UNIVERSIDADE FEDERAL DE GOIÁS
INSITUTO DE INFORMÁTICA

ISABELA FELIX FRANÇA

Identificação de traços de violência contra mulher em postagens do Twitter

Goiânia
2022

UNIVERSIDADE FEDERAL DE GOIÁS

INSITUTO DE INFORMÁTICA

AUTORIZAÇÃO PARA PUBLICAÇÃO DE MONOGRAFIA EM FORMATO ELETRÔNICO

Na qualidade de titular dos direitos de autora, **AUTORIZO** o Insituto de Informática da Universidade Federal de Goiás – UFG a reproduzir, inclusive em outro formato ou mídia e através de armazenamento permanente ou temporário, bem como a publicar na rede mundial de computadores (*Internet*) e na biblioteca virtual da UFG, entendendo-se os termos "reproduzir" e "publicar" conforme definições dos incisos VI e I, respectivamente, do artigo 5º da Lei nº 9610/98 de 10/02/1998, a obra abaixo especificada, sem que me seja devido pagamento a título de direitos autorais, desde que a reprodução e/ou publicação tenham a finalidade exclusiva de uso por quem a consulta, e a título de divulgação da produção acadêmica gerada pela Universidade, a partir desta data.

Título: Identificação de traços de violência contra mulher em postagens do Twitter

Autor(a): Isabela Felix França

Goiânia, 11 de Abril de 2022.

Isabela Felix França – Autora

Dra. Deborah Silva Alves Fernandes – Orientadora

Msc. Márcio Giovane Cunha Fernandes – Co-Orientador

ISABELA FELIX FRANÇA

Identificação de traços de violência contra mulher em postagens do Twitter

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

Área de concentração: Ciência da Computação.

Orientadora: Profa. Dra. Deborah Silva Alves Fernandes
Co-Orientador: Prof. Msc. Márcio Giovane Cunha Fernandes

Goiânia
2022

ISABELA FELIX FRANÇA

Identificação de traços de violência contra mulher em postagens do Twitter

Trabalho de Conclusão apresentado à Coordenação do Curso de Ciência da Computação do Insituto de Informática da Universidade Federal de Goiás como requisito parcial para obtenção do título de Bacharel em Ciência da Computação, aprovada em 11 de Abril de 2022, pela Banca Examinadora constituída pelos professores:

Profa. Dra. Deborah Silva Alves Fernandes
Insituto de Informática – UFG
Presidente da Banca

Prof. Msc. Márcio Giovane Cunha Fernandes
Instituto Acadêmico de Ciências Tecnológicas – UEG

Prof. Dr. Fabrízzio Alphonsus Alves de Melo Nunes Soares
Instituto de Informática – UFG

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ISABELA FELIX FRANCA

Identificação de traços de violência contra a mulher em postagens do Twitter

Trabalho de conclusão de curso apresentado à Universidade Federal de Goiás como parte dos requisitos para a obtenção do título de Bacharel em Ciências da Computação.

Orientador: Profa. Dra. Deborah Silva Alves Fernandes
Coorientador: Prof. Me. Márcio Giovane Cunha Fernandes

Aprovado em 11/04/2022.

BANCA EXAMINADORA

Profa. Dra. Deborah Silva Alves Fernandes
Universidade Federal de Goiás
Instituto de Informática

Prof. Me. Márcio Giovane Cunha Fernandes
Universidade Federal de Goiás
Instituto de Informática

Prof. Dr. Fabrízzio A. A. de M. N. Soares
Universidade Federal de Goiás
Instituto de Informática

Isabela Felix França

Durante a graduação, foi monitora de Compiladores, para os departamentos de Engenharia da Computação e Ciência da Computação da UFG; pesquisadora volutária do CNPq, em um trabalho de iniciação científica, para o departamento de Ciência da Computação e voluntária no projeto de extensão em Organização e Exposição do Museu do INF. Foi estagiária em Suporte ao Usuário, para Centro de Recursos Computacionais da UFG e para o Ministério Público do Estado de Goiás; e desenvolvimento Full Stack. Atualmente, trabalha como programadora.

Agradecimentos

Agradeço a vocês, que estiveram do meu lado, em todos momentos e nunca me deixaram desistir. Que mesmo eu não sabendo o caminho, nunca me deixaram sozinha nessa caminhada.

A minha orientadora e meu coorientador, obrigada pelo respeito, confiança, paciência, dedicação, ensinamentos e por todos conselhos que me permitiram não só finalizar este trabalho como me tornar uma pessoa e uma profissional mais qualificada.

Para sobreviver, nós nos prendemos a tudo que conhecemos e entendemos. Rotulando isso como realidade. Contudo, conhecimento e entedimento são imprecisos, esta realidade poderia ser uma ilusão. Todos os humanos vivem com suposições erradas. Porém, esta não é apenas outra forma de olhar para o assunto?

Itachi,
Naruto.

Resumo

França, Isabela. **Identificação de traços de violência contra mulher em postagens do Twitter**. Goiânia, 2022. 49p. Monografia de Especialização. Insituto de Informática, Universidade Federal de Goiás.

Em janeiro de 2022, [DATAREPORTAL], o Brasil possuía uma população de 214,7 milhões, onde 171,5 milhões possuíam redes socias, dentre os quais 19,05 milhões eram usuários do Twitter. Conforme [Instituto Patrícia Galvão, AASP, Mansuido, Nogueira], o uso da internet trouxe novos meios de violência contra mulher, dentre estas, sextorção, publicação de imagens íntimas, estupro virtual, perseguição, são algumas das novas formas de violência, incluindo, discurso de ódio.

Tendo em mente as situações de violência que uma mulher já enfrenta em seu dia a dia, este trabalho busca tentar oferecer mais segurança em um ambiente que vem cada vez mais sendo utilizado, as redes socias, mais precisamente, o Twitter. Neste aborda-se a identificação de violência contra mulher em tweets, através do uso de classificadores e *ensemble*, buscando validar qual seria a melhor abordagem para identificação desta.

Random Forest foi o classificador que obteve melhor pontução F1, sendo 81,5%. A abordagem *ensemble* obteve 78,9% de pontuação. Apesar de não apresentar melhor desempenho que um único classificador, foi possível verificar que uma abordagem com *ensemble* pode melhorar o resultado no quesito falso negativo e positivo.

Palavras–chave

Violência contra mulher, *ensemble*, Twitter.

Abstract

França, Isabela. **Identification of traces of violence against women in Twitter posts**. Goiânia, 2022. 49p. Monografia de Especialização. Insituto de Informática, Universidade Federal de Goiás.

In January 2022, [DATAREPORTAL], Brazil had a population of 214.7 million, of which 171.5 million had social networks, among which 19.05 million were Twitter users. According to [Instituto Patrícia Galvão, AASP, Mansuido, Nogueira], the use of the internet has brought new means of violence against women, among them, sextortion, publication of intimate images, virtual rape, persecution, are some of the new forms of violence, including hate speech.

Bearing in mind the situations of violence that a woman already faces in her daily life, this work tries to offer more security in an environment that is increasingly being used, social networks, more precisely, Twitter. This addresses the identification of violence against women in tweets, through the use of classifiers and ensemble, seeking to validate what would be the best approach to identify it.

Random Forest was the classifier that obtained the best F1 score, being 81.5%. The ensemble approach scored 78.9%. Despite not presenting better performance than a single classifier, it was possible to verify that an ensemble approach can improve the result in the false negative and positive aspects.

Keywords

Violence against women, ensemble, Twitter

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

1.1 Contextualização

De acordo com [Cetic], em 2020, o Brasil possuía 152 milhões de usuários de internet, sendo 85% do sexo feminino e 77% do sexo masculino, onde o principal dispositivo utilizado foi o celular. Dentre as atividades realizadas no aparelho, as principais elencadas foram envio de mensagens, chamada de voz ou vídeo e uso de redes sociais.

Em janeiro de 2022, conforme relatório disponibilizado pelo portal independente Datareportal[DATAREPORTAL], a quantidade de usuários de internet no Brasil alcançou os 77% da população (214,7 milhões). Em relação a quantidade de usuários em mídias sociais, são 171,5 milhões, isto reflete diretamente no amontoado de dados gerados por dia, dentre eles, vídeos, imagens, mensagens e etc. Essa quantidade de informação, permitiu que pesquisadores aplicassem técnicas de *machine learning* a fim de extrair caracterísiticas, como, opiniões e sentimentos. Tal fator pode auxiliar a tomada de decisão ou até mesmo identificação de determinados problemas que surgem com uso da internet, como, por exemplo, o discurso de ódio.

Dentre os vários aplicativos de redes sociais online, nos quais os usuários expressam opiniões através de mensagens, o Twitter têm sido popular no país. O mesmo possuía 19,05 milhões de usuário no ínicio do ano, ficando em quarto lugar no rank de países com mais usuários, [DATAREPORTAL]. Como se trata de uma rede com a maioria dos perfis públicos e com quantidade reduzida de caracteres, têm sido um ambiente para emissão de opiniões, comentários sobre situações do cotidiano no momento em que acontecem e também para espalhamento de discursos sobre diversos temas, incluindo racismo, ódio, violência e outros. Dentre essas diversas manifestações nas redes sociais, neste projeto final de curso a pesquisadora dedicou-se ao contexto de violência contra a mulher.

No Brasil, atualmente, a cada 7 horas, uma mulher morre de feminícidio. Quanto ao aspecto da violência física, 67% das vítimas, são mulheres, [Institute]. Du-

rante a pandemia, de acordo com o portal de notícias G1[1], uma em quatro mulheres foram vítimas de alguma tipo de violência, seja esta física, psicológica ou sexual. Já a violência dentro de casa, passou de 42% para 48,8%. Em 2020, os canais 100 e 180 receberam mais de 105 mil denúncias de violência contra mulheres, [Ministério da Mulher, da Família e dos Direitos Humanos].

Com relação ao mundo virtual, segundo [Instituto Patrícia Galvão, AASP, Mansuido, Nogueira], o uso da internet trouxe novos meios de violência contra mulher, dentre eles, podem ser citados: sextorção, publicação de imagens íntimas, estupro virtual, perseguição e discurso de ódio.

Tendo em mente as situações de violência que uma mulher já enfrenta em seu dia a dia, este trabalho busca promover o estudo e utilização de ferramentas computacionais para auxílio na identificação automática de traços de discurso de violência contra a mulher em textos publicados na plataforma Twitter. Tais atividades podem auxiliar na promoção de políticas públicas para esta população e também na detecção e auxílio à tomada de decisão para retenção de espalhamento desse discurso dentre outras possibilidades.

1.2 Definição do Problema

Neste projeto final de curso aborda-se a identificação de traços de violência contra mulher, através do discurso de ódio, em tweets. A identificação será dada através do uso de classificadores independentes e em *ensemble* (quando um conjunto de classificadores independentes são combinados para classificação). Busca-se validar qual seria a melhor abordagem para tal identificação.

1.3 Objetivo

O objetivo deste é indentificação de traços de violência contra mulher em tweets, através do estudo, implementação e verificação de técnicas de *machine learning* para classificação com algoritmos trabalhando em modo independente e em *ensemble*.

1.4 Apresentação da monografia

O texto desta monografia apresenta seções específicas para trabalhos relacionados, fundamentação teórica, experimento, resultados e conslusão.

[1]Uma em cada quatro mulheres foi vítima de algum tipo de violência na pandemia no Brasil, aponta pesquisa. 07/06/2021. https://g1.globo.com/sp/sao-paulo/noticia/2021/06/07/1-em-cada-4-mulheres-foi-vitima-de-algum-tipo-de-violencia-na-pandemia-no-brasil-diz-datafolha.ghtml

A seção de trabalhos relacionados - capítulo 2- foi divida em duas subseções, uma que trata da detecção de violência e outra sobre o apoio às vítimas. No capítulo 3, referente a fudamentação, é apresentando o que é violência contra mulher, a ocorrência no meio digital, assim como, leis brasileiras que abordam o tema. Na segunda parte deste mesmo capítulo são apresentados os classificadores que serão utilizados e as medidas de avaliação de desempenho a serem adotadas na etapa de avaliação. O experimento é apresentado no capítulo 4 que abordará a coleta de dados, o pré-processamento realizado, a rotulação de dados e a implementação. Por fim, têm-se os capítulos 5 e 6 os quais descrevem os resultados obtidos e a conclusão, respectivamente. No capítulo de resultados serão encontradas a avaliações dos classificadores e do *ensemble*. Já na conclusão, as considerações finais dos resultados e trabalhos futuros.

CAPÍTULO 2

Trabalhos Relacionados

Este capítulo apresentará trabalhos que tratam da violência contra mulher. Porém, também foram adicionados trabalhos que incluiam violência contra criança e familiar, inclusive formas de apoio às vítimas.

