor de ganho de informação, ou seja o teste que vai eliminar mais opções dentre as disponíveis. A fórmula para o ganho de informação é

$$Ganho(A) = B(\frac{p}{p+n}) - Resto(A) \tag{3-4}$$

sabe-se que Resto(A) equivale a

$$Resto(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} \cdot B(\frac{p_k}{p_k + n_k}) \tag{3-5}$$

e temos que

$$B(\Theta) = -(\Theta log_2 q + (1-\Theta) log_2 (1-\Theta)) \tag{3-6}$$

. Nas fórmulas '$B(\Theta)$' representa a entropia de uma variável aleatória booleana que é verdadeira com probabilidade Θ. 'P' representa o número de elementos positivos na amostra. 'N' representa o número de elementos negativos. 'A' representa um atributo, que possui 'D' valores distintos e divide o conjunto de treinamento 'E' em 'k' partes, como em $E_1, ..., E_k$. Cada conjunto 'E_k' possui seus próprios 'p_k' exemplos positivos e 'n_k' exemplos negativos.

Ainda assim, as árvores de decisão tendem ao *overfitting*, ou *super adaptação* em português brasileiro, quando a inteligência artificial se ajusta demais aos dados. Este fenômeno ocorre com maior frequência à medida que se expande o espaço de hipótese e o número de atributos da entrada.

Para combater este dilema surge a ideia por trás das *Random Forest* que segundo os autores de [30] são:

> "[...] essencialmente uma coleção de árvores de decisão, onde cada árvore é ligeiramente diferente das demais [...]"

Como maneira de se combater a tendência das árvores de decisão de super adaptação. Múltiplas árvores de decisão são produzidas e cada uma delas, classifica o texto, com suas próprias escolhas quanto a testes a sua própria maneira de tentar maximizar o ganho de informação de 3-4. Uma árvore de decisão construída, recebe os dados, executa sua decisão caminhando pela árvore e retorna uma resposta, que é o voto para alguma das classes, o modelo *Random Forest* então cria a média das respostas dadas por cada árvore de decisão, e a usa como resposta para a predição do texto recebido.

Support Vector Machines Segundo [36] uma *Support Vector Machine* (SVM), é uma inteligência artificial que constrói um separador de margem máxima, que realiza decisões por meio de um sistema de separação linear em hiperplanos. Os SVMs minimizam a perda da generalização esperada.

Ao escolher o separador de classes como sendo o mais distante entre os exemplos existentes. Produz-se um separador de margem máxima. Um separador simples se define por:

$$x : w \cdot x + b = 0 \tag{3-7}$$

Na fórmula 'w' representa o peso encontrado, 'x' representa o valor variante da entrada da reta, 'b' é o parâmetro de corte da reta. Uma forma para encontrar o espaço de 'w' e de 'b' é realizar a descida do gradiente maximizando a margem e mantendo a corretude da classificação dos valores. Os detalhes serão omitidos mas os parâmetros que maximizam a margem, mantendo as classificações corretas podem ser obtidos ao resolver:

$$argmax \sum_j \alpha_j - \frac{1}{2} \sum_j \alpha_j \alpha_k y_j y_k (x_j \cdot x_k) \tag{3-8}$$

Para tal, existem as restrições

$$\alpha \geq 0 \tag{3-9}$$

e

$$w = \sum_j \alpha_j y_j \tag{3-10}$$

Ao lidar com espaços não linearmente separáveis, as SVMs precisam de outras alternativas, logo se criou o 'Truque de Kernel' pelo qual se aplicam a pares de dados de entrada a avaliação quanto a fórmulas de kernel, capazes de tornar valores lineares em valores em espaços característicos de dimensões superiores, de forma a conseguir separar os dados de maneira mais eficiente. Ao mapear estes espaços lineares ao espaço de entrada original, podem ser geradas fronteiras sinuosas que comportem melhor os dados textuais em suas respectivas classes.

CAPÍTULO 4

Experimento

4.1 Desenho do Experimento

A criação da ontologia que se segue foi produzida por meio do método definido por Chaves em [18], sendo este uma variação das ações presentes nas metodologias *On-To-Knowledge*, *Menthology* e *Ontology Development 101*. A ferramenta escolhida foi o Protégé 2000, visto a sua alta disponibilidade e possibilidade de exportação em múltiplos formatos. A construção da ontologia se fez por meio da interface do software Protégé e a linguagem escolhida para exportação foi a sintaxe OWL/XML. Após sua criação o recorre-se a linguagem Python e a biblioteca Pandas para processar os *tweets* coletados.

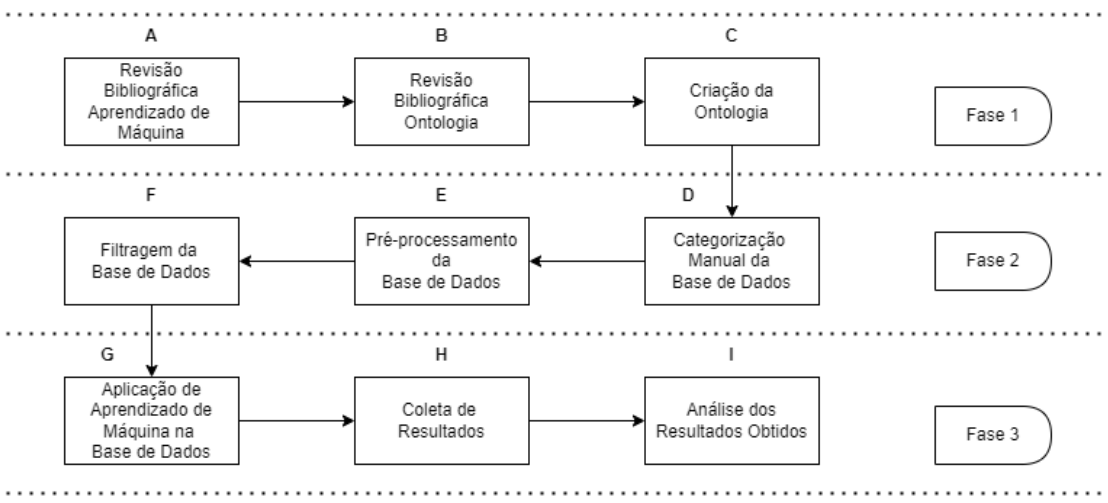


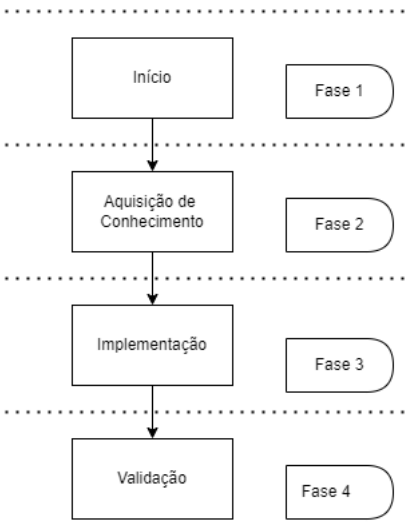
Figura 4.1: Ilustração das etapas do experimento.

Para uma melhor visualização do experimento como um todo, a Figura 4.1 ilustra as etapas decorridas. Há um particionamento das atividades em fases, a Fase 1 inicia com a realização da revisão bibliográfica quanto à ontologia (B) e ao aprendizado de

máquina (A), etapas contidas nos Capítulos 2 e 3 deste. Para a elaboração da ontologia (C) a Subseção 4.1.1 elucidará os processos utilizados.
Segue de maneira similar para as atividades referentes a Fase 2, a categorização manual (D) será explicada na subseção 4.1.2, e o pré-processamento (E) será descrito na subseção 4.1.3, a filtragem dos dados (F) está na subseção 4.1.4. Analogamente a aplicação do aprendizado de máquina nos dados (G) é apresentada na Subseção 4.1.5. Por fim, a obtenção de resultados (H) e sua análise (I) são expostas no Capítulo 5.

4.1.1 A criação da Ontologia

Para a geração da ontologia fez-se uso do processo descrito por Chaves em [18], a versão criada por Chaves é uma amálgama de três grandes Métodos, *Menthology*, *On-to-Knnowledge* e *Ontology Devolpment* 101. Sua execução é particionada em etapas: Início, Aquisição de conhecimento, Implementação e Validação, o que pode ser contemplada pela figura 4.2.



Elaborado pelo autor

Figura 4.2: Esquema de etapas para a produção de ontologias segundo a metodologia de Chaves

Durante a fase inicial (1), deve-se elaborar um cronograma com quais atividades serão realizadas e quanto tempo cada fase levará, em seguida deve decidir o propósito a qual a ontologia gerada seguirá, para esclarecer a razão de querer criar esta ontologia. O passo seguinte será discernir o domínio, que descreve a área de atuação da ontologia e o escopo, que descreve o quão abrangente a ontologia deve ser. Por fim, há a busca por ontologias existentes, para possível adaptação e reuso.