2.1 Detecção de Violência

Em [Corrêa e Faria 2021], têm-se um trabalho brasileiro, ao qual propõe um método de análise de relatos de violência contra mulheres por meio de técnicas de mineração de texto para caracterizar a agressão ou o agressor. A solução proposta pelas autoras é a aplicação de mineração de texto em relatos das vítimas escritos em língua portuguesa. Como resultado, o classificador adotado obteve 71,2% de acerto para violência constante ou esporádica. Em relatos constantes, boa parte foram associados a alguém próximo à vítima. Ainda, para agressões constantes, as palavras falar, dizer, denunciar e casa são frequentes. Já no LDA, tópicos de abuso familiar por parte do irmão ou pai e violência sexual, está contido como agressão. Já nos relatos esporádicos, também foi identificado, casos de violência fora de casa e abuso familiar.

Um modelo de processamento de linguagem natural para classificação de mensagens do Twitter e Reddit, que continham violência de gênero, é apresentado em [Soldevilla e Flores 2021]. Os pesquisadores coletaram 113910 Reddits e 30377 tweets. Neste foram aplicados critérios de expressões regulares e contagem de caracteres. Por fim, foram tokenizados através do *BERT tokenizer*, resultando em 4465 mensagens do Reditts e 11956 tweets. Os pesquisadores utilizaram um modelo pré-treinado *BERT*, sendo a versão escolhida *bert-base-uncased*. Os dados foram divididos em três conjuntos, 64% para treinamento, 16% para validação e 20% para avaliação, fazendo uso de *cross-validation* com *fold* igual a 5. Para este, o melhor resultado foi obtido quando o modelo foi treinado utilizando os dados da duas redes sociais em conjunto, resultando em 96% de acurácia.

Em [Kumar e Aggarwal 2019] têm-se o objetivo de verificar a segurança das mulheres na Índia. Para chegar a conclusão de qual cidade é mais segura, realizaram coleta de tweets. Esses foram processados através de técnicas de análise de sentimento,

aprendizado de máquina e processamento de linguagem natural. Após avaliações dos textos, obtiveram várias conclusões sobre os mesmos incluindo que, se a quantidade de tweets neutros forem altos, o assunto em questão não é muito significativo difícultando a análise. Para os dados analisados, concluiram que as cidades de Chennai e Delhi são a mais e menos segura, respectivamente.

Os pesquisadores de [González, Gabarrot e Cantu-Ortiz 2020] utilizaram de técnicas de *Data Science*, afim de detectar padrões, desenvolver modelos e realizar uma classificação automática. Estes optaram por uma representação vetorial e para neutralizar a dispersão usaram *bag of words*. As técnicas adotadas foram mineração de texto, análise de sentimentos, reconhecimento de padrões e aprendizagem não supervisionada, sendo que para técnica foram utilizados *Representative-Based Algorithms*, *Probabilistic Text Clustering Algorithms*, *Co-Clustering* e *Probabilistic Latent Semantic Analysis*, onde estes passaram por adaptações. Para teste do *framework* desenvolvido, os mesmos utilizaram duas fontes de dados, *SemEval* e os dados originados de outro trabalho, ao qual foi chamado de Zenodo. Este geraram 4 fontes de dados, onde os dados do *SemEval 2019 Task 5*[1] foram divididos em 3 partes, para cada parte deste, foi obtido 64%, 68,8% e 78,9% de acurácia. Já para Zenodo, obteve-se 78,3%. Os autores concluíram que o problema abordado é de difícil classificação, visto que obtiveram resultados abaixo de 80%.

O modelo desenvolvido em [Saha et al. 2018] objetivou a identificação automática de misógenia fazendo uso de aprendizado de máquina. O trabalho desenvolvido foi submetido ao EVALITA 2018[2], ao qual foi dado como vencedor. Neste haviam duas tarefas, A e B. A primeira era a identificação de twitter em misóginos e não misóginos, onde tiveram primeiro lugar com precisão de 70,4%, e a segunda era determinar ao qual categoria os misóginos pertenciam e o alvo. Neste último alcançaram o quinto lugar com F1-score de 0,37. Os dados utilizados foram os disponibilizados pelo próprio evento, sendo este rotulados, divididos em misoginia, categorias de misoginia e alvo e não rotulados. O pré-processamento realizado envolvia *Embeddings* de frases, Vetores *TF-IDF* e Vetores *BOW*. As técnicas foram utilizadas em conjunto devido o melhor resultado alcançado. Foram utilizados três modelos, regressão logística, que foi o classificador que lhes renderam o primeiro lugar para tarefa A, XGBoost e CatBoost. Já para tarefa B foi utilizado CatBoost. É relatado que o baixo resultado na tarefa B é devido ao alto desequilíbrio de dados. Este modelo foi tornado público.

Estudar a relação de mídias sociais que, ao mesmo tempo em que servem como canal de denúncia também são utilizadas para disseminação de discurso de ódio, é o objetivo de [Rodrigues, Júnior e Lobato 2019]. Para tal, foi analisada a tentativa de

[1] https://aclanthology.org/S19-2007/
[2] https://www.evalita.it/evalita-2018/tasks/

feminicídio de Elaine Caparroz. Os dados utilizados advêm da coleta de informações básicas de notícias referente ao assunto, assim como comentários da notícia coletada. Das notícias foram extraídos título, resumo, data da publicação, link. Já dos comentários foram o autor, data da publicação, conteúdo, *likes* e *dislikes*. Após coleta, os dados passaram por pré-processamento através do uso da biblioteca NLTK[3]. Estes ainda foram classificados quanto sua polaridade fazendo uso de intervalos numéricos. Para esta análise de sentimento foi utilizada a biblioteca Polyglot. Para modelagem de tópicos foi utilizado *Latent Semantic Analysis (LSA)* e *Latent Dirichlet Allocation (LDA)*. Foi ainda construída uma matriz de pesos fazendo uso de *Term Frequency-Inverse Document Frequency (TF-IDF)*. Os resultados indicaram que a culpa foi atribuída à vítima. A análise de sentimento mostrou que a quantidade de comentários negativos é maior que a neutra e a positiva, tendo porcentagens de 40.58%, 37,4% e 22,01%, respectivamente. Além disso, foi verificado que a maioria dos comentários negativos foram publicados por pessoas com escolaridade de 5ª a 8ª série do ensino fundamental. Outra característica levantada sobre os dados foi a identificação de posicionamento machista nos comentários.

Realizar modelagem de tópicos em textos sobre violência doméstica foi o objetivo de [Xue, Chen e Gelles 2019]. Para isso o mesmo faz uso de aprendizado de máquina não supervisando com *Latent Dirichlet Allocation - LDA*. Os dados utilizados foram obtidos Twitter, através de busca por tweets que continham a palavra chave "*domestic violence*". O LDA, foi configurado para gerar 20 distribuições de tópico usando unidade de bigramas estrutural. Os dados passaram por um pré-processamento que consistiu em remoção de #, citação, caracteres não ingleses e urls; conversão do twitter em matriz de termo usando *CountVectorizer*, determinação do número para entrada do LDA. Como resultado, obtiveram os 20 pares de palavras mais comuns, que eram: "*violence awareness (conscientização da violência)*", "*greg hardy (Greg resistente)*", "*awareness month (mês de conscientização)*", "*victims domestic (violência doméstica)*", "*stop domestic (parar doméstico)*"e "ronda rousey", sendo que para "ronda rousey"e "greg hardy", está relacionado a determinados casos de violência que estavam em alta no momento. Os autores citam que fazer uso de apenas uma palavra chave ao fazer a coleta de dados pode limitar o resultado, sugerindo a adição de "*intimate and sexual violence-related*".

A finalidade de [Sahi, Kilic e Saglam 2018] é a detecção automática de discurso de ódio e para isso fez uso de aprendizado de máquina. Para tal, coletaram do Twitter dados que continham a *hashtag* #kiyafctimckai'isma (tirem as mãos da minha roupa). Os tweets colhidos foram rotulados manualmente, separados por discurso ou não de ódio. Passaram também por rotinas de pré-processamento, e por fim, pelo algoritmo de lematização Zemberek. Utilizaram unigrama, bigrama e trigrama e ponderação através de

[3]https://www.nltk.org/

TF-IDF e, *Flesch- Kincaid Grade Level* e *Flesch Reading Ease* para pontuar a qualidade. Aplicaram as abordagens de *Support Vector Machines (SVM)*, com os *kernels*, polinomial e *RBF*, J48, *Naive Bayes*, *Random Forest* e *Random Tree*, com *cross-validation* de *fold* igual a 4. Como resultado, os melhores classificadores foram *Naive Bayes* e *SVM*, sendo que *SVM* teve melhor resultado, porém com baixo valor de *recall*. O valor baixo para *recall* deu-se devido os falso positivo. Um outro problema identificado foi a dificuldade de classificação, ou seja, o discernimento do que era ou não discurso de ódio.

Em [Khatua, Cambria e Khatua 2018], os pesquisadores buscaram classificar agressões sexuais por localização e por seus agressores, em tweets, por meio de aprendizado profundo. A separação da agressão consiste em: local de trabalho por colegas, instituição de ensino por professores ou colegas, locais públicos por estranhos, casa por um familiar, múltiplas ocorrências de agressões ou um tweet genérico sobre violência sexual. Os dados coletados são pertecentes a *#Metoo*[4]. Nestes foram aplicadas técnicas de *Multilayer Perceptron (MLP)*, *convolutional Neural Network (CNN)*, *Long Short-Term Memory (LSTM)* e *Bidirectional LSTM (Bi-LSTM)*. Quanto à precisão dos modelos, os melhores resultados foram obtidos usando *CNN* (0,83) sendo 0,82 para *LSTM*, 0,81 para *Bi-LSTM* e 0,77 para *MLP*. Também foi possível concluir uma relação de 1 em 4 agressões ocorrem por desconhecidos, logo as agressões por conhecidos ocorrem em maior quantidade. Observou-se que quando um tweet pode ser incluído em mais de uma categoria e é genérico, gera uma classificação errada.

2.2 Apoio às Vítimas

Um sistema para a identificação de postagens com conteúdo de abuso é proposto por [Subramani, Vu e Wang 2017]. A intenção é promover ações que ajudem as vítimas. Para tal, utilizaram de *machine learning* para classificação de intenções em textos reconhecendo-as como abuso ou conselho/opinião. Os dados foram extraídos do Facebook, pré-processados e rotulados pelos pesquisadores. Para a realização do experimento utilizaram duas abordagens, análise de características psicolinguísticas com *Linguistic Inquiry and Word Count (LIWC)* e modelo de classificação de classe e termo com *TF-IDF* em conjunto com algoritmo qui-quadrado que resulta em *bag-of-tokens*. Por fim, aplicaram *Support Vector Machine (SVM)*, *k-Nearest Neighbor (KNN)*, *Naive Bayes* e *Decision Tree*. Para análise LIWC, o melhor classificador foi o *SVM* com 97% de precisão. Já para o modelo de interação a melhor precisão foi com *Naive Bayes*, 82%.

Construir um novo conjunto de dados e realizar experimentos deste em arquiteturas de *Deep Learning* é o objetivo de [Subramani et al. 2019]. Para o experimento

[4]Movimento para conscientização contra assédio sexual.

coletaram dados do Facebook fazendo uso das palavras violência doméstica e abuso doméstico. Dentre os dados colhidos, 3000 foram escolhidos aleatoriamente e anotados manualmente. Os que possuíam imagens ou links foram excluídos. Como resultado obtiveram 1654 postagens catagorizadas em: conscientização, empatia, angariação de fundos, história pessoal e geral. A extração de recursos foi realizada através de *word embeddings*. As técnicas de *Deep Learning* utilizadas foram *Convolutional Neural Networks (CNNs)*, *Recurrent Neural Networks (RNNs)*, *Long Short-Term Memory networks (LSTMs)*, *Gated Recurrent Units (GRUs)* e *Bidirectional LSTMs (BLSTMs)*. O desempenho dos modelos foram comparados com abordagens comuns de *Machine Learning*, sendo estas *Support Vector Machine (SVM)*, *Random Forest (RF)*, *Logistic Regression (LR)* e *Decision Trees (DT)*. Os modelos de *Deep Learning* com *word embeddings* obtiveram desempenho superior em comparação com as abordagens tradicionais de Machine Learning (exceto *RNNs*). As *BLSTMs* mostraram-se vantajosas para identificação de história pessoal. Já as *GRUs* na captação de recursos e conscientização.

Em [Kaur e Sharma 2020] é apresentado um modelo com uso de análise de sentimento e aprendizado de máquina para verificar a popularidade das *#Women* e *#Metoo*. O objetivo era verificar a opinião das pessoas sobre as questões sociais das mulheres. Os dados foram obtidos do Twitter com a *hashtag Women* e *Metoo* e pré-processados para a remoção de ruído, fazendo uso da biblioteca NLTK. Para verificar as palavras com maior frequência, foi realizada uma nuvem de palavras dos tweets após a limpeza. As postagens foram classificadas em positivo, negativo e neutro através da ferramenta TextBlob. As técnicas de classificação utilizadas foram *Naive Bayes*, *Support Vector Machine*, *Logistic Regression* e *Random Forest*. Como resultado, concluíram que a *hashtag* mais popular é *Women* e está mais relacionada ao sentimento positivo. A *Metoo* ao sentimento neutro. O melhor classificador foi *SVM* com precisão de 92,42% e 95,62%, para *#Metoo* e *#Women*, respectivamente.

CAPÍTULO 3

Fudamentação

3.1 Violência digital contra mulher

De acordo [Pedrosa e Zanello 2016], a violência contra mulher, consiste em qualquer ato, baseado em gênero, que cause danos ou sofrimento físicos e/ou psicológicos e/ou sexuais. Segundo [Conselho Nacional De Justiça], é qualquer conduta, ação ou omissão de discriminação, agressão ou coerção, gerada devido à vítima ser mulher e que cause dano, morte, constrangimento, limitação, sofrimento físico, sexual, moral, psicológico, social, político ou econômico ou perda patrimonial. Segundo [Lei Nº 11.340 2006, Enfrentando a Violência On-line Contra Adolescentes 2020], trata-se de qualquer ato violento cometido contra a mulher devido esta ser do sexo feminino, podendo manifestar-se de forma piscológica, moral, patrimonial, física ou sexual. Logo, as formas de manifestações podem ser diversas, como, por exemplo, espancamento, manipulação, estupro, assédio, xingamentos, tortura e ameaças.

Uma das formas desta violência ser manifestada, é através do meio digital. Conforme [What is Digital Gender Violence? 2021], esta caracteriza-se por atos cometidos instigados ou agravados, em parte ou totalmente, pelo uso de Tecnologias da Informação e Comunicação, plataformas de mídia social e e-mail. De acordo com [Enfrentando a Violência On-line Contra Adolescentes 2020, What is Digital Gender Violence? 2021], esse tipo de violência pode resultar em problemas físicos e emocionais, como: depressão, automutilação, afastamento da vida social, suicídio, ansiedade, estresse, cansaço, perda ou excesso de apetite e autocensura. A Tabela 3.1 apresenta os tipos mais comuns de violência de acordo com [What is Digital Gender Violence? 2021, CODING RIGHTS; INTERNETLAB São Paulo, 2017].

Infelizmente, a deteção de violência digital não é simples. De acordo com [CODING RIGHTS; INTERNETLAB São Paulo, 2017], a dificuldade da detecção de violência de gênero online se dá devido ao reconhecimento de fato da violência, ou seja, se tal ato é violento, de atribuição de culpa à vítima e diminuição da violência psicológica.

Tipos de Violência Digital	
Tipos	Detalhes
Disseminação não consentida de imagens íntimas.	Humilhar, expor, constranger alguém por divulgação de imagens íntimas.
Controle e Manipulação de Informação	
Discurso violento	Fazer uso de discurso agressivo; comentários abusivos, xingamentos de cunho sexista; comentários misógionos, transfóbicos, racistas.
Vigilância Eletrônica ou Espionagem Eletrônica	Vigiar as ações e monitorar conversas por meio eletrônico ou plataformas digitais.
Sextorsão	Expor e/ou chantagear e/ou extorquir através do uso de imagens íntimas.
Extorsão	Forçar a agir conforme vontade de outra pessoa, por meio de ameaça e intimidação.
Cyberbullying	Ofender e/ou agredir alguém em ambientes digitais.
Assédio	
Perfil Falso	Quando, através de identidades falsas, são cometidos crimes que prejudicam a vítima ou que resultam em ganhos pessoais pelo falsificador.
Perseguição (cyberstalking)	Ato de perseguir, assediar ou ameaçar a mulher, causando-lhe medo; Interações não solicitadas e/ou abusivas.
Censura ou Controle no ambiente digital	Proibição de uso dos meios digitais; Bloqueio de post, perfis.
Hacking	
Ataques Coordenados	Ataques realizados por uma ou mais pessoas em uma publicação específica ou em uma página.
Ameaça de Violência Física	Ameaças/Intimidação via mensagem privada.

Tabela 3.1: Exemplos de violência digital mais comuns.
Fonte: [What is Digital Gender Violence? 2021, CODING RIGHTS; INTERNETLAB São Paulo, 2017]

3.1.1 Leis para crimes digitais contra mulher

No Brasil, em termo de leis que regulamentam o meio digital, no que diz respeito à violência contra mulher, não existe legislação específica. O que resulta na aplicação de uma ou mais leis que possam melhor envolver o assunto. Apesar da Lei nº 11.340, também conhecida como Lei Maria da Penha, não ser de âmibito digital, a mesma pode

ser aplicada sozinha ou em combinação com outras. Isso se deve ao tipos de violência presentes nos artigos desta. Esta lei é responsável por evitar, enfrentar e punir a violência doméstica e familiar contra a mulher.

Na Tabela 3.2, com base em [CODING RIGHTS; INTERNETLAB São Paulo, 2017], temos punições jurídicas que podem ser aplicadas em determinados casos.

3.2 Aprendizado de Máquina

Machine Learning ou Aprendizado de Máquina, pode ser visto como a junção das áreas de estatística, computação e inteligência artificial [Müller e Guido 2017]. Pois, extrai conhecimento dos dados por meio de técnicas utilizadas conjuntamente, pertencentes às áreas citadas. Para [Ertel 2017], são algoritmos de computadores que melhoraram sua performance conforme adquirem experiência e para [Faceli Ana Carolina Lorena 2011], é o processo de indução de uma hipótese a partir de uma experiência passada.

Aplicações que fazem uso de *Machine Learning*, são cada vez mais comuns e úteis no cotidiano. Podem ser encontradas em aplicativos para predição de qual filme um usuário gostaria de assistir, em sistemas para auxílio no tratamento de câncer, análise de sequenciamento de DNA, busca por novos planetas, definição de quais clientes seriam mais aptos a receber determinados tipos de propagandas conforme suas compras, detecção de fraudes em cartões, reconhecimento de palavras faladas, dentre outros [Müller e Guido 2017, Faceli Ana Carolina Lorena 2011].

Os algoritmos adotados para aprendizagem de máquina podem ser divididos em supervisionados, não supervisionados e de reforço. Os supervisionados necessitam de dados para treinamento que estejam rotulados a priori [Medhat, Hassan e Korashy 2014, Coppin, Ertel 2017]. Uma outra forma de entender este treinamento, seria pensar que jutamente com a entrada do algoritmo, também são fornecidas as saídas esperadas [Müller e Guido 2017]. O não supervionado não faz uso de dados rotulados, estes são capazes de adquirir conhecimento apenas com os dados de entrada fornecidos [Müller e Guido 2017, Coppin]. Para reforço, o aprendizado dar-se a partir de uma série de reforço, seja recompensas ou punições. Neste também não são fornecidos dados de treinamento, [Coppin, Kubat 2017, Ertel 2017]. O modelo pode melhorar as decisões a partir dos reforços fornecidos.

Conforme [Faceli Ana Carolina Lorena 2011], separando os algoritmos de Aprendizado de Máquina de acordo com o paradigma de aprendizado, a tarefa de aprendizado podem ser divididas em: preditivas e descritivas. Para predição, a ideia é encontrar

Leis Aplicáveis		
Tipo	Lei	Detalhes
Censura	Arts 186, 187 e 927 do Código Cível	Manutenção ou liberação de conteúdo, tendo ou não pedido de indenização com danos materiais ou morais
Censura	Lei Maria da Penha	Aplicação de medidas preventivas.
Ofensa / Incitação ao Ódio ou Crime, também conhecido como discurso de ódio ou Hate Speech	Lei n. 7716/1989, também conhecida como Lei Antirracista e o Código Penas Brasileiro. Pode ser visto como crime contra honra ou de incitação ou apologia ao crime	A Lei Antirracista não apresenta diferenciação por gênero e/ou sexualidade.
Ameaças de Violência	Artigo 147 do Código Penas e Lei Maria da Penha	A aplicação da Maria da Penha exige ressalvas visto que exige questões a mais para que seja nela classificada.
Stalking (Perseguição)	Decreto - Lei n. 3688 art. 65 e Lei 11.340/2006	Caso passível de aplicação da Lei 11.340/2006, poderia encaixar em violência psicólogica.
Impersonation	Código Penal, art. 307 Crime de Falsa Identidade	
Exposição de Dados Pessoais	Lei 12737/2021 também conhecida como Lei Carolina Dieckmann	Dependendo pode ser solicitado indenização e/ou aplicado o Marco Civil da Internet.
Utilização Indevida da Imagem	Arts. 186, 187 e 927 do Código Civil (danos morais); Arts 138, 139 e 140 do Código Penal (crimes de honra); difamação ou injúria arts 139 e 140 do Código Penal; Lei Maria da Penha	
Disseminação não Consentida de Imagens	Artigos 138, 139 e 140 do Código Penal; Art, 147 (ameaça); art, 158 (extorsão) e art. 213 (estupro)	
Invasão/Hacking	Lei Carolina Dickmann - Lei 12.737/2021, Código Penal, art.154-A	
Ataque Coordenado	Não há leis em específico que o trate.	

Tabela 3.2: Legislações brasileiras para crimes digitais.

uma função a partir dos dados de treinamento e assim conseguir caracterizar um novo exemplo. Esses algoritmos seguem o paradigma do aprendizado supervisionado. Para descrição, busca-se explorar ou descrever um conjunto. Neste não se tem conhecimento da saída, portanto aprendizado é não supervisionado.

Para desenvolvimento da solução, juntamente com algoritmos de aprendizado de máquina, faremos uso de análise de sentimento. Conforme [Medhat, Hassan e Korashy 2014], trata-se do estudo computacional de opiniões, atitudes e emoções de uma pessoa sobre uma entidade. A mesma apresenta três formas de classificação, que seriam a nível de documento, validando o que é neste expressado de forma negativa ou positiva; a nível de sentença, analisa cada a sentença; e aspecto, onde a classificação é feita de acordo com determinada característica.

A solução proposta para o problema definido neste projeto final de curso, utiliza algoritmos de aprendizado de máquina supervisionados com análise à nível de documento para a realização de classificação de tweets. As classes nas quais os dados serão classificados foram definidas em "contém traços de violência contra a mulher"e "não contém". Na subseção 3.2.1 serão apresentados os classificadores *Naive Bayes (NB)*, *K-Nearest Neighbors (k-NN)*, *Maximum Entropy*, *Random Forest* e *Support Vector Machine*, visto que os mesmos foram escolhidos para compor a solução proposta, separados por classes ao qual pertecem. Na subseção 3.2.2, é esclarecido formas de mensurar a qualidade dos classificadores.

3.2.1 Modelos Preditivos

Conforme citado anteriormente, nos modelos preditivos são empregadas funções que, dada uma entrada, conseguem estimar a saída. Nestes, há a possibilidade de trabalhar com dois domínios: conjunto nominal ou um conjunto infinito e ordenado. Para valores nominais, temos um problema de classificação. Já em caso de um conjunto finito e ordenado, têm-se um problema de regressão [Faceli Ana Carolina Lorena 2011].

Abaixo estão dispostas as decrições sobre os classificadores separados em categorias.

Métodos Baseados em Distâncias

Nos métodos baseados em distância, há a premissa de que dados que possuem informações semelhantes estarão próximos uns dos outros [Faceli Ana Carolina Lorena 2011]. Um algoritmo dessa classe, é o k-NN *(k-Nearest Neighbors)*.