Como resposta aos questionamentos produzidos pela Fase 1 da metodologia produzida por Chaves. Temos a definição do propósito da ontologia que é de reconhecer

em uma base dados, *tweets* que possuam traços de violência contra a mulher, desta forma sendo auxílio para a remoção de ruídos nos dados apresentados. Seu domínio é o de violência contra a mulher e seu escopo são as palavras contidas pelo idioma português brasileiro.

Na fase de aquisição de conhecimento (2), desenvolve-se questões de competência, sendo estas um guia sobre como a ontologia ideal deveria se comportar. O passo que se segue é a enumeração de termos importantes, no qual os termos mais frequentes daquele domínio, são os termos desejados elencar, nesta etapa requer atenção quanto a criação de contradições e ambiguidades entre os termos. Em seguida, se selecionam termos reutilizáveis, sejam estes oriundos do vocabulário popular ou de outras ontologias. Sucede a classificação dos termos, definindo-os quanto a ser um atributo, uma classe, uma instância ou uma propriedade. Em seguida, é definida a hierarquia de classes, ao associar duas ou mais classes. Elucidar possíveis generalizações de termos, é a etapa responsável por definir se a criação da ontologia será *top-down*, de termos mais genéricos para os termos mais específicos ou *bottom-up*, de termos mais específicos para os termos mais genéricos. Posteriormente se tem a criação de propriedades que as classes possui. Logo há o mesmo processo em relação a quais são atributos e a que classe estes pertencem. A última etapa desta fase é crucial, onde por meio de representação visual se cria um modelo conceitual para agrupar as proposições anteriores, neste se usou da UML.

Como resposta aos questionamentos da Fase 2 da metodologia se gerou as seguintes Questões de Competência.

- O texto é ofensivo?
- Qual a origem da ofensa proferida?
- O texto possui alguma variação de escrita?

Após análise sobre as palavras com maior destaque quando o assunto é violência, se encontrou como considerável reuso pela mídia as palavras: Ofensa, agressão, violência. A partir deste momento, se decidiu avaliar cada palavra ofensiva encontrada para depois agrupá-las com as classes que melhor comportem as suas características, o que tornou a hierarquia de classe empregada como sendo a *bottom-up*.

Ao iniciar o desenvolvimento de classes ficou decidido por usar uma grande classe para todos os tipos de palavras, a superclasse Termo. Como subclasse de Termo existe a classe Ofensa, que armazena ofensas misóginas independente do tipo de ofensa empregada. Como subclasses de Ofensa temos: Capacidade, Etnicidade, Fisicalidade, Moralidade, Nacionalidade, Sanidade, Sexualidade, como tentativa de agrupar adequadamente todo tipo de agressão contra a mulher encontrada, será atribuída a alguma destas subclasses.

O próximo passo foi o de descrever relações que se adequem a estas classes. Para tal se decidiu criar a relação 'é uma', que associa um Termo a uma Ofensa. Em seguida

se optou por criar a relação 'quanto a', responsável por ligar uma Ofensa a uma das suas subclasses.

A próxima atividade desta fase foi discernir regras para os dados, e para permitir uma avaliação mais simples, apenas uma regra foi estabelecida, ficou decidido então que cada membro instanciado de uma das subclasses, será sempre disjunto das demais, ou seja uma ofensa pertence unicamente a uma das subclasses existentes, desta forma a análise sobre qual classe cada palavra ocupa é mais direta.

A etapa que se seguiu foi a de produção de atributos para as classes, e a ideia inicial foi conceder à classe Termo o atributo numérico 'Quantidade de palavras' responsável por contar o número de palavras em suas subclasses. Para a classe Ofensa foi atribuído o atributo do tipo lista, 'Variações de escrita', responsável por enumerar as possíveis grafias das ofensas. Para exprimir todo o progresso da fase 2, se produziu a tabela 4.1 que é uma adaptação do modelo conceitual proposto por Chaves.

A terceira fase é a da implementação (3), no qual se converte o modelo conceitual da fase anterior em uma ontologia computável, para tal é necessário a escolha de uma linguagem de criação de ontologias e de uma ferramenta para tal.

Durante a fase 3 da metodologia, o processo foi apenas de transportar as propostas realizadas pelas demais fases para a interface da ferramenta de criação de ontologias Protegé 2000 e o resultado do processo, ou seja a ontologia gerada pode ser vista na Figura 4.3.

A fase final é a validação (4), momento para comparar o conteúdo presente na ontologia e nas fontes de conhecimento, para garantir conceitos bem definidos e não-dúbios. Segue a validação perante usuários, onde o autor testa a ontologia com seus possíveis usuários em busca de notar inconsistências e possíveis melhorias. Ao final se avalia a eficiência da ontologia em responder às questões de competência definidas na fase de aquisição de conhecimento.

A fase 4 da metodologia foi avaliar os resultados obtidos quanto às questões de competência, onde cada questão levantada foi relizada para cada termo a ser adicionado na ontologia, a análise das respostas obtidas, auxiliou na construção de cada classe, portanto a ontologia atende sim, as questões de competências estabelecidas. Por cumprir o propósito do estudo sobre a violência e ter auxiliado na identificação de *tweets* que contenham traços de violência contra a mulher, a ontologia produzida foi considerado um sucesso.

Classe	Subclasse	Relação	Regra
Termo	Ofensa	é uma	-
Ofensa	Capacidade, Etnicidade, Fisicalidade, Moralidade, Nacionalidade, Sanidade, Sexualidade	quanto a	Disjunção mútua entre todas as subclasses

Tabela 4.1: Tabela de relação entre classes da ontologia

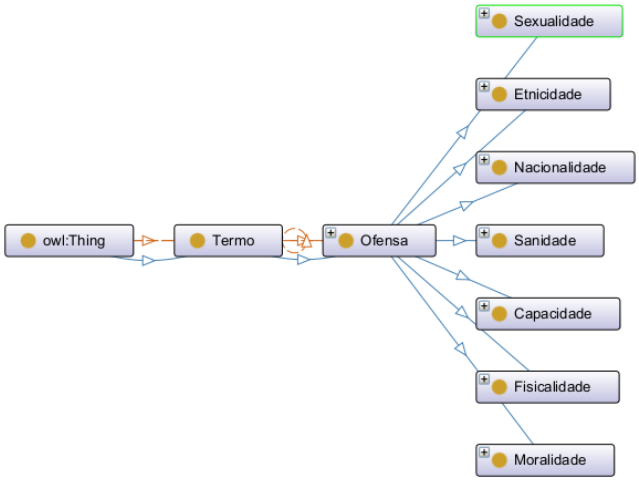


Figura 4.3: Ilustração da ontologia gerada.

4.1.2 Classificação manual dos dados

Para o treinamento de uma inteligência artificial de forma supervisionada emerge a necessidade de uma coleção de dados de treinamento e dados de teste que permitam relacionar uma questão à sua solução. A base de treinamento para este trabalho foi gerada por meio da parceria com um outro aluno do grupo de pesquisa guiado pela professora orientadora, o discente Heinrych M. G. Andrade. Para tal, foram selecionados a quantidade de 1 mil *tweets* os quais foram classificados em duas categorias: Não contém traços de violência contra a mulher e Contém traços violência contra a mulher, levando em consideração os seguintes critérios.

Exemplos de texto marcado como 'sem traços de violência':

- *"que nunca tenha um baldasso no gremio"*
- *"ba vim so com um moletom e ta tri frio, puta merda"*
- *"poxa jao ja era!"*

Exemplo de texto marcado como 'com traços de violência':

- *"@pessoa*[1] *posta tt, mais nao me responde vagabunda"*

[1] @pessoa é a maneira de preservar a privacidade da citação entre usuários.

- *"@pessoa[1] eque gente burra."*
- *"rt @pessoa[1]: fico puta quando uma garota feia e mais bonita que eu"*

Os critérios de seleção foram a presença de ofensas contra mulheres, os pontos decisivos de escolha são a presença de linguagem chula e a direção da violência. Quanto a existência de xingamentos no corpo do texto não são um indicativo direto de violência como em:

- *"eu nao tenho um dia de paz, puta que pariu!!"*
- *"psg montando um puta time viu.."*

No texto acima há presença de termos chulos para a referência ao gênero feminino mas que não se direcionam a nenhuma pessoa em específico, sendo um dos desafios do Processamento de Linguagem Natural compreender essa variação linguística oriunda de padrões sutis da comunicação informal, termos como este foram um dos desafios para avaliação de classificadores humanos, se mantendo desafiador para classificadores automáticos.
De forma similar, frases prejudiciais ditas pelo autor do tweet para referir a si mesmo, auto-depreciação, foram consideradas como baixa-estima e não violência contra todas as mulheres. Como em:

- *"mas eu sou burra viu"*
- *"ta para nascer mais otaria que eu, puta que pariu, eu me supero a cada dia uo"*

4.1.3 Pré-processamento

Para um aprendizado de máquina eficiente, temos de saber como dito pela autora de [10] "Se seus dados são ruins, suas ferramentas de *Machine Learning* são inúteis". Em outras palavras, a qualidade do aprendizado possível para sua Inteligência Artificial é diretamente relacionada e limitada à qualidade dos dados de entrada fornecidos.