A Figura 3.1 apresenta o algoritmo para o KNN, este determina os pontos de dados do conjunto a partir da comparação do ponto que está sendo analisado com o

ponto mais próximo, ou seja, o vizinho próximo. Logo, conforme [Müller e Guido 2017], a indentificação do novo ponto de dados é dada a partir do ponto que está mais próximo do novo ponto. O k refere-se à quantidade de vizinhos que serão analisados para chegar a decisão sobre o novo ponto, fazendo com que a decisão da classe pertencente seja dada a partir da classe majoritária entre os vizinhos. A classe majoritária, conforme [Faceli Ana Carolina Lorena 2011], pode ser ilutrada na fórmula 3-1, onde $F(x_t)$ é a hipótese aplicada em x_t e $f(x_k)$ é a função objetivo aplicada ao objeto x_k.

$$F(x_t) \leftarrow moda(f(x_1), f(x_2), ..., f(x_k)) \tag{3-1}$$

Figura 3.1: Algoritmos de k-NN

Algoritmo para *Edit* k-NN: eliminação sequencial

Entrada: Um conjunto de treinamento $\mathbf{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$
Saída: Um conjunto de treinamento $\mathbf{D}' = \{(\mathbf{x}_i, y_i), i = 1, \ldots, m; m < n\}$

para cada *exemplo* $(\mathbf{x}_i, y_i) \in \mathbf{D}$ **faça**
  se $(\mathbf{x}_i, y_i)$ *é corretamente classificado por* $\mathbf{D}\backslash\{(\mathbf{x}_i, y_i)\}$ **então**
    /* **Remove** $(\mathbf{x}_i, y_i)$ **de** $\mathbf{D}$ */ ;
    $\mathbf{D} \leftarrow \mathbf{D}\backslash\{(\mathbf{x}_i, y_i)\}$;
  fim
fim
Retorna: $\mathbf{D}$;

Algoritmo para *Edit* k-NN: inserção sequencial

Entrada: Um conjunto de treinamento $\mathbf{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$
Saída: Um conjunto de treinamento $\mathbf{D}' = \{(\mathbf{x}_i, y_i), i = 1, \ldots, m; m < n\}$

$\mathbf{D}' \leftarrow \{\}$;
para cada *exemplo* $(\mathbf{x}_i, y_i) \in \mathbf{D}$ **faça**
  se $(\mathbf{x}_i, y_i)$ *é incorretamente classificado por* $\square'$ **então**
    /* **Acrescenta** $(\mathbf{x}_i, y_i)$ **a** $\mathbf{D}'$ /* ;
    $\mathbf{D}' \leftarrow \mathbf{D}' \cup \{(\mathbf{x}_i, y_i)\}$;
  fim
fim
Retorna: $\mathbf{D}'$

Fonte: [Faceli Ana Carolina Lorena 2011]

Para [Faceli Ana Carolina Lorena 2011], escolher o número de vizinhos não é uma tarefa fácil e para problemas de classificação, normalmente são utilizados k pequeno e ímpar. Duas estratégias comuns é estimar k por validação cruzada e associar um peso na contribução de cada vizinho.

Algumas características sobre o KNN podem ser citadas, de acordo com [Faceli Ana Carolina Lorena 2011]:

- Trata-se de um algoritimo baseado em memória;
- É de fácil treinamento.

- Devido a forma com que realiza as aproximações, em caso de funções complexas, isto pode ser uma boa característica.
- É aplicável a problemas complexos.
- É um algoritmo incremental.
- O erro do k-NN tende para o erro do *Bayes* ótimo.
- A predição pode ter custo alto, devido ao cálculo das distâncias de seus vizinhos.
- É afetado pela presença de atributos redundantes e irrelevantes.
- A classificação pode ser um processo lento.

Métodos Probabilísticos

Naive Bayes

O algoritmo *Naive Bayes*, é baseado no Teorema de Bayes. Este teorema refere-se que dado um evento A ocorrer em um evento B, não depende apenas da relação entre A e B, mas também de A ocorrer sem B, [Faceli Ana Carolina Lorena 2011], ou seja, qual a probabilidade de um evento ocorrer dado que se tem conhecimento de parte da condição deste ocorrer.

Naive Bayes, possuem esse nome devido a assumir que as variáveis envolvidas no problema, ou seja, os atributos dentro da classe, ocorrerão sempre de forma independente, [Aggarwal 2017]. A equação deste pode ser vista em 3-2, onde $P(y_i|x)$ refere-se a probabilidade de x pertercer a classe y_i.

$$log(P(y_i|x)) \propto log(P(y_i)) + \sum_j log(P(x_j|y_i)) \tag{3-2}$$

Bernoulli e Multinomial, são alguns dos modelos de algoritmos possíveis para implementação deste. Bernoulli é uma rede Bayesina e faz uso de palavras binárias, tendo melhor desempenho com baixa variedade de vocubulários. Já Multinomial, é utilizado quando a frequência que a palavra aparece é importante. Este faz uso de unigramas e tem melhor desempenho em grande variedade de vocabulários, [Bhuta et al. 2014].

Segundo [Müller e Guido 2017], o algoritmo Naive Bayes é bastante eficiente e mais rápido no treinamento que outros classificadores lineares. Sua eficiência está ligada a forma com a qual ele aprende. Este é utilizado apenas para problemas de classificação.

Abaixo a descrição de algoritmo para Naive Bayes adaptado de [Kubat 2017]. Para este classificador, sendo a entrada X = (x_1, ..., x_n) a ser classificada, considera-se:

1. Para cada x_i e classe c_j, calcule a probabilidade $P(x_i \mid c_j)$, como a frequência relativa de x_i entre o conjunto de treinamento que pertence a c_j.
2. Para cada classe c_j, faça:

(a) Calcular a P(c_j) como a frequência relativa da classe no conjunto de treinamento.

(b) Calcule a probabilidade condicional de P(x | c_j) assumindo "`ingenuamente`" independência mútua dos atributos.

3. Escolha a classe com maior valor, sendo a frequência da classe no conjunto vezes a probabilidade condicional de P(x | c_j).

Maximum Entropy

Outro método pertencente a esta classe é o de Entropia Máxima ou *Maximum Entropy*. Este faz uso de estimativa de distruição de probabilidades trazendo a ideia de que quanto menos se sabe sobre os dados, mais uniforme será a distribuição logo, resultando assim, em máxima entropia. As restrições, obtidas dos dados de treinamento, farão com que a não uniformidade seja miníma [Bhuta et al. 2014]. Este classificador satisfaz todas as restrições conhecidas e não faz suposições subjetivas sobre o desconhecido, na previsão da probabilidade de um evento aleatório, gerando assim uma distribuição uniforme, [Yang e Chen 2017]. Devido a não fazer suposições sobre o conjunto de dados, esta pode ter melhor desempenho se as suposições de independência condicional não são atendidas, [Khairnar e Kinikar 2013].

Para uso deste, primeiro deve-se definir as características que serão incluídas e o valor esperado para cada recurso que pode ser usado como restrição. Este valor esperado pode ser utilizado como restrição para geração da entropia máxima, [Bhuta et al. 2014].

A fórmula para cálculo da Entropia Máxima pode ser vista em 3-3, onde o peso da característica f_i é expresso pelo parâmetro correspondente λ_i, resultando na entropia máxima p(y|x), [Yang e Chen 2017].

$$p_\lambda(y|x) = \frac{1}{Z_\lambda(x)} exp(\sum_i \lambda_i f_i(x,y)) \tag{3-3}$$

Métodos Baseados em Otimização

Os métodos baseados em otimização formulam o problema como se fossem de otimização, ou seja, maximizando ou minimizando a função objetivo [Faceli Ana Carolina Lorena 2011]. Uma técnica comum desta classe é a *Support Vector Machine* (SVM).

As SVMs, usam separação de hiperplanos como limite de decisão entre duas classes, [Aggarwal 2017], ou seja, é realizada a criação de um hiperplano ideal com os dados de entrada ou de treinamento e o plano em novos exemplos de curvas de categorias [Das e Behera 2017]. A ideia é encontrar um ou mais limites que separam os dados [Khairnar e Kinikar 2013]. Estas, objetivam maximizar a separação entre objetos de diferentes classes [Faceli Ana Carolina Lorena 2011].

As SVMs podem ser divididas em lineares e não lineares, com margens rígidas ou suaves. Quando com margens rígidas, são impostas restrições na geração do hiperplano, de modo a assegurar que não se tenha dados de treinamento entre os limites de separação das classes, ou seja, os dados para esta devem ser linearmente separáveis. Já as suaves, permitem que hajam dados entre as separação das classes, fazendo com que seja também permitido erro de classificação, por isto o nome de margem suaves. Devido nem sempre ser possível separar os dados linearmente, as SVMs não lineares mapeiam os dados do seu espaço original para um espaço com maior dimensão [Faceli Ana Carolina Lorena 2011].

Geralmente são utilizados para problemas de classificação binárias, porém também pode ser generalizado para casos de multiclasse [Aggarwal 2017]. É eficiente em conjunto de dados médio, onde as características dos dados tem significados semelhantes, requer escalonamentos de dados e é sensível aos parâmetros, [Müller e Guido 2017].

Métodos Baseados em Procura

Metódos baseados em procura são aqueles que buscam a solução em um conjunto de soluções possíveis. Dada uma forma de representar os dados e uma função para avaliar esta, a busca é realizada de acordo com a forma que os dados estão representados. A árvore de decisão é um modelo pertecente a esta classe [Faceli Ana Carolina Lorena 2011]. Um algoritmo para esta modelo é apresentado na Figura 3.2.

Figura 3.2: Algoritmos para contrução de árvores de decisão

Algoritmo para construção de uma árvore de decisão

Entrada: Um conjunto de treinamento $\mathbf{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$
Saída: Árvore de Decisão
1 /* **Função GeraÁrvore(D)** */ ;
2 **se** *critério de parada*(**D**) = *Verdadeiro* **então**
3 **Retorna:** um nó folha rotulado com a constante que minimiza a função perda ;
4 **fim**
5 Escolha o atributo que maximiza o critério de divisão em **D** ;
6 **para cada** *partição dos exemplos* $\mathbf{D}_i$ *baseado nos valores do atributo escolhido* **faça**
7 Induz uma subárvore $Árvore_i = GeraÁrvore(\mathbf{D}_i)$;
8 **fim**
9 **Retorna:** *Árvore* contendo um nó de decisão baseado no atributo escolhido, e descendentes $Árvore_i$;

Fonte: [Faceli Ana Carolina Lorena 2011]

O método *Random Forest* é constituído por árvores de decisão no qual cada árvore é direfente da outra, mesmo que seja de forma mínima [Müller e Guido 2017, Aggarwal 2017]. Estas recebem este nome devido a inserção de aleatoriedade na construção da árvore, onde está pode ser realizada de duas formas, selecionado os pontos

de dados usados e os recursos. O resultado é gerado através da predição de cada árvore presente na floresta. Elas ainda podem serem utilizadas tanto para classificação quanto para regressão, sendo o resultado dado através da realização de uma predição "*soft*"e pela média dos resultados, respecitivamente [Müller e Guido 2017].

3.2.2 Avaliação dos Modelos

A avaliação de modelos de apendizado de máquina pode ser realizada considerando diferentes características tais como acurácia, compreensibilidade do conhecimento extraído, tempo de aprendizado e requisitos de armazenamento [Faceli Ana Carolina Lorena 2011]. Como este projeto dinal de curso trata de um problema de classificação, apenas os métodos para esta classe serão explanados e serão abordadas as medidas de desempenho.

Taxa de erro, acurácia e matriz de confusão são algumas das formas de avaliação de desempenho dos modelos [Faceli Ana Carolina Lorena 2011]. A taxa de erro é equivalente a quantidade de dados que o classificador $\hat{f}$ categorizou erroneamente dentro de um conjunto de dados n. A equação para cálculo das classificações erradas, [Faceli Ana Carolina Lorena 2011] pode ser vista em 3-4, onde x_i é a classe conhecida, y_i a classe predita e $I(y_i \neq \hat{f}(x_i))$ é 1 para verdadeiro e 0 para falso. A acurácia ou taxa de acerto é o complemento da taxa de erro. A equação desta, [Faceli Ana Carolina Lorena 2011] pode ser vista em 3-5. Para erro, valores próximos a 0 são melhores, enquanto que para acurácia, são melhores, valores próximos a 1.

$$err(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i)) \tag{3-4}$$

$$ac(\hat{f}) = 1 - err(\hat{f}) \tag{3-5}$$

A matriz de confusão, onde as linhas são as classes verdadeiras e as colunas as preditas, ilustra o número de classificações corretas e incorretas, apresentado medidas quantitativas sobre a dificuldade de classificação do algoritmo, [Faceli Ana Carolina Lorena 2011]. A figura 3.3, ilustra um exemplo desta. As siglas VP, VN, FP, FN representam verdadeiro positivo, verdadeiro negativo, falso positivo, falso negativo, respectivamente.

Na tabela 3.3, adaptada de [Faceli Ana Carolina Lorena 2011] é possível observar algumas definições de medidas de desempenho juntamente com suas equações para n amostras.

Figura 3.3: Representação da Matriz de confusão para avaliação de modelos.

		Classe predita	
		+	-
Classe verdadeira	+	VP	FN
	-	FP	VN

Equações para medir desempenho		
Tipo	Descrição	Equação
Taxa de erro na classe positiva	Falsos negativos, ou seja, classe positiva classificada como classe negativa.	$err_+(\hat{f}) = \frac{FN}{VP+FN}$
Taxa de erro na classe negativa	Falso positivo, ou seja, itens da classe negativa classificados como positivos.	$err_-(\hat{f}) = \frac{FP}{FP+VN}$
Taxa de erro total	Soma da diagonal secundária dividido pela soma todos elementos	$err(\hat{f}) = \frac{FP+FN}{n}$
Taxa de acerto ou acurácia toal	Soma da diagonal principial dividido pela soma de todos elementos da matriz	$ac(\hat{f}) = \frac{VP+VN}{n}$
Precisão	Quantidade de itens positivos classificados corretamente.	$prec(\hat{f}) = \frac{VP}{VP+FP}$
Sensibilidade ou revocação ou taxa de verdadeiro positivo	Taxa de acerto na classe positiva	$sens(\hat{f}) = rev(\hat{f}) = TVP(\hat{f}) = \frac{VP}{VP+FN}$
Especificidade	Taxa de acerto da classe negativa.	$esp(\hat{f}) = \frac{VN}{VN+FP}$
F_1	Combinação entre a medida de precisão e revocação. Considerando peso 1, pois ambas possuem mesmo grau de importância.	$F_1(\hat{f}) = \frac{2*prec(\hat{f})*rev(\hat{f})}{prec(\hat{f})+rev(\hat{f})}$

Tabela 3.3: Equações para aferir desempenho

CAPÍTULO 4

Experimento

O experimento realizado nesta monografia trata-se da reprodução do estudo contido no artigo [Fauzi 2018], este objetiva verificar se um *ensemble* apresenta melhores resultados para identificação de discurso de ódio, na língua indonésia, do que os classificadores em modo independente. Porém, neste trabalho de conclusão de curso, têm-se o foco na detecção de violência contra a mulher em tweets na língua portuguesa.

Conforme pode ser visualizado na Figura 4.1, o experimento pode ser divido em seis partes: coleta dos dados, treinamento e teste, pré-processamento, a transformação do texto em *bag of words*, a classificação pelos classificadores em separado e a classificação pelo *ensemble*, resultando na categorização dos tweets como contendo ou não traços de violência contra a mulher. As subseções 4.1, 4.2, 4.3 e 4.4, descrevem cada uma das etapas.

4.1 Coleta dos dados

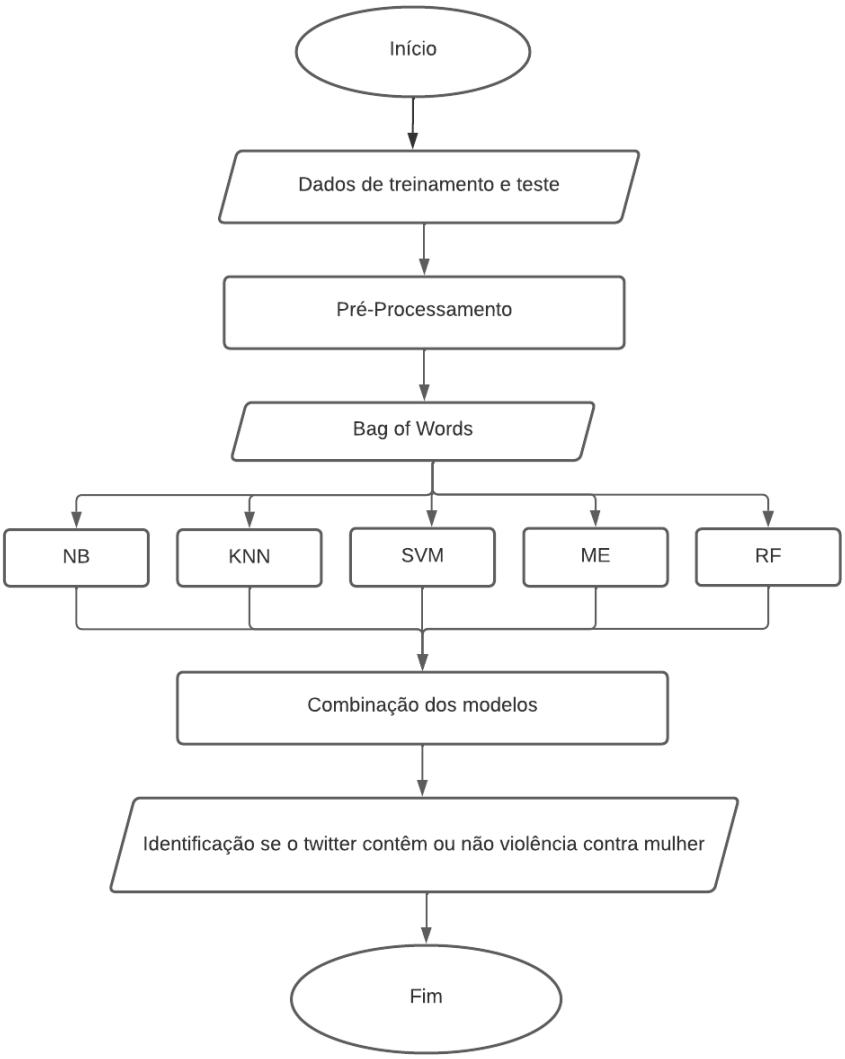
A base de dados utilizada neste exprimento foi coletada e idealizada por [Batista 2021]. Conforme o autor, os tweets foram coletados em tempo real por meio de um filtro de palavras. Estas palavras faziam parte de uma lista constituída através de pesquisas realizadas pelo autor (leitura de testemunhos, artigos, blogs, notícias e aplicação de questionários) sobre o contexto de violência contra a mulher no Brasil.

Os tweets foram coletados através de uma aplicação *crawler* que conectada à API do Twitter capturou e armazenou os dados em um banco PostgreSQL. A coleta foi realizada durante 15 dias, resultando em 7 milhões de tweets.

4.2 Pré-processamento

Após a coleta, devido o interesse do trabalho ser para língua portuguesa, o autor [Batista 2021] filtrou os tweets pela localização do emissor, resultando em 1 milhão de

Figura 4.1: Fluxograma do experimento realizado para identificação de traços de violência contra a mulher em tweets de língua portuguesa. Desenho adaptado de [Fauzi 2018].



tweets. Dentre este, dois mil foram selecionados de forma aleatória para a fase de pré-processamento deste trabalho.

Para esta monografia, o primeiro passo foi a transformação dos dados. Por possuírem caracteres que não estavam em padrão UTF-8, estes foram convertidos para este padrão. Alguns destes não puderam ser corrigidos. Portanto foi realizado a coleta de palavras que possuíam os caracteres com problema a fim de detectar padrões. Com isso, foi possível fazer a conversão para a palavra certa. Quando não foi possível identificar padrões que pudesse corrigir a palavra, os tweets que as continham, foram deletados.

O próximo passo foi reservado à criação de rotinas em Python para: transformar todos os caracteres da mensagem em minúsculo e abrevições em palavras por extenso (veja Tabela 4.1); remoção de padrões mais conhecidos de risadas, acentos, caracteres não alfanuméricos e tweets repetidos. A rotina para remoção de caracteres não alfanuméricos,

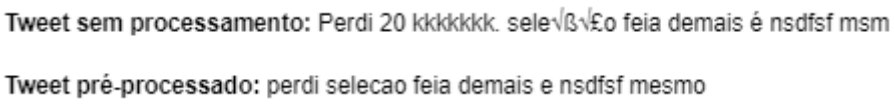


Figura 4.2: Exemplo do pré-processamento realizado em um tweet

removeu também links e arroba das mensagens, deixando apenas os *rt(retweets)* e o nome do perfil de usuário contido nas mensagens. Na Figura 4.2, temos um exemplo de um tweet e este após o processamento.

Por fim, foram aplicadas remoção de *stopwords*. As *stopwords* escolhidas foram obtidas do padrão português junto a biblioteca NLTK[1]. NLTK é um conjunto de bibliotecas da linguagem Python para processamento de linguagem natural. As *stopwords* também tiveram seus acentos retirados através da rotina de retirada de acento. Um trecho das *stopwords* utlizadas podem ser vistas na tabela 4.2.

Mais duas fazes de pré-processamento foram realizadas, tokenização e filtragem. Estas são realizados no Orange[2], durante a execução do modelo. A fase de filtragem retira as *stopwords* do tweet e em tokenização, foi utilizado *word & ponctuation*. Nesta, a frase será quebrada em palavras e as pontuações são mantidas, [Demšar et al. 2013], porém, como não têm-se pontuações, o que importa para este tópico é a criação de tokens.

[1]NLTK - https://www.nltk.org/
[2]https://orangedatamining.com/

Abreviações corrigidas	
Abreviação	Palavra
msm	mesmo
n	nao
q	que
p, pra	para
tbm, tb	tambem
qnd	quando
vc	voce
vcs	voces
hj	hoje
fdp	filha da puta
crl, crlh, krl	caralho
mds	meu deus
tnc	tomar no cu
pqp	puta que pariu
pq	porque
agr	agora
vsf	vai se fuder
pf, pfv	por favor
cmg	comigo
mt	muito
mlk	moleque
c	com
algm	alguem

Tabela 4.1: Abreviações identificadas e corrigidas durante o pré-processamento.

Conjunto de *Stopwords*			
teriam	teriamos	teria	nos
terao	teremos	tera	qual
terei	tiverem	tivermos	elas
tiver	tivessem	tivessemos	pelos
tivesse	tenham	tenhamos	numa
tenha	tiveramos	tivera	minha
tiveram	tivemos	teve	as
tive	tinham	tinhamos	meu
tinha	tem	temos	suas
tem	tenho	seriam	nem
de	a	o	num
que	e	do	essa
da	em	um	voce
para	com	nao	eles
no	os	uma	esse
por	na	se	me
dos	as	mais	nas
ao	mas	como	quem
a	das	ele	seus
ou	sua	seu	aos
nos	muito	quando	mesmo
ja	eu	tambem	sem
pela	pelo	so	depois
ela	isso	ate	entre

Tabela 4.2: *Stopwords*

4.3 Rotulação

Após a fase de pré-processamento dos dois mil tweets colhidos aleatoriamente da base, foram filtrados 1916 para a etapa de rotulação. Isto ocorreu devido à eliminação de dados repetidos. A categorização utilizada foi a separação por contém ou não traços de violência contra a mulher.Nos parágrafos seguintes serão apresentadas algumas observações realizadas sobre os dados.

Durante a rotulação manual dos tweets para a composição da base de treinamento para a etapa seguinte, pôde-se notar algumas categorias nas quais os tweets se enquadravam, abaixo alguns exemplo, tais como:

- Conteúdo de autodepreciação: Tweets contendo baixa autoestima e/ou depressão.
  - `continuo uma vagabunda!`
- Opiniões políticas:
  - `Verdade - ao contrário de todas as manifestações do PT, sempre tem um graninha suja por fora - senão apenas pão com mortadela!`
- Presença de memes:
  - `Vai responder não, puta?`
- Conteúdo adulto/sexual: foram incluído neste tweets que continam palavras que remetiam a contéudo pornográfico
  - `Corno cheio de tesao vendo de perto sua esposa safada sentando no comedor sem camisinha`
- Violência verbal contra a mulher:
  - `vai trabalhar vagabunda`
- Não pertencente a nenhuma das anteriores.
  - `o dia que eu surtei no trabalho pq chamaram a minha garrafinha de água de feia`

Os tweets que continham meme e conteúdo sexual, tornou mais difícil a identificação de violência contra mulher. Para o conteúdo sexual/adulto, apesar das palavras conterem alguma forma de violência contra mulher, sejam ela diminuição, objetificação, por exemplo, foi difícil tomar uma decisão a respeito da sua classe devido a não ficar claro se este tinha a intenção de ser um ato violento contra a mulher ou se era um conteúdo voltado para um público em específico. Já para o meme, deu-se pela dificuldade de identificação de ironia sem se ter um contexto prévio.

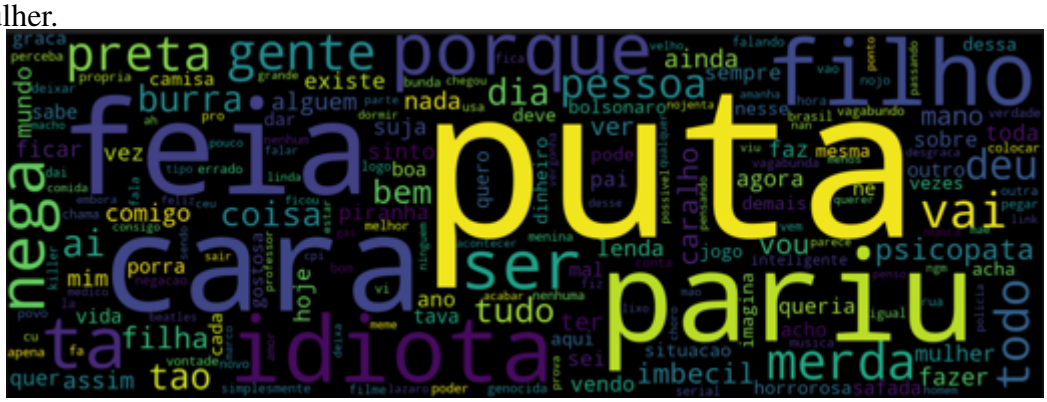
Na época da coleta, era comum um meme com a frase: "vai responder não, puta?". Isto fez com que vários tweets aparececem citando essa referência, o que dificultou saber se era uma citação ao meme ou alguma forma de violência. Porém, de antemão, não era sabido sobre tal questão, até que foi notado devido a grande repetição. Tal questão resultou em uma nova rotulação.