Para minimizar os ruídos dos dados brutos, as técnicas de pré-processamento selecionadas para este trabalho foram: remoção de *hiperlinks* e caracteres especiais, seguida pela transformação do texto em minúsculas, a tokenização e lematização do texto por intermédio da biblioteca python **nltk** [2], por fim houve a remoção das *stopwords* intrínsecas ao Português Brasileiro. As nuvens de palavras deste processo são apresentadas no Capítulo 5 nas Figuras 5.1 e 5.2.

[2] **Natural Language Toolkit** plataforma especialista em linguagem python para lidar com Processamento de Linguagem Natural

4.1.4 Filtragem nos dados

Durante a etapa de filtragem de dados, realizou-se o agrupamento de *tweets* que continham dados com possíveis traços de violência contra as mulheres. Para esta separação, fez-se o uso da ontologia, descrita na Seção 4.1.1, criada para tal. A sua adaptação para a linguagem Python pode ser vista na Figura 4.4. As nuvens de palavras deste processo são apresentadas no Capítulo 5 nas Figuras 5.1 e 5.2.

```
termos_ofensa = ['capacidade', 'etnicidade', 'fisicalidade', 'moralidade', 'nacionalidade', 'sanidade', 'sexualidade']
termos_capacidade = ['anta', 'burra', 'cadela fracassada', 'desprezivel', 'fracassada', 'idiota',
                     'imbecil', 'mongoloide', 'mulher inutil', 'otaria']
termos_etnicidade = ['criola', 'macaca', 'mulata', 'nariz de nega', 'nega', 'negrice', 'preta']
termos_fisicalidade = ['ana', 'baleia', 'baranga', 'feia', 'gorda', 'gorila', 'horrorosa', 'mocreia', 'nojenta']
termos_moralidade = ['desviada', 'farsante', 'filha da mae', 'filha da puta', 'filha de chocadeira', 'golpista', 'mal amada',
                     'meretriz', 'mulher facil', 'perua', 'piranha', 'pistoleira', 'puta', 'rameira', 'safada', 'vadia', 'vagabunda']
termos_nacionalidade = ['baianagem', 'carcamana', 'resto de porra', 'sangue ruim']
termos_sanidade = ['histerica', 'louca', 'maluca', 'doida']
termos_sexualidade = ['bambi', 'viada']
```

Figura 4.4: Palavras contidas na ontologia.

Ao realizar rotinas de seleção na base de dados, exemplificadas nas imagens 4.5 e 4.6, buscando pelos termos contidos na ontologia (Figura 4.4), dos 1032852 itens houve uma remoção de 945804 desses, o que representava 91.57% da base de dados original. Os elementos selecionados compõem então, um arquivo para consulta durante a pesquisa. As nuvens de palavras deste processo são apresentadas no Capítulo 5 nas Figuras 5.3 e 5.4.

```
def checa_se_contem_termos(index, tweet, termos_classe_principal, termos_classe_relacionada):
    resultado = 1
    for termo_principal in termos_classe_principal:

        for termo_relacionado in termos_classe_relacionada:

            if (termo_principal in tweet and termo_relacionado in tweet) or (termo_relacionado in tweet):
              lista_tweets_violencia_contra_mulher.append(index)
              print(index, termo_principal, termo_relacionado)
              with open("caminho/lista_tweets_violencia_contra_mulher.txt", "w") as fp:
                json.dump(lista_tweets_violencia_contra_mulher, fp)

              with open("lista_tweets_violencia_contra_mulher.txt", "w") as fp:
                json.dump(lista_tweets_violencia_contra_mulher, fp)

              resultado = 0
              return resultado

    return resultado
```

Figura 4.5: Rotina de busca de um termo específico da ontologia.

```
for index, tweet in df.iterrows():
    tweet = tweet.text
    flag = 1
    if flag == 1:
      checa_se_contem_termos(index, tweet, termos_ofensa, termos_capacidade)
      if flag == 1:
        checa_se_contem_termos(index, tweet, termos_ofensa, termos_etnicidade)
        if flag == 1:
          checa_se_contem_termos(index, tweet, termos_ofensa, termos_fisicalidade)
          if flag == 1:
            checa_se_contem_termos(index, tweet, termos_ofensa, termos_moralidade)
            if flag == 1:
                checa_se_contem_termos(index, tweet, termos_ofensa, termos_nacionalidade)
                if flag == 1:
                    checa_se_contem_termos(index, tweet, termos_ofensa, termos_sanidade)
                    if flag == 1:
                        checa_se_contem_termos(index, tweet, termos_ofensa, termos_sexualidade)

with open("lista_tweets_violencia_contra_mulher_pos_filtragem.txt", "w") as fp:
    json.dump(lista_tweets_violencia_contra_mulher, fp)
```

Figura 4.6: Rotina que realiza a busca de todos os termos da ontologia.

Ao observar os *tweets* listados abaixo, nota-se a presença de palavras que contém sub-palavras contidas na ontologia, observável em 4.4.

- *"so da maneira de distribuir dinheiro p deputado e empregar sobrinho ne"*
- *"efeito bolsonaro bate turbulencia americana e ibovespa fecha em alta https:link* [3]

Para contornar esta situação apresentada, na qual há textos contendo parte das palavras selecionadas na ontologia, optou-se por realizar uma nova filtragem com a remoção desses elementos indicados. Para isso, implementou-se rotinas a partir da biblioteca Python de expressões regulares 're', para identificar tais palavras e removê-las da lista, o código equivalente se encontra em 4.7 e 4.8, sua aplicação diminuiu uma vez mais esta proporção. Na base, houve a remoção de 2936 *tweets* o que representava aproximadamente 3.37%[4] desta nova listagem, sendo portanto a base de dados final para a pesquisa.

```
pattern = '\Wputa\W'
contador = 0
for indice in conjunto_exemplo:
  resposta = re.search(pattern, df['text'][indice])
  if resposta is not None:
    conjunto_final[contador] = -1
  else:
    print('.')
  contador += 1
```

Figura 4.7: Rotina de remoção para subpalavras contendo 'puta'.

```
pattern = '\Wana\W'
contador = 0
for indice in conjunto_exemplo:
  if conjunto_final[contador] != -1:
    resposta = re.search(pattern, df['text'][indice])
    if resposta is not None:
      conjunto_final[contador] = -1
      print(df['text'][indice])
    else:
      print('.')
  contador += 1
```

Figura 4.8: Rotina de remoção para subpalavras contendo 'ana'.

O resultado da aplicação destas duas etapas de filtragem é ilustrada na Figura 4.9, onde se nota a queda no volume de dados a se analisar, assentindo com a hipótese de que a aplicação de ontologia pode reduzir o ruído de grandes volumes de dados.

[3] **https:link** é a maneira abordada aqui para representar a presença de um *hyperlink* no texto, enquanto se mantêm a privacidade dos autores do *tweet*

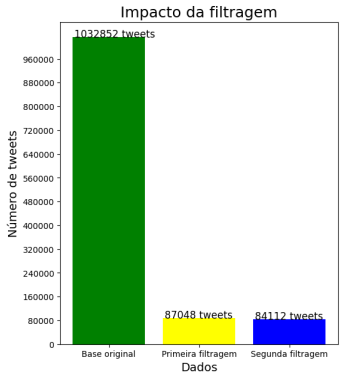


Figura 4.9: Impacto das rotinas de filtragem nos dados.

4.1.5 Aplicação de Aprendizado de Máquina

Baseados na recomendação de [36] e nos trabalhos de [37] e [20] as técnicas de Inteligência Artificial aplicadas neste foram: *Logistic Regression*, *Multinomial Naïve Bayes*, *Random Forest* e *Support Vector Machines*, vide 3.4. De maneira a possibilitar seu uso, converteu-se os dados textuais (*tweets*) para vetores esparços presentes na biblioteca **numpy** [4] por meio do método de contagem de palavras *CountVectorizer*, mais conhecido como *bag-of-words*, implementado na bilioteca *scikit-learn* [34] [5].

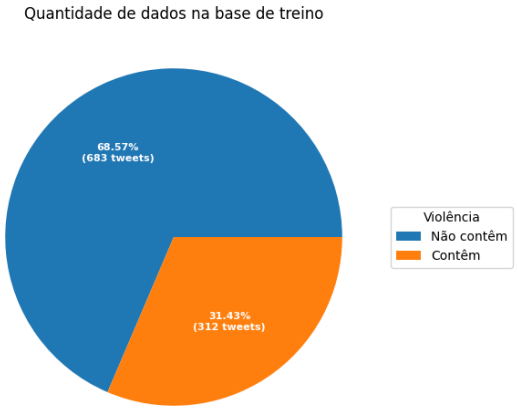


Figura 4.10: Disposição dos dados de treino

Para o treinamento foram 996 *tweets* catalogados manualmente, como descrito em 4.1.2. A distribuição dos dados foi de 68.57% contendo traços de violência contra a mulher e de 31.43% não a contendo, como pode ser observado em 4.10. O treino

[4] **Numpy:** Projeto *open source* que auxilia na computação numérica em linguagem Python. Disponível em: https://numpy.org/

[5] **Sklearn:** Plataforma de ensino de inteligência artificial. Disponível em: https://scikit-learn.org/stable/index.html

dos modelos ocorreu com 70% da base, balanceados artificialmente perante a cláusula *stratify* do módulo *train_test_split* contidos em *sklearn.model_selection*, e tendo sua etapa de treino e teste como é descrito por 4.11, 4.12 e 4.13 em busca pela assertividade. As configurações adotadas para cada modelo testado foram as padrões, exceto pela presença do parâmetro *random_state* a qual atribui-se valor arbitrário 10, para permitir replicabilidade.

```
xa = data['texto_processado']
ya = data['target_final']
```

```
xa_novo = []
preprocessar(data, 'texto_processado', xa_novo)
```

```
x_traina, x_testa, y_traina, y_testa = train_test_split(xa_novo, ya, test_size=0.3, random_state=10, stratify=y)

cv = CountVectorizer(max_features=576)

x_train_em_float = cv.fit_transform(x_traina)
```

Figura 4.11: Divisão e treino com os dados catalogados.

```
cvc = CountVectorizer(max_features=576)
x_train_em_float_576 = cvc.fit_transform(x_traina)
```

```
w = np.asarray(original['target_final'])
```

Figura 4.12: Processo de vetorização das palavras.

```
def plot_cm_treinada(modelo):
  modelo.fit(x_train_em_float_576, y_traina)
  z_test_em_float = cvc.transform(original['texto_processado'])
  z_predita = modelo.predict(z_test_em_float)
  cm = confusion_matrix(w, z_predita)
  plot_confusion_matrix(cm)

  print('O modelo possui: ')
  print(f'Acurácia:\t{modelo.score(z_test_em_float, w)}')
  print(f'F1-Score:\t{f1_score(w, z_predita, average="macro")}')
  print(f'Recall:  \t{recall_score(w, z_predita, average="macro", zero_division=0)}')
  print(f'Precision:\t{precision_score(w, z_predita, average="macro", zero_division=0)}')
```

Figura 4.13: Rotina de impressão de métricas do modelo.

As métricas de validação dos dados a serem adotadas são: acurácia, *F1-Score*, *Recall* e *Precision*.Se um texto for classificado como possuidor de traços de violência contra a mulher e ele realmente o possui, será categorizado como um Verdadeiro Positivo (VP). Se for identificado como possuidor de misoginia, e não a contiver, será um Falso Positivo (FP). Se o texto for minuciado como não possuidor de violência, e realmente não apresentar, será Verdadeiro Negativo (VN), por fim caso o texto seja caracterizado como não apontar misoginia, mas a relatar, será um Falso Negativo (FN).

As métricas de validação de aprendizado analisam estes valores, cada um à sua própria maneira. A **'Acurácia'** correlaciona os termos por meio da fórmula:

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \tag{4-1}$$

A acurácia é utilizada como correlação entre os valores positivos encontrados e todos os valores encontrados, se foca no quanto o modelo acertou no caso geral.

Para calcular o valor de ***'Recall'*** é necessário aplicar:

$$Recall = \frac{VP}{VP + FP} \tag{4-2}$$

O *Recall* é utilizado como medida para minimizar a valorização de falsos negativos, em contextos de detecção de doenças é mais importante acertar quem está doente (VP), do que considerar alguém saudável como doente (FP).

Quando se deseja calcular o valor de ***'Precision'*** é necessário aplicar:

$$Precision = \frac{VP}{VP + FN} \tag{4-3}$$

O *Precision* utilizado como medida para minimizar a valorização de falsos positivos, em contextos de detecção pesquisas meteorológicas, é mais importante acertar quando realmente irá ocorrer um furacão (VP), do que os casos com fortes indícios mas não ocorreu (FN).

As métricas *Recall* e *Precision* são ligeiramente opostas, não sendo possível atingir altos valores em ambas ao mesmo tempo.

O valor da métrica ***'F1-Score'*** utiliza da abordagem:

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall} \tag{4-4}$$

Mediante uso das métricas citadas acima, é uma boa alternativa para um panorama geral de acertos do modelo, ao ser uma visão geral sobre as reações do modelo quanto a seus erros, se utiliza esta métrica quando se deseja que seus erros de predição sejam mínimos tanto positivamente (FP) quanto negativamente (FN).

CAPÍTULO 5

Resultados

No capítulo 4 se apresentou o método proposto em busca do objetivo desta pesquisa, a aplicação de ontologia para a remoção de ruído, ao explicitar as etapas para sua criação e utilização juntamente do aprendizado de máquina. Neste capítulo se exibirá os resultados obtidos ao realizar tais atividades. Sendo primeiramente mostradas nuvens de palavras dos dados sem pré-processamento e com pré-processamento dos dados existentes, para a base de dados original, para a base com a primeira filtragem e para a base com a segunda filtragem e se discutirá o efeito do pré-processamento em cada etapa. Em seguida, se ilustrará os resultados obtidos com o aprendizado de máquina para a base original e para a base final após todos os filtros, com argumentação das motivações para estes resultados.

5.1 Visualização e pré-processamento de dados

Nesta seção serão apresentandos os resultados obtidos com a realização das etapas E e F de pré-processamento e filtragem da base de dados, ambas na Fase 2 mostrada na Figura 4.1 do Capítulo 4.

As Figuras 5.1 e 5.2 apresentam as nuvens de palavras para a base de *tweets* original contendo os 1032852 *tweets*, sem nenhum tratamento e após a realização das atividades de pré-processamento explicitadas em 4.1.3, respectivamente.

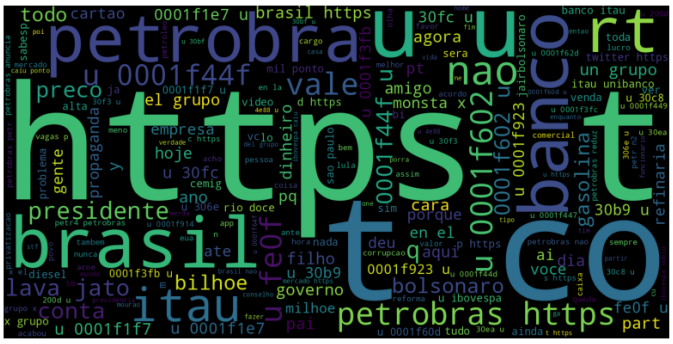


Figura 5.1: Nuvem de palavras da base de *tweets* original.

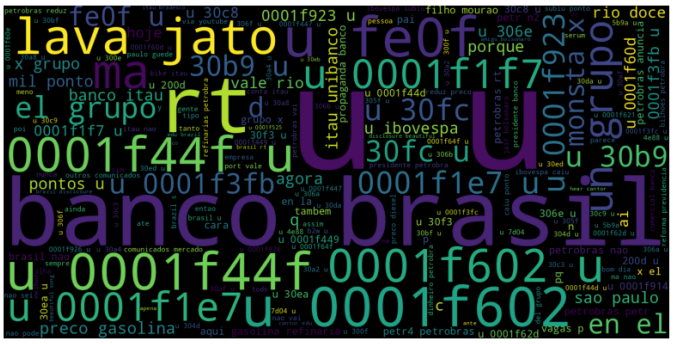


Figura 5.2: Nuvem de palavras da base de *tweets* original pré-processada.

As palavras que possuem maior destaque na Figura 5.1 são 'https', 'petrobras', 'brasil' e 'banco', além dos caracters 't' e 'co', possivelmente pela discussão acalorada entre eleitores de direita e esquerda sobre o futuro da empresa petrolífera brasileira, durante o ano da coleta de dados desta base, além do compartilhamento de *links* externos a rede social, no caso do 'https'.

Enquanto se ressaltam na Figura 5.2 se mantem em destaque 'banco' e 'brasil' mas as demais palavras mudam, com a presença de 'lavajato', 'rt'. A aparição de 'rt' tem o signo de reencaminhamento de algum *tweet*, sendo este o *retweet* ou na sigla 'rt'. Ocorrem algumas sequências de números que em primeiro olhar não parecem relevantes, entretanto estas são representações hexadecimais em UTF-32 para *emoji*. Como '0001f602' que representa o emoji de 'sorrindo com lágrimas', a entrada '0001f44f' que representa o *emoji* de 'batendo palmas' e o valor '0001f1f7' que representa um *emoji* incomum um 'símbolo de indicador regional, a letra R'. A presença marcante destes valores hexadecimais são um bom indício de uma grande presença de *emojis* perante a base de dados original.

As Figuras 5.3 e 5.4 apresentam as nuvens de palavras obtidas para a base de dados após a primeira filtragem. Isto é, a base gerada ao selecionar dentre os 1032852 *tweets* da base original, os 87048 *tweets* que continham traços de violência contra a mulher, que foram descobertos por meio da aplicação da ontologia na remoção de ruído. Para maiores detalhes vide Subseção 4.1.4 no Capítulo 4.