Outras questões como, a presença dos erros de escrita, ironia e não se ter conhecimento sobre o contexto da mensagem, tornou mais difícil o processo de rotulação. As palavras "puta"e "vagabunda"são bastante utilizadas para atos violentos, sendo que a primeira, muita vezes é utilizada como superlativo para determinadas expressões. As Figuras 4.3 e 4.4 apresentam as nuvens de palavras referentes aos dados da fase de rotulação. Nestas já estão consideradas a sepração entre as classes com e sem traço de violência, respectivamente.

A catalogação dos dados resultou em 1514 rotulados como não violência e 402 com violência. Com a finalidade de obter um balanceamento para teste e treinamento, realizou-se subamostragem. Através de seleção aleatória, foram elegidos 402 tweets da categoria não violência, obtendo assim, 804 tweets no total.

Figura 4.3: Nuvem de palavras para base contendo traços de violência contra a mulher.

Figura 4.4: Nuvem de palavras para base que não contém traços de violência contra a mulher.



4.4 Implementação dos modelos classificadores

A implementação do modelo foi realizada através do Orange Data Mining[3]. O software possibilita a implementação de modelos de aprendizado de máquina, pré-processamento, visualização e mineração de dados, entre outras funcionalidades. A implementação é realizada de forma simples, visto que toda a contrução é dada através de blocos, bastando apenas selecionar e configurar o que se deseja utilizar [Demšar et al. 2013].

Para inicialização dos experimentos, foram utilizados os tweets balanceados e não balanceados, na intenção de verificar o impacto na diferença das classes no resultado. Os dados foram separados através do modelo *cross-validation* em 10 partes, onde 9 destas são utilizadas para treinamento e 1 para teste.

Os dados processados são entradas da *Bag of Words*. Está pode ser entendida, como uma representação simplicafica de uma frase, onde cada uma torna-se um "saco"de suas palavras. Devido o uso da tokenização, cada palavra de uma frase será transformada em um token. Apesar de no trabalho utilizado como baseline para este utilizar configuração de Frequência de Termo e Frequência Inversa do Documento (TF/IDF)[Fauzi 2018], neste trabalho também foram realizados testes utilizando apenas a configuração de Frequência de Termo. Conforme [Wendel Melo], TF/IDF combina a frequência de termo com a frequência inversa do documento, onde quanto mais um termo estiver no texto mais ele descreve o conteúdo e quanto maior a raridade de um termo mais importante ele será para descrevê-lo, respectivamente. Portanto, o TF/IDF atribui peso ao termo através da quantidade de vezes que ele aparece na base e de sua raridade.

Os dados foram testados com cada um dos modelos *Naive Bayes, K-Nearest Neighbors, Logistic Regression, Radom Forest* e *Support Vector Machine* de modo independente. À partir da saída desses classificadores, realizou-se a implementação do modo *ensemble*.

No *ensemble*, cada classificador tem uma opinião sobre a classe a qual o twitter pertence. A classe final é escolhida de duas formas, voto suave e voto difícil. Para o voto difícil o resultado é aferido através da escolha da classe que mais se repetiu entre os classificadores. Já no suave, é calculada a média entre os valores retornados para cada classe. A classe que teve maior média é a escolhida para classificar o tweet.

[3]Orange Data Mining - https://orangedatamining.com/

CAPÍTULO 5

Resultados

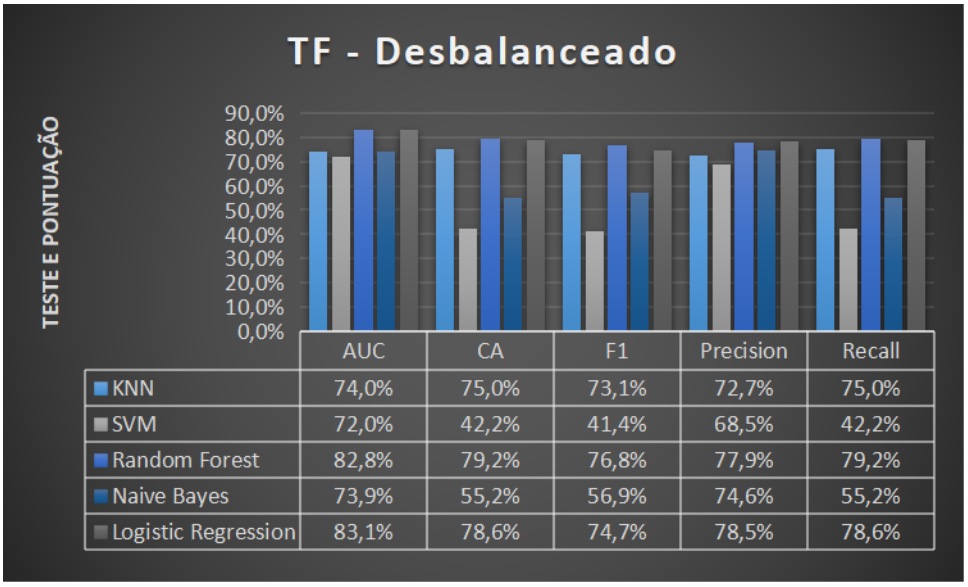
Este capítulo será dividido em duas seções, uma abordará os resultados dos classificadores em modo independente, a outra em modo ensemble, 5.1 e 5.2, respectivamente.

5.1 Classificadores em modo independente

No total, restaram 1916 dados rotulados e processados, sendo destes 402 para a classe contendo traços de violência. Portanto, inicia-se a coleta dos resultados, analisando como os modelos se comportavam para os dados desbalanceados.

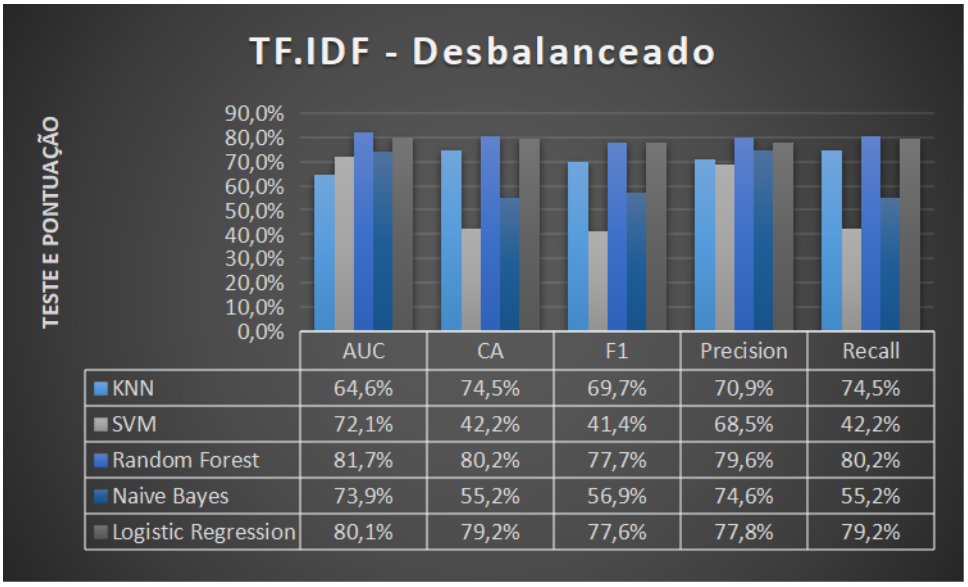
Nas figuras 5.1, 5.2, 5.3, e 5.4 são apresentados os resultados dos classificadores, incluindo as medidas: área embaixo da curva (AUC), acurácia (CA), *F1 score*, precisão (*precision*) e sensibilidade (*recall*). Analisando o *F1 score*, o classificador com melhor resultado foi o *Random Forest*, havendo apenas diferença de pontuação para cada tipo de dados e configuração de *bag of words*. Para TF e TF/IDF, têm-se 76,8% e 77,7% para dados desbalanceados e 81,4 e 80,2% para balanceados. Para TF/IDF, em desbalanceados, a diferença entre *Random Forest* e *Logistic Regression* foi de 0,01%. Os modelos com piores resultados variam de acordo com a configuração e o conjunto de dados. SVM apresentou 41,4% para dados desbalanceados com TF e TF/IDF. Já, para dados balanceados, em TF/IDF o KNN apresentou o pior resultado com 48,3% e, para TF o SVM com 55,5%.

Para auxiliar a tomada de decisão quanto a melhor configuração, também foram geradas matrizes de confusão. As Figuras 5.5 e 5.6, referem-se aos dados balanceados, onde com exceção do *Random Forest* e *Logistic Regression* para configuração TF, todos idenficam mais falso negativos. Porém, a diferença entre *Logistic Regression*, em TF.IDF, e KNN, em TF, é igual a uma classificação. Para os dados desbalanceados, observa-se que independente da configuração, KNN, *Logistic Regression* e *Random Forest*, identificam mais falsos positivos, levando a um *overfit* dos dados. Já SVM e *Naive*, possuem mesma quantidade de falso positivo e negativo para as duas configurações de *bag of word*, porém prevalecendo uma maior ocorrência de falso negativo.



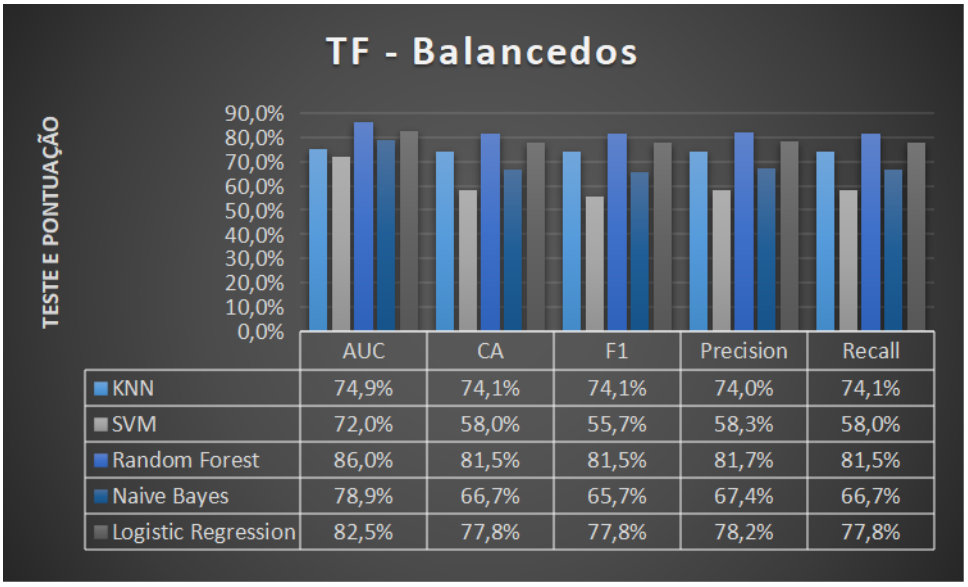
	AUC	CA	F1	Precision	Recall
KNN	74,0%	75,0%	73,1%	72,7%	75,0%
SVM	72,0%	42,2%	41,4%	68,5%	42,2%
Random Forest	82,8%	79,2%	76,8%	77,9%	79,2%
Naive Bayes	73,9%	55,2%	56,9%	74,6%	55,2%
Logistic Regression	83,1%	78,6%	74,7%	78,5%	78,6%

Figura 5.1: Resultados classificadores - TF - Desbalanceados



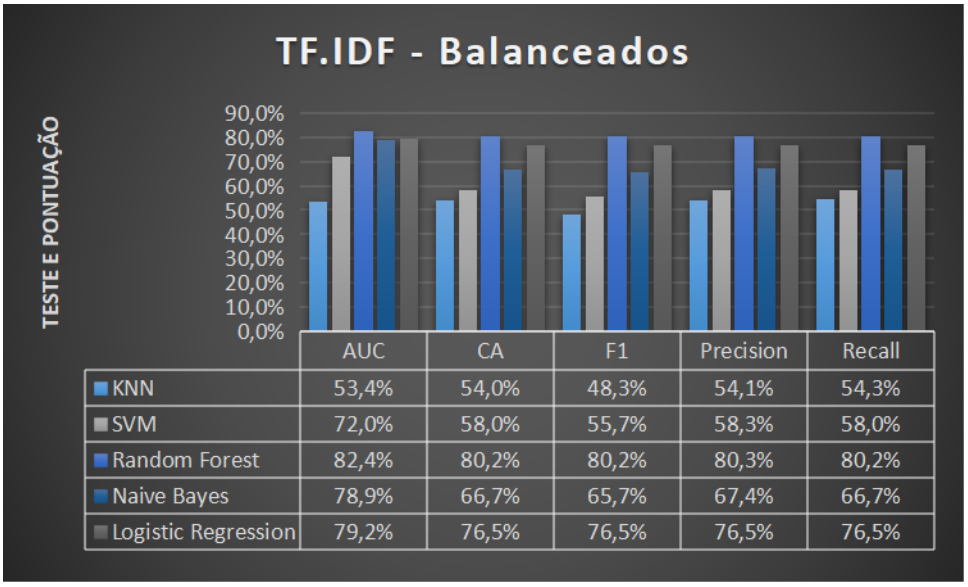
	AUC	CA	F1	Precision	Recall
KNN	64,6%	74,5%	69,7%	70,9%	74,5%
SVM	72,1%	42,2%	41,4%	68,5%	42,2%
Random Forest	81,7%	80,2%	77,7%	79,6%	80,2%
Naive Bayes	73,9%	55,2%	56,9%	74,6%	55,2%
Logistic Regression	80,1%	79,2%	77,6%	77,8%	79,2%

Figura 5.2: Resultados classificadores - TF.IDF - Desbalanceados



	AUC	CA	F1	Precision	Recall
KNN	74,9%	74,1%	74,1%	74,0%	74,1%
SVM	72,0%	58,0%	55,7%	58,3%	58,0%
Random Forest	86,0%	81,5%	81,5%	81,7%	81,5%
Naive Bayes	78,9%	66,7%	65,7%	67,4%	66,7%
Logistic Regression	82,5%	77,8%	77,8%	78,2%	77,8%

Figura 5.3: Resultados classificadores - TF - Balanceados



	AUC	CA	F1	Precision	Recall
KNN	53,4%	54,0%	48,3%	54,1%	54,3%
SVM	72,0%	58,0%	55,7%	58,3%	58,0%
Random Forest	82,4%	80,2%	80,2%	80,3%	80,2%
Naive Bayes	78,9%	66,7%	65,7%	67,4%	66,7%
Logistic Regression	79,2%	76,5%	76,5%	76,5%	76,5%

Figura 5.4: Resultados classificadores - TF.IDF - Balanceados

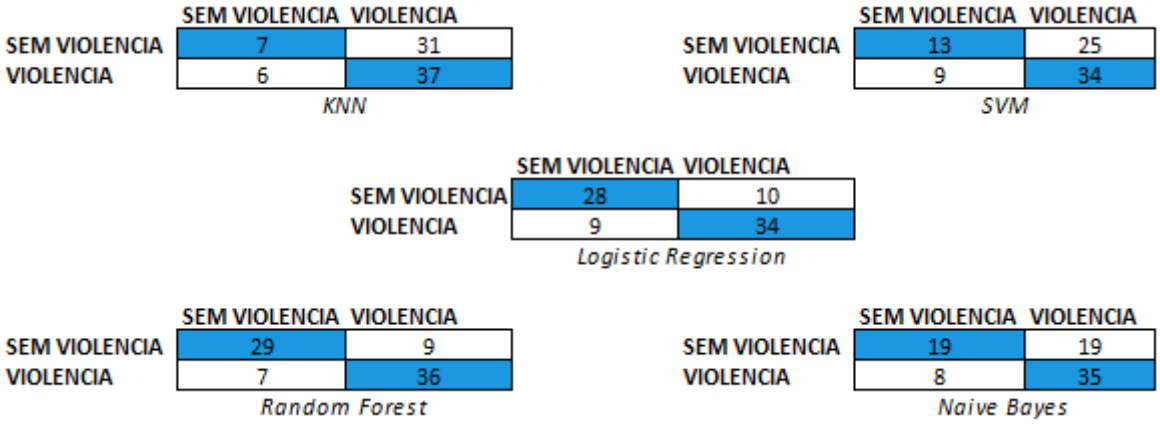


Figura 5.5: Matriz de Confusão - TF.IDF - Balanceados

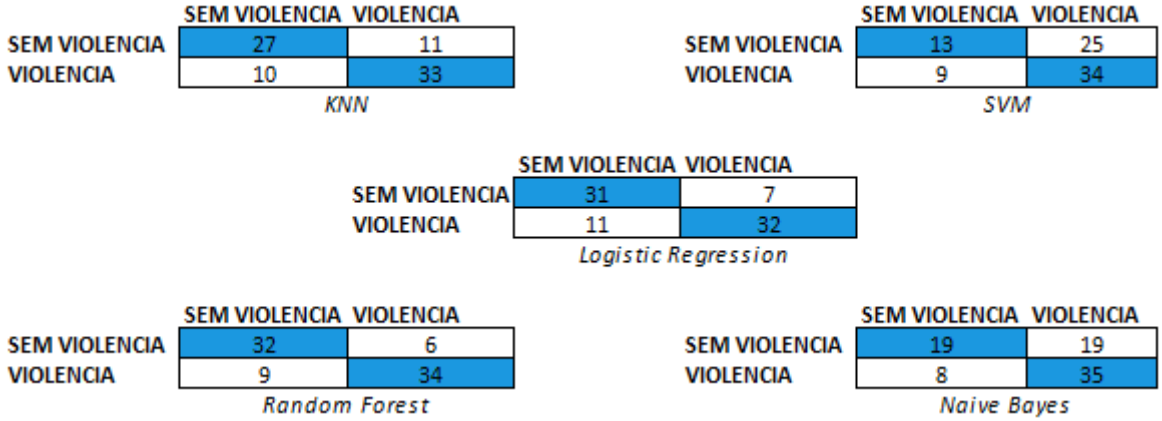


Figura 5.6: Matriz de Confusão - TF - Balanceados

Em geral, nota-se uma identificação de não violência como violência, onde foi concluído que tal fator ocorre devido a dificuldade de separação dos dados, o que pode ser reforçado na visualização da nuvem de palavras. Por exemplo, a palavra "puta"sozinha é violência, porém "puta que pariu"é uma expressão, logo com a criação dos *tokens*, ao identificar determinado tipo de ocorrência, a classificação será realizada erroneamente. Tal fator dificulta a classificação dos algoritmos. No caso do SVM por exemplo, dificultará a criação do hiperplano, pois não há boa separação dos dados. Para o *Naive Bayes*, será assumida indepêndencia, devido a característica do *Naive*, mesmo que esta não exista de fato.

Visto que os dados balanceados com configuração termo de frequência, e desbalanceados com TF/IDF apresentam melhores resultados, foi realizada uma comparação entre seus resultados. De forma geral, os dados balanceados com termo de frequência

apresentaram melhores resultados de acurácia, F1 e área abaixo da curva AUC, e a diferença entre falsos positivos e negativos não são grandes. Os maiores valores são apresentandos por *Naive Bayes* e SVM na identificação de sem violência como violência. Portanto, apesar do artigo base seguir a configuração TF/IDF, devido ao resultados obtidos aqui, o modo *ensemble* foi configurado com TF.

5.2 Classificadores em modo *Ensemble*

Em modo *ensemble*, assim como no artigo adotado como *baseline* [Fauzi 2018], foram realizados dois experimentos: um contendo todos os classificadores testados (*ensemble* completo); o outro, à partir dos resultados obtidos da seção anterior, contendo os três que apresentaram o melhor resultado (*ensemble* de 3), sendo *Logistic Regression, Random Forest* e KNN.

Os três modelos de melhores resultados *Logistic Regression, Random Forest* e KNN, apresentaram, respectivamente, 82,5%, 86,0% e 74,9% para área embaixa da curva; 77,8%, 81,5% e 74,1% de acurácia; 77,8%, 81,5% e 74,1% para F1; *precision* de 78,2%, 81,7% e 74,0%, e *recall* de 77,8%, 81,5% e 74,1%. Quanto suas matrizes de confusão, *Random Forest* e *Logistic Regression* apresentam maior identificação de falso negativo, quanto KNN apresenta mais falso positivo.

A decisão da classe pertecente pelo *Ensemble*, é realizada através de duas votações, suave e difícil. Para difícil, a classe escolhida será dada pela classe que foi mais identificadas pelos classificadores. A suave é dada pela média gerada durante a escolha para cada classe pelo classificador. Essas médias são somadas, sendo escolhida a classe que obteve maior valor para a média.

As matrizes de confusão para os modelos *Ensemble* implementados tanto para votação suave quanto para difícil estão disponíveis nas Figuras 5.7 e 5.8.

	SEM VIOLENCIA	VIOLENCIA
SEM VIOLENCIA	25	13
VIOLENCIA	7	36

Votação Suave

	SEM VIOLENCIA	VIOLENCIA
SEM VIOLENCIA	30	8
VIOLENCIA	8	35

Votação difícil

Figura 5.7: Matriz de confusão para o *ensemble* completo.

O desempenho dos *ensembles*, tanto completo quanto de 3, pode ser observado na figura 5.9. São apresentados os valores de medidas para *F1-score, precision, recall* e *CA* (Tabela 3.3).

Analisando os resultado do modelo *ensemble* completo, observa-se que o melhor resultado foi obtido por votação difícil, tendo também melhores resultados de falso

	SEM VIOLENCIA	VIOLENCIA
SEM VIOLENCIA	25	13
VIOLENCIA	6	37

Votação Suave

	SEM VIOLENCIA	VIOLENCIA
SEM VIOLENCIA	29	9
VIOLENCIA	8	35

Votação difícil

Figura 5.8: Matriz de confusão para o *ensemble* de 3.

negativo e positivo. A votação suave apresenta maior identificação de falso negativo. Tal fator pode ter ocorrido devido a forma que a votação é feita, visto que a maioria dos modelos possuem maior ocorrência de falso negativo o que influencia diretamente na média geradas de cada classe. Votação difícil alcançou 78,9% de F1 contra 71,4%.

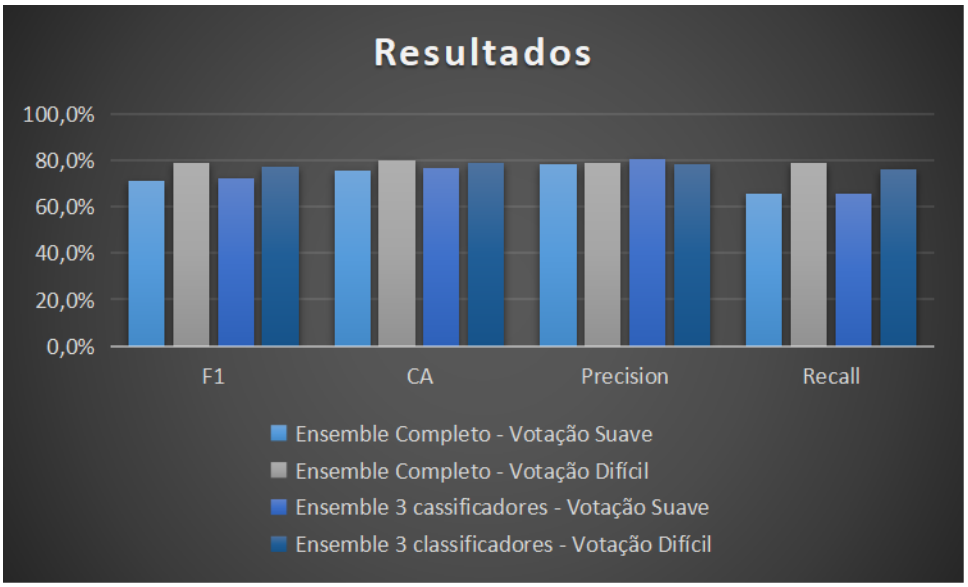


Figura 5.9: Resultado geral dos modelos *ensembles*.

Já, para o *ensemble* de 3, como os classificadores de melhor resultado, *Random Forest*, *Logistic Regression* e KNN, temos novamente melhores resultados para votação difícil, porém neste caso, ambos identificaram maior ocorrência de faso negativo, porém para votação difícil, a diferença é de uma classificação. Votação suave apresentou melhores valores para precisão, com 80,6% contra 78,4%, entretanto apresentou baixo valor de *recall*, o que fez com que votação difícil tivesse melhores resultados. Portanto, novamente obteve-se as melhores pontuações para votação difícil, com 77,3% contra 72,5%.

Finalmente, comparando os dois *ensembles*, têm que o completo apresenta melhores resultados, porém não chega a ter se quer uma diferença de 2% entre o *F1 score* dos dois modelos. Novamente, o fator resultante na diferença foi o *recall*. Ao comparar as matrizes, é persecptível que os valores entre ambos divergem em uma classificação.

CAPÍTULO 6

Conclusão

Conforme pôde ser visto através da nuvem de palavras contida na Figura 4.3, podemos concluir que as palavras "piranha", "puta", "burra", "vagabunda", "vadia"e "feia"são bastante utilizadas para se referir a mulheres em atos de violência. Porém, estes termos também são visualizados com grande frequência na nuvem de palavras, Figura 4.4, para a base rotulada como não contendo traços de violência contra a mulher. Essa situação dificultou a classificação por parte dos modelos, aumentando assim a quantidade de falsos positivos e negativos. Acredita-se que tal questão deu-se devido a quantidade de autodepreciação na base. Com autodepreciação refere-se a quantidade de tweets onde usuário diminui a si mesmo, suas capacidades ou ações, levando a conclusão de que a base possui mais tweets relacionado a depressão do quê a violência contra mulher.