Figura 5.3: Nuvem de palavras da base de dados após primeira filtragem com ontologia

Figura 5.4: Nuvem de palavras da base de dados após primeira filtragem com ontologia pré-processada

As palavras em maior realce para 5.3 são as mesmas de 5.1 acrescidas de ’não’, ’semana’ e aparece o primeiro xingamento ’puta’, indicativo de que a filtragem colaborou com a hipótese de que ontologia, auxilia na remoção de ruído para a detecção de traços de violência.

Enquanto que as palavras de maior distinção para 5.4 são ’rt’, ’banco’, ’brasil’, ’vale3’, ’petr4’, ’nao’, ’analise’, ’ibov’. A presença de termos como ’banco’ e ’brasil’ se explicam pela aparente indignação dos usuários com intituições bancárias. A aparição de ’vale3’, ’petr4’ suplantam a ideia de conversas sobre o mercado financeiro, sendo estas nomes de ações das empresas Vale S. A. e Petrobras. A presença de ’ibov’ também assente quanto a isto, visto que é a abreviatura do indíce Ibovespa.

As Figuras 5.5 e 5.6 apresentam as nuvens de palavras obtidas para a base de dados para a segunda filtragem. Isto é, a base gerada ao selecionar dentre os 87048 *tweets* da base de primeira filtragem, os 84112 *tweets* que não contenham em sua descrição subpalavras da ontologia. Para maiores detalhes vide Subseção 4.1.4 no Capítulo 4.

Figura 5.5: Nuvem de palavras da base de dados após segunda filtragem com ontologia

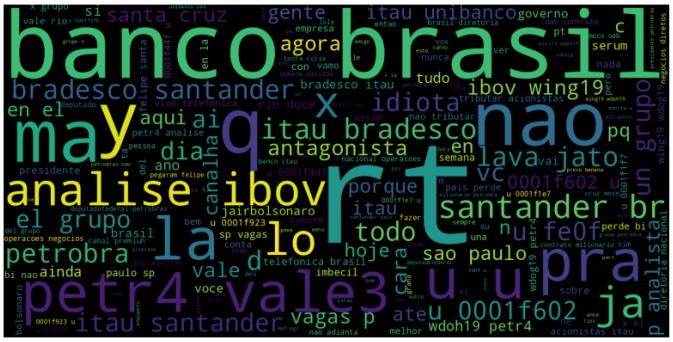


Figura 5.6: Nuvem de palavras da base de dados após segunda filtragem com ontologia pré-processada

Para Figura 5.5 as palavras que se sobressaem são 'https', 'petrobras', 'vale', 'banco', 'brasil', 'não' e 'semana'. A motivação para a presença das cinco primeiras já foram discutidas neste Capítulo. Todavia a presença de 'nao' e 'semana' não conferem muita informação sobre o contexto para sua produção.

Por sua vez, na Figura 5.6 as palavras que se acentua são 'rt', 'banco', 'brasil', 'nao', petr4', 'vale3, 'analise', 'pra'. De forma similar, as únicas palavras não discutidas são 'analise' que remete possivelmente para analise financeira e a versão informal da preposição "para".

O destaque de palavras do sistema financeiro é decorrente da alta violência no discurso financeiro, a inquietude quanto aos resultados ruins de algum investimento, resultam em uma série de xingamentos em seus relatos de *tweets*. Entretanto, sua maioria não contem cunho violento contra mulheres, mas pela presença notável de palavras de baixo calão femininas, acabam por parecer em nossa base de dados.

Os termos com maior ocorrência para todas as nuvens de palavras foram portanto 'banco', 'brasil', 'petroras', 'petr4' e 'vale3', o que induz a pensar na tendência da base obtida, em comentar sobre temas econômicos brasileiros. Ao fim, é perceptível que a etapa de filtragem observada na Subseção 4.1.4 e a de pré-processamento na Subseção

4.1.3, ambas Subseções do Capítulo 4 deste, se mostraram capazes de reduzir o ruído quanto a notável presença de *emojis* na base de dados original.

5.2 Classificação dos dados da base original

Nesta seção há a apresentação dos resultados obtidos ao classificar os dados da base original. Logo, esta uma seção apresenta a execução das etapas G, H e I da fase 3 descrita pela Figura 4.1 no Capítulo 4.

Os dados expostos aqui foram produzidos por meio de 'plot_cm_treinada' ilustrada em 4.13 apontando as métricas e matrizes de confusão resultantes, para cada um dos quatro modelos de inteligência artificial utilizados, sendo estes *Logistic Regression*, *Multinomial Naïve Bayes*, *Random Forest* e *Support Vector Machine*, juntamente da discussão para os resultados obtidos. Os algoritmos selecionados são categorizados como Aprendizagem Supervisionada, exigindo uma base de dados previamente rotulada para poderem realizar seu treino. O processo de rotulação foi descrito na Subseção 4.1.2 no Capítulo 4.

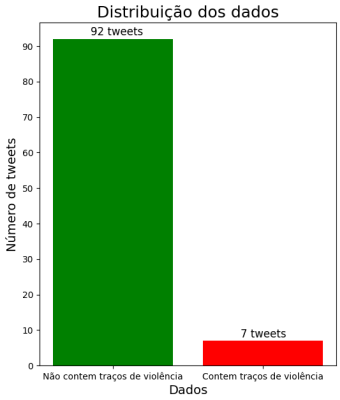


Figura 5.7: Distribuição dos dados de teste para a base original.

Subsequente ao processo de treino dos modelos de aprendizado de máquina descritos em 4.1.5 ocorreu a testagem do aprendizado para a base de dados original, para tal testagem foram selecionados aleatoriamente 100 *tweets* contidos na base original. Analisou-se então, a possível presença de traços de violência nestes, manualmente catalogou-se quanto a isto, a distribuição dos dados resultante pode ser observada na Figura 5.7. O processo seguinte foi o de utilizar o modelo treinado com a base catalogada dos 1 mil *tweets* e testar se realmente houve algum aprendizado. Ao inserir estes 100 novos dados e avaliar o que o modelo preveu, e comparar com os valores reais anotados quanto a presença ou não de traços de violência contra a mulher, se pode avaliar a efetividade do modelo. Os resultados deste processo se encontram disponíveis na Tabela 5.1.

Nesta, podem ser visualizadas as métricas alcançadas por cada um dos algoritmos, sendo elas Acurácia, *F1-score*, *Recall* e *Precision*.

Modelo de Dados	**Acurácia**	**F1-Score**	**Recall**	**Precision**
Logistic Regression	0.92929	0.48167	0.5	0.46464
Multinomial Naïve Bayes	0.85858	0.46195	0.46195	0.46195
Random Forest	0.91919	0.47894	0.49456	0.46428
Support Vector Machine	0.92929	0.48167	0.5	0.46464

Tabela 5.1: Métricas obtidas para os modelos de aprendizado de máquina adotados.

Por coincidência, dois modelos alcançaram os mesmos valores em todas as métricas, sendo os melhores durante a classificação, dos dados contidos nesta pesquisa, são eles *Logistic Regression* e *Support Vector Machine*, com uma acurácia de 92.92%, e *F1-Score* de 48.16%, o modelo *Random Forest* por sua vez obteve 91.91% de acurácia e um *F1-Score* de 47.89%, por fim o modelo com o pior desempenho foi o *Multinomial Naïve Bayes* com 85.85% de acurácia e 46.19% de *F1-Score*. A motivação suposta para uma acurácia ao redor dos 90% em todos os modelos, porém com um valor abaixo de 50% quanto a *F1-Score* foi a tendência dos modelos em predizer todo texto como não contendo traços de violência. Pois, ao sempre alegar que não existia violência, e na maior parte da base não existir realmente (Verdadeiros Negativos), se criou uma inflação nos valores de acurácia. Entretanto dado o contexto vigente, os melhores resultados seriam os que acertassem corretamente a classificação dos *tweets* que realmente continham traços de violência contra a mulher. Dito isto, o melhor dos modelos é o que possuir o valor de *Recall* mais alto, pois uma das características que se destacam nesta métrica é a preferência por aceitar Falsos Positivos, em detrimento de permitir uma maior acerto quanto aos Verdadeiros Positivos. Ou seja, é razoável considerar um *tweet* como possuidor de traço de violência, mesmo que ele não a possua. Desde que quando um *tweet* possuidor de traços de violência seja encontrado, ele seja predito corretamente como tal. Ao analisar os dados comparando os valores de *Recall* os dois melhores se mantêm como *Logistic Regression* e *Support Vector Machine* com 0.5, seguidos por *Random Forest* com 49.45% e por *Multinomial Naïve Bayes* com 46.19%