Observa-se que apesar do *ensemble* não ter tido resultado superior ao melhor classificador, *Random Forest*, ressalta-se a proximidade de resultados, assim como, quando comparado ao F1 dos outros modelos, todos são superados pelo *ensemble* de votação difícil.

Através das figuras apresentadas no capítulo de resultados foi possível perceber que determinada configuração de *bag of words* pode influenciar em melhores resultados para os classificadores, impactanto assim, na quantidade de falsos positivos e negativos. Outro fator determinante é o balanceamento ou não dos dados, devido o *overfit* gerado para último.

Em comparação dos resultados obtidos neste experimento com os do trabalho utilizado como base, nota-se divergência entre modelos com os melhores resultados. No [Fauzi 2018], com dados desbalancedos, o melhores resultados foram obtidos através de *Naive Bayes* e SVM, com 78,3% e 78,1%. Neste, nas duas configurações de *bag of words* utilizadas, os dois modelos foram os que obtiveram piores resultados. Em ambos os trabalhos, o *ensemble* não superou o classificador de melhor resultado.

Outra diferença entre os trabalhos é que para o autor, votação suave traz melhores resultados gerais, enquanto que neste foi a votação difícil. Outra questão é que os três melhores classificadores no [Fauzi 2018] foram *Naive Bayes*, SVM e *Random Forest*. Neste foram KNN, *Random Forest* e *Logistic Regression*. Para o autor de [Fauzi 2018],

o melhor resultado para o *ensemble* foi encontrado fazendo uso dos três melhores classificadores. Já, para este projeto final de curso o melhor foi o *ensemble* completo.

Em ambos os trabalhos, os dados balanceados tiveram melhores resultados e apesar de uma diferença não muito grande entre classificadores e os resultados do *ensemble*, nota-se que o metódo traz melhorias para seleção da classe.

Para fins de melhoraramento dos resultados obtidos com este experimento, como trabalhos futuros, sugere-se aumento dos dados de treinamento, como tentativa de suavizar o fato das mesmas palavras com violência serem identificadas como não violência, e auxílio de um especialista para melhorar a qualidade da rotulação. Outro fator que pode trazer diferenças nos resultados da classificação é fazer o uso de *n-grams*, visto que há dependências entre as palavras.

Outra oportunidade para trabalhos futuros seria a identificação de tweets com conteúdo sexual/adulto, pois conforme pôde ser observado visualmente pelo pesquisador, estas mensagens quase sempre são acompanhadas de traços de violência contra mulher.

Referências Bibliográficas

[AASP]AASP. *Crimes sexuais pela internet: a violência contra a mulher entre o real e o virtual*. Disponível em: <https://www.aasp.org.br/noticias/crimes-sexuais-pela-internet-violencia-contra-mulher-entre-o-real-e-o-virtual/>.

[Aggarwal 2017]AGGARWAL, C. C. *Data Mining: The Textbook*. [S.l.]: Springer, 2017.

[Batista 2021]BATISTA, P. *Detecção de Traços de Exposição à Violência contra mulher no Twitter.* 2021. Monografia (Bacharel em Sistemas de Informação), UFG (Universidade Federal de Goiás), Goiânia, Goiás, Brazil.

[Bhuta et al. 2014]BHUTA, S. et al. 2014.

[Cetic]Cetic. *TIC DOMICÍLIOS 2020*. Disponível em: <https://cetic.br/media/analises/tic$_{d}$*omicilios*$_{2}$020$_{c}$*oletiva*$_{i}$*mprensa.pdf*>.

[CODING RIGHTS; INTERNETLAB São Paulo, 2017]CODING RIGHTS; INTERNETLAB. *Violências contra mulher na internet: diagnóstico, soluções e desafios. Contribuição conjunta do Brasil para a relatora especial da ONU sobre violência contra a mulher*. São Paulo, 2017. Disponível em: <https://www.internetlab.org.br/wp-content/uploads/2017/11/Relatorio$_{V}$*iolenciaGenero*$_{O}$*NU.pdf*>.

[Conselho Nacional De Justiça]Conselho Nacional De Justiça. *Forma de violência contra a mulher*. Disponível em: <https://www.cnj.jus.br/programas-e-acoes/violencia-contra-a-mulher/formas-de-violencia-contra-a-mulher/>.

[Coppin]COPPIN, B. *Artificial Intelligence Illuminated*. [S.l.]: Jones and Bartlett Publishers.

[Corrêa e Faria 2021]CORRêA, I. T.; FARIA, E. R. An analysis of violence against women based on victims' reports. In: *XVII Brazilian Symposium on Information Systems*. New York, NY, USA: Association for Computing Machinery, 2021. (SBSI 2021). ISBN 9781450384919. Disponível em: <https://doi.org/10.1145/3466933.3466968>.

[Das e Behera 2017]DAS, K.; BEHERA, R. N. A survey on machine learning: concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering*, v. 5, n. 2, p. 1301–1309, 2017.

[DATAREPORTAL]DATAREPORTAL. *DIGITAL 2021: BRAZIL*. Disponível em: <https://datareportal.com/reports/digital-2021-brazil>.

[Demšar et al. 2013]DEMsAR, J. et al. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, v. 14, p. 2349–2353, 2013. Disponível em: <http://jmlr.org/papers/v14/demsar13a.html>.

[Enfrentando a Violência On-line Contra Adolescentes 2020]ENFRENTANDO a Violência On-line Contra Adolescentes: No contexo da pandemia de covid-19. [S.l.], 2020. Disponível em: <https://www.gov.br/mdh/pt-br/navegue-por-temas/politicas-para-mulheres/publicacoes-1/68ENFENTANDO$_V$*IOLENCIA*$_O$*NLINE*.*pdf*>.

[Ertel 2017]ERTEL, W. *Introduction to Artificial Intelligence*. 2. ed. [S.l.]: Springer, 2017.

[Faceli Ana Carolina Lorena 2011]FACELI ANA CAROLINA LORENA, J. G. A. C. P. d. L. F. d. C. K. *Inteligência Artificial: Uma Abordagem de Apredizado de Máquina*. [S.l.]: LTC, 2011.

[Fauzi 2018]FAUZI, A. Y. M. A. Ensemble method for indonesian twitter hate speech detection. *Indonesian Journal of Electrical Engineering and Computer Science*, v. 11, n. 1, p. 294–299, 2018.

[González, Gabarrot e Cantu-Ortiz 2020]GONZáLEZ, G. A. R.; GABARROT, M.; CANTU-ORTIZ, F. J. Understanding violence against women in digital space from a data science perspective : Full/regular research papers - csci-isna. In: *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. [S.l.: s.n.], 2020. p. 263–269.

[Institute]INSTITUTE, J. W. B. *Fighting Gender-Based Violence in Brazil*. Disponível em: <https://www.wilsoncenter.org/article/fighting-gender-based-violence-brazil: :text=Brazil%20consistently%20ranks%20as%20one,global%20rates%20for%20the%20year.>.

[Instituto Patrícia Galvão]Instituto Patrícia Galvão. *VIOLÊNCIA DE GÊNERO NA INTERNET*. Disponível em: <https://dossies.agenciapatriciagalvao.org.br/violencia/violencias/violencia-de-genero-na-internet/>.

[Kaur e Sharma 2020]KAUR, C.; SHARMA, A. Social issues sentiment analysis using python. In: *2020 5th International Conference on Computing, Communication and Security (ICCCS)*. [S.l.: s.n.], 2020. p. 1–6.

[Khairnar e Kinikar 2013]KHAIRNAR, J.; KINIKAR, M. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, Citeseer, v. 3, n. 6, p. 1–6, 2013.

[Khatua, Cambria e Khatua 2018]KHATUA, A.; CAMBRIA, E.; KHATUA, A. Sounds of silence breakers: Exploring sexual violence on twitter. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.: s.n.], 2018. p. 397–400.

[Kubat 2017]KUBAT, M. *An Introduction to Machine Learning*. 2. ed. [S.l.]: Springer, 2017.

[Kumar e Aggarwal 2019]KUMAR, D.; AGGARWAL, S. Analysis of women safety in indian cities using machine learning on tweets. In: *2019 Amity International Conference on Artificial Intelligence (AICAI)*. [S.l.: s.n.], 2019. p. 159–162.

[Lei Nº 11.340 2006]LEI Nº 11.340. [S.l.], 2006. Disponível em: <https://legislacao.presidencia.gov.br/atos/?tipo=LEInumero=11340ano=2006ato=4b0gXTU5kMRpWT5c7>.

[Mansuido]MANSUIDO, M. *Violência de gênero na internet: o que é e como se defender*. Disponível em: <https://www.gov.br/pt-br/noticias/assistencia-social/2021/03/canais-registram-mais-de-105-mil-denuncias-de-violencia-contra-mulher-em-2020>.

[Medhat, Hassan e Korashy 2014]MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, v. 5, n. 4, p. 1093–1113, 2014. ISSN 2090-4479. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2090447914000550>.

[Ministério da Mulher, da Família e dos Direitos Humanos]Ministério da Mulher, da Família e dos Direitos Humanos. *Canais registram mais de 105 mil denúncias de violência contra mulher em 2020*. Disponível em: <https://www.gov.br/pt-br/noticias/assistencia-social/2021/03/canais-registram-mais-de-105-mil-denuncias-de-violencia-contra-mulher-em-2020>.

[Müller e Guido 2017]MüLLER, A. C.; GUIDO, S. *Introduction to Machine Learning with Python*. 1. ed. [S.l.]: O'Reilly Media, 2017.

[Nogueira]NOGUEIRA, L. d. R. *MÍDIAS SOCIAIS: UMA NOVA PORTA DE ENTRADA PARA A VIOLÊNCIA CONTRA A MULHER*. Disponível em: <http://ihs.sites.uff.br/wp-content/uploads/sites/47/2019/08/MIDIAS-SOCIAIS-porta-de-entrada-para-violencia-contra-mulher-de-LucianaRezende.pdf>.

[Pedrosa e Zanello 2016]PEDROSA, M.; ZANELLO, V. (in)visibilidade da violência contra as mulheres na saúde mental. *Psicologia: Teoria e Pesquisa*, Instituto de Psicologia, Universidade de Brasília, v. 32, n. spe, 2016. ISSN 1806-3446.

[Rodrigues, Júnior e Lobato 2019]RODRIGUES, L.; JúNIOR, J. da S.; LOBATO, F. A culpa é dela! É isso o que dizem nos comentários das notícias sobre a tentativa de feminicídio de Elaine Caparroz. In: *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*. Porto Alegre, RS, Brasil: SBC, 2019. p. 47–58. ISSN 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/6547>.

[Saha et al. 2018]SAHA, P. et al. Hateminers : Detecting hate speech against women. *CoRR*, abs/1812.06700, 2018. Disponível em: <http://arxiv.org/abs/1812.06700>.

[Sahi, Kilic e Saglam 2018]SAHI, H.; KILIC, Y.; SAGLAM, R. Automated detection of hate speech towards woman on twitter. In: *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. [S.l.: s.n.], 2018. p. 533–536.

[Soldevilla e Flores 2021]SOLDEVILLA, I.; FLORES, N. Natural language processing through bert for identifying gender-based violence messages on social media. In: *2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE)*. [S.l.: s.n.], 2021. p. 204–208.

[Subramani et al. 2019]SUBRAMANI, S. et al. Deep learning for multi-class identification from domestic violence online posts. *IEEE Access*, v. 7, p. 46210–46224, 2019.

[Subramani, Vu e Wang 2017]SUBRAMANI, S.; VU, H. Q.; WANG, H. Intent classification using feature sets for domestic violence discourse on social media. In: *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*. [S.l.: s.n.], 2017. p. 129–136.

[Wendel Melo]Wendel Melo. *Faculdade de Computação - Universidade Federal de Uberlândia*. Disponível em: <http://www.facom.ufu.br/ wendelmelo/ori201901/4$_p$*onderacao*$_d$*e*$_t$*ermos.pdf*>.

[What is Digital Gender Violence? 2021]WHAT is Digital Gender Violence? [S.l.], 2021. Disponível em: <https://violenciadigital.tedic.org/indexEng.html#>.

[Xue, Chen e Gelles 2019]XUE, J.; CHEN, J.; GELLES, R. Using data mining techniques to examine domestic violence topics on twitter. *Violence and gender*, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 6, n. 2, p. 105–114, 2019.

[Yang e Chen 2017]YANG, P.; CHEN, Y. A survey on sentiment analysis by using machine learning methods. In: *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. [S.l.: s.n.], 2017. p. 117–121.