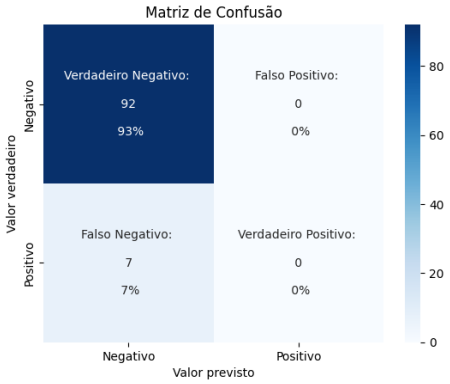


Figura 5.8: Matriz de Confusão para o *dataset* original com *Logistic Regression.*

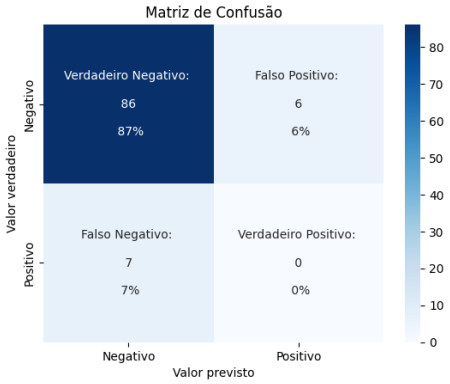


Figura 5.9: Matriz de Confusão para o *dataset* original para *Multinomial Naïve Bayes.*

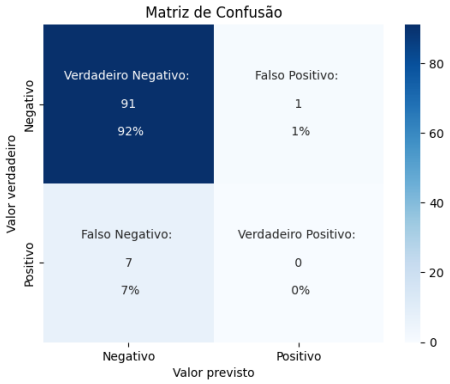


Figura 5.10: Matriz de Confusão para o *dataset* original com *Random Forest.*

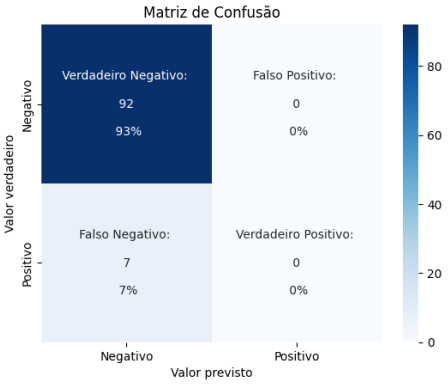


Figura 5.11: Matriz de Confusão para o *dataset* original com *Support Vector Machine.*

Nas Figuras 5.8, 5.9, 5.10 e 5.11 são espoxtas as matrizes de confusão para os modelos aplicados *Logistic Regression*, *Multinomial Naïve Bayes*, *Random Forest* e *Support Vector Machine*. A matriz de confusão apresenta quatro campos de informação os valores Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). Valores Verdadeiros Positivos são obtidos quando o modelo prediz corretamente quanto a presença de traços de violência em um *tweet*. Valores Verdadeiros Negativos são obtidos quando o modelo prediz corretamente quanto a ausência de traços de violência em um *tweet*. Quando o modelo prediz que existem traços de violência, mas está errado, se gera um Falso Positivo. Quando prediz que não existem traços de violência, mas estes traços existem, se gera um Falso Negativo. A matriz de confusão possui 4 quadrados, representando cada um das quantidades, o quadrado da esquerda no topo mostra a quantidade para Verdadeiro Negativo, o quadrado da esquerda no fundo indica o Falso Negativo. O quadrado da direita do topo ilustra a quantidade de Falso Positivo e quadrado da direita no fundo indica a quantidade de Verdadeiro Positivo. Os valores em porcentagem em cada quadrado relatam quantos porcento dos valores testados aquele quadrado representa em relação a todos os valores testados. A cor contida na barra lateral indica visualmente a quantidade de valores obtidos, quanto mais escuro o azul, em maior quantidade aquele valor foi predito. Quanto mais claro o azul, em menor quantidade aquele valor foi predito. O uso de matrizes de confusão tem a função de auxiliar na visualização do comportamento geral do modelo preditivo, quais são seus pontos fortes e fracos, sendo portanto uma ferramenta útil para tomar decisões mais embasadas sobre qual modelo utilizar dado o contexto de uso necessário.

Graças a predição com mesmas quantidade em Verdadeiros Negativos (92) e em Falsos Negativos (7) dos modelos *Logistic Regression* na matriz da Figura 5.8 e *Support Vector Machine* na matriz da Figura 5.11, as matrizes de confusão dos dois modelos (assim como suas outras métricas) acabaram por serem iguais. A presença apenas de valores no lado esquerdo das matrizes de confusão, responsável por descrever os valores

Negativos obtidos, informa que o modelo está tendencioso para considerar quaisquer *tweet* como não possuindo traços de violência, o que não está correto segundo a base de teste aplicada, sendo os piores valores quanto a esta métrica.

Quanto ao modelo *Random Forest* com predições descritas na Figura 5.10 acabou por produzir um resultado similar aos dois anteriores, produzindo 91 Verdadeiros Negativos, 7 Falsos Negativos e 1 único Falso Positivo. Levantando uma vez mais a suspeita quanto a tendência de apenas considerar que todo *tweet* processado não contem violência, sendo o segundo melhor quanto a esta métrica.

Por outro lado, o modelo *Multinomial Naïve Bayes* apresentado na Figura 5.9 com sua predição estatística resultou em 86 Verdadeiros Negativos, 7 Falsos Negativos e 6 Falsos Positivos, indicando que mesmo de forma falha, houve a tentativa de predizer alguns dos *tweets* da base, como contendo traços de violência. O que sugere um verdadeiro aprendizado para este modelo. Sendo o melhor dos 4, para esta métrica.

5.3 Classificação dos dados da base final

Nesta seção há a apresentação dos resultados obtidos ao classificar os dados da base final, após os dois processos de filtragem descritos na Subseção 4.1.4 no Capítulo 4. Esta seção apresenta a execução das etapas G, H e I da fase 3 descrita pela Figura 4.1 no Capítulo 4.

Os dados aqui contidos são decorrentes da chamada da função 'plot_cm_treinada' que pode ser visualizada na Figura 4.13 do Capítulo 4. Nesta Seção se utilizou dos métodos de Aprendizado Supervisionado *Logistic Regression*, *Multinomial Naïve Bayes*, *Random Forest* e *Support Vector Machine*, seu treinamento fora realizado pelos dados obtidos pela Subseção 4.1.2 do Capítulo 4.

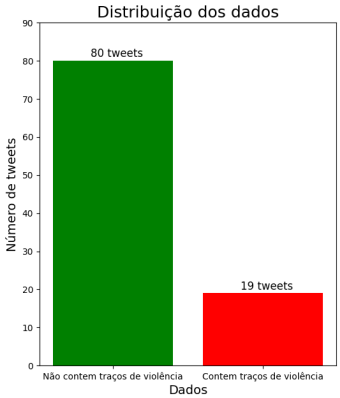


Figura 5.12: Distribuição dos dados de teste para a base final.

Consecutivo ao processo de treino dos modelos de aprendizado de máquina descritos em 4.1.5 ocorreu a testagem do aprendizado para a base de dados final, obtida após a etapa de filtragem descrita na Subseção 4.1.4 no Capítulo 4. De forma a testar os modelos treinados se seleciou aleatoriamente 100 *tweets* contidos na base final, a distribuição dos dados pode ser observada na Figura 5.12. Ao inserir estes 100 dados como entrada nos modelos previamente treinados de *Logistic Regression*, *Multinomial Naïve Bayes*, *Random Forest* e *Support Vector Machine* se produziu métricas, como Acurácia, *F1-score*, *Recall* e *Precision*, e estas estão dispostas na Tabela 5.2.

Modelo de Dados	**Acurácia**	**F1-Score**	**Recall**	**Precision**
Logistic Regression	0.82828	0.61086	0.59276	0.75268
Multinomial Naïve Bayes	0.80808	0.56508	0.56019	0.66397
Random Forest	0.81818	0.60152	0.58651	0.70419
Support Vector Machine	0.79797	0.44382	0.49375	0.40306

Tabela 5.2: Tabela sobre as métricas obtidas pelos modelos de Aprendizado de Máquina perante o *dataset* pós tratamento com ontologia

Ao observar os valores de Acurácia obtida temos que o método com melhores resultados para estes dados foi o *Logistic Regression* com 82.82%, seguido pelo *Random Forest* com 81.81%, *Multinomial Naïve Bayes* com 80.80% e o *Support Vector Machine* com 79.79%. Entretanto como já fora discutido na Seção 5.2 deste Capítulo, o parâmetro mais interessante para a análise da presença de traços de violência contra a mulher é o *Recall* por indicar que o modelo tende a predizer corretamente os casos de Verdadeiros Positivos, onde existem traços de violência, ele é capaz de identificar. Um fato interessante é de que mesmo ao mudar a ótica de análise, a ordem de qualidade de modelos se mantêm, confirmando que neste caso o melhor modelo foi o *Logistic Regression* com 59.27%, seguido pelo *Random Forest* com 58.65%, do *Multinomial Naïve Bayes* com 56.50% e o *Support Vector Machine* com 49.37%.

A seguir estão as matrizes de confusão produzidas pelos modelos de Aprendizado Supervisionado quando atribuídos os 100 *tweets* catalogados que foram selecionados aleatoriamente da base de dados final. A descrição do funcionamento de matrizes de confusão foi realizada na Seção 5.2 e não será replicada aqui.

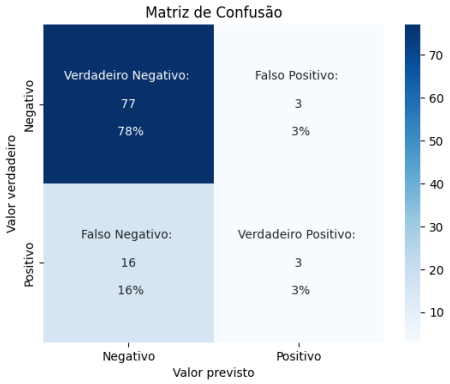


Figura 5.14: Matriz de Confusão para o *dataset* após processamento com ontologia com *Multinomial Naïve Bayes*

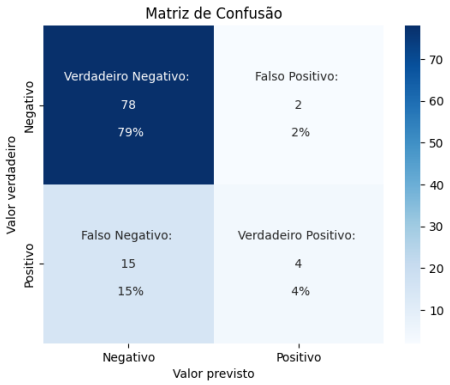


Figura 5.13: Matriz de Confusão para o *dataset* após processamento com ontologia com *Logistic Regression*

A matriz para o modelo *Logistic Regression* exibe 78 valores Verdadeiros Negativos, 15 valores Falsos Negativos, 2 valores Falsos Positivos e 4 Verdadeiros Positivos. O que indica um aprendizado na detecção dos *tweets* que contenham traços de violência contra a mulher, mas se mantêm a tendência em predizer quaisquer *tweet* como não contendo traços de violência, pois dos 19 *tweets* violentos enumerados pela Figura 5.12, a técnica acertou somente 4 deles.

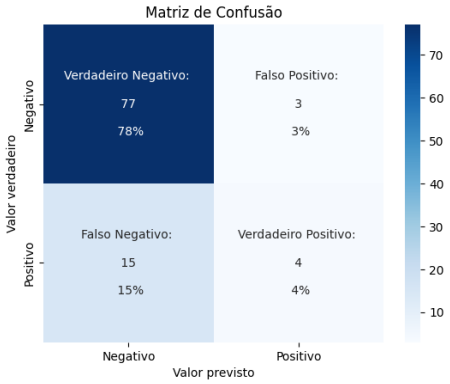


Figura 5.15: Matriz de Confusão para o *dataset* após processamento com ontologia com *Random Forest*

Os modelos *Multinomial Naïve Bayes* e *Random Forest* geraram matrizes muito similares, com os mesmos valores para Verdadeiro Negativo (77) e Falso Positivo (3), com uma ligeira diferença entre Falso Negativo e Verdadeiro Positivo, sendo 16 para o *Multinomial Naïve Bayes* e 15 *Random Forest* no Falso Negativo e 4 para o *Multinomial Naïve Bayes* e 3 para o *Random Forest* no Verdadeiro Positivo. Esta pequena variação em 1 predição para os Verdadeiros Positivos, aumentou o valor do *Recall* do *Random Forest* colocando-o na segunda colocação dos modelos em relação a esta métrica.

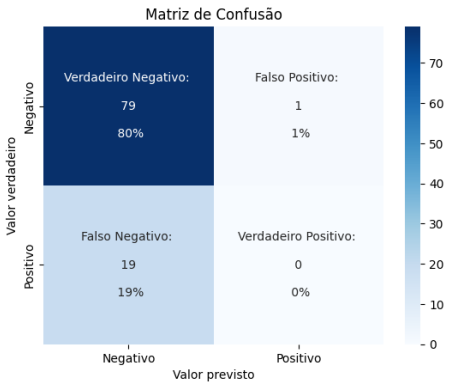


Figura 5.16: Matriz de Confusão para o *dataset* após processamento com ontologia com *Support Vector Machine*

Por fim temos a matriz produzida pelo modelo *Support Vector Machine* com seus 79 em Verdadeiro Negativo, 19 em Falso Positivo e 1 em Falso Negativo. No qual se manteve a indesejada tendência a predizer quaisquer *tweet* da amostra como não possuidor de traços de violência.

Considerando a fragilidade ao tema da pesquisa e sua relevância perante a sociedade, os valores de acurácia obtidos se assemelham aos localizados nos trabalhos dos pesquisadores de [3] com 0.75 de acurácia e pelos cientistas de [14] com 0.789, nos quais

pesquisaram sobre violência contra a mulher no idioma espanhol. Ainda assim possui pontuação mais baixa do que os 0.97 de acurácia obtido pelos autores de [37] buscando por misoginia no idioma turco. Todavia não são suficientes para permitir uma análise automática em quaisquer situação pois, há uma tendência em presumir que não existe violência nos dados, fato confirmado pela baixa presença de falsos positivos e verdadeiros positivos demonstrados nas matrizes de confusão 5.13, 5.14, 5.15 e 5.16 que possuem um leve incremento quanto aos acertos em casos positivos, quando comparados as matrizes 5.8, 5.9, 5.10 e 5.11 da base de dados original. Para lidar com violência seria mais valioso ter mais Verdadeiros Positivos do que Verdadeiros Negativos, o que não é o caso.

Mediante os resultados obtidos se pode dizer que o processo da criação de ontologia auxilia na remoção de ruído quando diante de um grande volume de dados, graças a seu processo de seleção mais elegante de palavras e um afunilamento para permear a observação mais direta quanto a uma característica. Ao incrementar a assertividade das detecções de padrões, fato ressaltado pelo aumento dos valores *Recall* quando se observa a tabela 5.1 e 5.1. Todavia este progresso é custoso para o ganho obtido, e fica a cargo do leitor a decisão quanto a sua aplicabilidade em seu próprio contexto.

5.4 Conclusão

O dilema presente é um volume de ruído alto, encontrado em *tweeets* coletados que abordam o tema de violência contra a mulher. Para remediar tal impasse se propôs a criação de uma ontologia com o mesmo tema, com ensejo de remover o ruído encontrado. A forma de avaliar o impacto da ontologia utilizada foi o uso de aprendizado de máquina supervisionado.

O texto proveniente de redes sociais é informal e precisou ser pré-processado por meio de remoção de caracteres especiais, *hyperlinks* e de *stopwords*, da transformação do texto em minúsculas, da sua respectiva tokenização e lematização. Realizou-se a filtragem dos dados obtidos, em primeiro plano diminuindo a base de 1032852 *tweets* para 87048 *tweets*, e após notar que algumas das palavras selecionadas foram filtradas erroneamente por conterem subpalavras que estavam na ontologia, uma nova etapa de filtragem ocorreu descendo dos 87408 *tweets* para 84112, ao remover os *tweets* que as continham. Essas atividades mostraram que a base de dados detinha de muitos dados pertencentes ao mercado financeiro e originalmente de muitos *emojis*, que ao executar os passos ditos aqui, o tema economia se manteve relevante, enquanto o ruído dos *emojis* desapareceu.

Conforme testes do aprendizado de máquina supervisionado aplicado, o método com os melhores resultados encontrados na base de estudo foi o *Logistic Regression*, que foi consistente em todos os testes, mesmo não sendo o melhor em alguns, sempre foi no mínimo o segundo lugar. Enquanto que o pior foi o *Support Vector Machines* que oscilou

bastante, algumas vezes era o melhor dos modelos e outras foi o pior. Foi um desafio de pesquisa rotular os dados para que os modelos supervisionados pudessem treinar, e mesmo com ajuda da equipe de pesquisa, apenas 1 mil dados de treino puderam ser produzidos e 100 dados de teste para cada base de dados. A baixa presença de valores com violência foi um agravante, pois o rebalanceamento artificial dos dados não surtiu o efeito desejado, tornando os modelos piores do que sem a sua realização. Outro fator relevante foi a ampla variabilidade dos modelos quanto a hiper-parâmetros e para não endossar nenhuma característica que pudesse prejudicar o desempenho das demais, se optou a avaliar a versão mais simples de cada modelo.

A execução da solução proposta tornou claro o potencial benéfico do uso de ontologias, como uma etapa a mais na remoção de ruído textual em textos de violência contra a mulher. O processo para tal realização é custoso, entretanto o ganho recebido em troca compensa. Das dificuldades encontradas se ressaltam a tribulação em localizar ontologias extensas para estudo, o processo moroso de classificar dados à mão e a limitação de *hardware* por conta do autor que impossibilitou testes massivos.

A pesquisa realizou uma análise comparativa quanto o aprendizado de máquina supervisionado, testando os resultados em uma base classificada manualmente com um montante aproximado de 10% (100 *tweets*) perante o volume dos dados de treino 1000 *tweets*, visto a limitação temporal e social de produzir uma grande base de dados catalogada para tal, emerge como possibilidade para estudos futuros a ampliação do volume dos dados de treino e teste. A monografia optou pela simplicidade em analisar separadamente as respostas obtidas por cada método de aprendizado de máquina, entretanto uma nova abordagem possível é a de realizar *ensemble* de técnicas de aprendizado de máquina, onde cada técnica avaliaria o *tweet* quanto a presença de traços de violência contra a mulher e realizaria seu voto, ao realizar uma média dos resultados o *ensemble* classificaria o *tweet* como violento ou não. Um recurso produzido pela monografia foi a ontologia que categoriza a violência contra a mulher, em algumas classes e instancia as ofensas quanto a estas classes. Uma expansão do estudo seria em avaliar as classes produzidas adaptando-as, adicionando ou removendo classes e palavras, como tentativa de refino, com posterior avaliação de impacto.

Referências Bibliográficas

[1] **(in)visibility of violence against women in mental health1**. *Psicologia: Teoria e Pesquisa*, 32:1–7, 2016.

[2] **Racist and sexist hate speech detection: Literature review**. p. 95–99. Institute of Electrical and Electronics Engineers Inc., 10 2020.

[3] **Understanding violence against women in digital space from a data science perspective : Full/regular research papers - csci-isna**. p. 263–269. Institute of Electrical and Electronics Engineers Inc., 12 2020.

[4] **Ensemble based hinglish hate speech detection**. p. 1800–1806. Institute of Electrical and Electronics Engineers Inc., 5 2021.

[5] **Lei carolina dieckmann: 10 anos da lei que protege a privacidade dos brasileiros no ambiente virtual**, jul 2023.

[6] **Quantidade de homens e mulheres**, jul 2023.

[7] **Violência contra a mulher gera prejuízo de r**1*bilhoparaeconomiabrasileira*, *jul*2023.

[8] 7GRAUS. **Ruido**, jul 2023.

[9] ALMEIDA, A. B. M.; BAX, M. P. **Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção**. Technical report, 2003.

[10] CHAI, C. **The importance of data cleaning: Three visualization examples**. *CHANCE*, 33:4–9, 01 2020.

[11] DEFERSHA, N. B.; KEKEBA, K.; KALIYAPERUMAL, K. **Tuning hyperparameters of machine learning methods for afan oromo hate speech text detection for social media**. p. 596–604. Institute of Electrical and Electronics Engineers Inc., 2021.

[12] EIJST, G. V. H.; CHREIBER, H. S.; IELINGA, B. J. W. **Using explicit ontologies in kbs development**, 1997.

[13] FILHO, W. L.; LEAL, W.; ANABELA, F. ; AZUL, M.; BRANDLI, L.; LANGE, A.; WALL, S. T. **Encyclopedia of the un sustainable development goals series editor: Gender equality**.

[14] GONZALEZ, G. A. R.; CANTU-ORTIZ, F. J. **A sentiment analysis and unsupervised learning approach to digital violence against women: Monterrey case**. p. 18–26. Institute of Electrical and Electronics Engineers Inc., 3 2021.

[15] GRUBER, T. R. **A translation approach to portable ontology specifications**. Technical report, 1993.

[16] GRÜNINGER, M.; FOX, M. **Methodology for the design and evaluation of ontologies**. 07 1995.

[17] GUARINO, N. **Formal ontology in information systems**. Technical report, 1998.

[18] HENRIQUE, P.; CHAVES, B.; OLIVEIRA, E. C. **Desenvolvimento de ontologia para estruturas organizacionais do governo brasileiro**. Technical report, 2015.

[19] INSTITUTE, B. **Fighting gender-based violence in brazil**, jul 2023.

[20] KOMPUTER, U. D. P. S. S.; OF ELECTRICAL ENGINEERING, U. D. D.; OF ELECTRICAL, I.; SECTION, E. E. I.; OF ELECTRICAL, I.; ENGINEERS, E. **2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE).**

[21] LOPEZ, F. **Overview of methodologies for building ontologies**. In: *International Joint Conference on Artificial Intelligence*, 1999.

[22] LYNN, T.; ENDO, P. T.; ROSATI, P.; SILVA, I.; SANTOS, G. L.; GING, D. **A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary; a comparison of machine learning approaches for detecting misogynistic speech in urban dictionary**. Technical report, 2019.

[23] MACHADO, J. **Pesquisa aponta aumento de violência contra a mulher no brasil em 2022 e integrantes do comitê de equidade comentam os números**, jul 2023.

[24] MARPAUNG, A.; RISMALA, R.; NURRAHMI, H. **Hate speech detection in indonesian twitter texts using bidirectional gated recurrent unit**. p. 186–190. Institute of Electrical and Electronics Engineers Inc., 1 2021.

[25] MARR, B. **How much data do we create every day? the mind-blowing stats everyone should read**, jul 2023.

[26] MARX, K. **Para a Crítica da Economia Política**. Abril Cultural, São Paulo, 1974.

[27] MIZOGUCHI, R.; WELKENHUYSEN, J.; IKEDA, M. **Task ontology for reuse of problem solving knowledge**. *Towards Very Large Knowledge Bases*, 01 1995.

[28] MOHAN, J.; GUPTA, A.; OF INFORMATION TECHNOLOGY UNIVERSITY, J. I.; OF ELECTRICAL, I.; CHAPTER, E. E. U. P. S. S. J.; OF ELECTRICAL, I.; ENGINEERS, E. **2019 International Conference on Signal Processing and Communication (ICSC) : 07-09 March 2019, Jaypee Institute of Information Technology, NOIDA**.

[29] MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens**, 2007.

[30] MÜLLER, A. C.; GUIDO, S. **Introduction to machine learning with python a guide for data scientists introduction to machine learning with python**.

[31] NOY, N.; MCGUINNESS, D. **Ontology development 101: A guide to creating your first ontology**. *Knowledge Systems Laboratory*, 32, 01 2001.

[32] OF ELECTRICAL, I.; ENGINEERS, E. **2019 National Information Technology Conference (NITC).**

[33] OF ELECTRICAL, I.; ENGINEERS, E.; ON SOCIAL IMPLICATIONS OF TECHNOLOGY, I. S. **2019 IST-Africa Week Conference : 08-10 May 2019, Nairobi, Kenya.**

[34] PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[35] RINI.; UTAMI, E.; HARTANTO, A. D. **Systematic literature review of hate speech detection with text mining**. Institute of Electrical and Electronics Engineers Inc., 10 2020.

[36] RUSSELL, S. J. S. J.; NORVIG, P.; DAVIS, E. **Artificial intelligence : a modern approach**.

[37] SAHI, H.; KILIC, Y.; SAGLAM, R. B. **Automated detection of hate speech towards woman on twitter; automated detection of hate speech towards woman on twitter**. Technical report, 2018.

[38] SALGADO, K. R.; PRODÓCIMO, E. **Bullying e cyberbullying:: duas faces da mesma realidade**. *Revista de Estudos Universitários - REU*, 42(2), fev. 2017.

[39] SANTOS, J. S. D.; PAES, A.; BERNARDINI, F. **Combining labeled datasets for sentiment analysis from different domains based on dataset similarity to predict electors sentiment**. p. 455–460. Institute of Electrical and Electronics Engineers Inc., 10 2019.

[40] SMITH, D. R. **The design of divide and conquer algorithms**. *Science of Computer Programming*, 5:37–58, 1985.

[41] STAAB, S.; STUDER, R.; SCHNURR, H.-P.; SURE-VETTER, Y. **Knowledge processes and ontologies**. *IEEE Intelligent Systems*, 16:26–34, 02 2001.

[42] UNIVERSITY, P.; SILIWANGI, U.; OF ELECTRICAL, I.; CHAPTER, E. E. I. S. C. J.; OF ELECTRICAL, I.; ENGINEERS, E. **ISCECC 2019 : International Conference on Sustainable Engineering and Creative Computing : proceedings : Bandung, 20-22 August 2019**.

[43] USCHOLD, M.; KING, M. **Towards a methodology for building ontologies**. 1995.

[44] USCHOLD, M.; GRUNINGER, M. **Ontologies: Principles, methods and applications**, 1996.

[45] USCHOLD, M.; JASPER, R. **A framework for understanding and classifying ontology applications proceedings of the ijcai-99 workshop on ontologies and problem-solving methods (krr5)**, 1999.

[46] VOLPETTI, C. **Hybrid deep learning for sentiment analysis and hate speech detection**.