Identificando e categorizando linguagem ofensiva em redes sociais

Cardeque Henrique B. A. Borges[1], **Nádia F. Felix**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)

Cardequeh@gmail.com, nadia@inf.ufg.br

Abstract. *The presence of texts with hate speech on social networks becomes more evident every day. The manual removal of such texts becomes unfeasible due to the volume of publications made daily. Machine learning techniques can be used to automate the detection and removal of such texts. This paper aims to identify and categorize the presence of hate speech in texts on the social network Twitter, using the classifiers Random Forest, SVM Linear, RBF e Sigmoid, Multinomial Naive Bayes and Decision Tree. An F1 score of 0.71 was achieved in the hate speech identification task and an F1 score of 0.54 and 0.49 for the hate speech categorization tasks.*

Resumo. *A presença de textos que apresentam discursos de ódio em redes sociais se torna mais evidente a cada dia. A remoção manual de tais textos passa a ser inviável devido ao volume de publicações feitas diariamente. Técnicas de aprendizado de máquina podem ser utilizadas com o objetivo de automatizar a detecção e remoção de tais textos. Esse trabalho tem como objetivo identificar e caracterizar a presença de discurso de ódio em textos da rede social Twitter, utilizando os classificadores Random Forest, SVM linear, RBF e Sigmoid, Naive Bayes multinomial e Árvore de Decisão. Foi alcançado um F1 score de 0.71 na tarefa de identificação de discurso de ódio e um F1 score 0.54 e 0.49 nas tarefas de categorização do discurso de ódio.*

1. Introdução

O número crescente de usuários nas redes sociais e a sensação de anonimidade causada por elas, promovem o aumento de discursos de ódio, atitudes ofensivas e perigosas, presença de linguagem agressiva e o *cyberbullying*. Atitudes, como essas, podem afetar a vida de pessoas, podendo ocasionar traumas, transtornos mentais e, em casos extremos até o suicídio.

Em 2015, a *Intel Security*[1] realizou uma pesquisa no Brasil, entrevistando 507 crianças e adolescentes. No levantamento, 66% dos entrevistados relataram ter presenciado situações de agressões nas redes sociais. Outros 21% afirmaram que já vivenciaram episódios de *cyberbullying*.

Uma das principais formas de *cyberbullying* é através de ofensas e agressões verbais em mídias sociais. A filtragem manual destes tipos de textos, dentre todos presentes nestes ambientes, se torna inviável, devido a quantidade de postagens sendo publicadas

[1]https://www.intel.com/content/www/us/en/security/hardware/hardware-security-overview.html

todos os dias. O uso de ferramentas para automatizar a filtragem de textos ofensivos é de extrema importância e já vêm sendo estudado por redes sociais como o Facebook[2], Youtube[3] e o Twitter[4].

A tarefa de automatizar a detecção de discurso de ódio, entretanto, tem como desafio o frequente problema em definir o que é considerado um discurso de ódio [Waseem et al. 2017]. Diferentes definições são utilizadas por vários pesquisadores. [Davidson et al. 2017], por exemplo, enfatizam que as palavras que possuem semântica ofensiva por si só não são suficientes ou bons indicativos de discurso de ódio, ou seja, tais termos demandam a análise do contexto em que são usadas. Enquanto [Waseem and Hovy 2016] utiliza uma definição mais ampla de discurso de ódio, ou seja, os autores levam em consideração gírias raciais e sexistas, críticas a minorias e *hashtags* específicas para avaliar um texto como discurso de ódio.

Para tanto, como inspiração deste trabalho é utilizada a competição SemEval 2019[5] e a definição de linguagem ofensiva proposta por eles. Sendo que, essa define como ofensiva qualquer forma de linguagem não-aceitável (profanidade) ou ofensa a um alvo, sendo ela velada ou direta. Um conjunto de 13.240 *tweets* anotados foi utilizado, providenciado por [Zampieri et al. 2019].

Neste trabalho, tendo em vista os desafios mencionados, o foco está no problema de identificação e classificação de traços de discurso de ódio na língua inglesa. Tal problema foi dividido em 3 tarefas, a fim de reduzir sua complexidade:

- Tarefa A: Classificar se uma mensagem do Twitter é considerada "ofensiva" ou "não-ofensiva";
- Tarefa B: Uma vez que a mensagem é considerada "ofensiva", o objetivo passa a ser, classificar se as ameaças e ofensas são direcionadas a um alvo em específico ou não;
- Tarefa C: Uma vez que as ameaças e ofensas são direcionadas a um alvo, o objetivo agora passa a ser, identificar quem é o alvo dentre 3 possíveis: (i) um indivíduo; (ii) um grupo de pessoas; ou (iii) eventos, organizações ou situações.

Neste trabalho, pretende-se resolver as tarefas mencionadas anteriormente com classificadores textuais clássicos, como o *SVM, Árvore de Decisão, Random Forest e Naive Bayes.*

Este artigo está organizado da seguinte forma: a Seção 2 discorre sobre os trabalhos relacionados, a Seção 3 apresenta os principais desafios e discorre sobre a automatização do processo de classificação de discurso de ódio. A Seção 4 apresenta uma descrição dos experimentos realizados e metodologia aplicada na execução dos mesmos. A Seção 5 aborda os resultados alcançados. Por fim, a Seção 6 apresenta as considerações finais e uma descrição dos trabalhos futuros.

2. Trabalhos relacionados

Em [Djuric et al. 2015], os autores têm como objetivo a detecção de discurso de ódio em comentários *online*. Eles definem discurso de ódio como "discurso abusivo tendo como

[2]https://www.facebook.com/
[3]https://www.youtube.com/
[4]https://twitter.com/
[5]https://competitions.codalab.org/competitions/20011

alvo um grupo de características, como etnicidade, religião ou gênero". O estudo foca na vantagem de se utilizar a representação *paragraph2vec*, a qual se baseia nas palavras ao redor para predizer a palavra central. O trabalho valida que a utilização de *paragraph2vec* é relevante e pode trazer resultados melhores do que representação tradicionais como *Bag of words*.

O trabalho de [Waseem and Hovy 2016] dá um foco ao discurso de ódio na forma de comentários racistas e sexistas. No estudo foi utilizada um definição mais ampla de discurso de ódio, onde os autores caracterizam como ofensivo qualquer texto que contenha pelo menos uma de várias características citadas. O estudo fez o uso de um conjunto de dados de 16.914 *tweets* anotados com diferentes combinações de atributos.

Como dito anteriormente, [Waseem et al. 2017] tem como objetivo mostrar os problemas que as diferentes formas de se referir a discurso de ódio podem causar, como pesquisadores se referindo a problemas diferentes utilizando a mesma expressão ou utilizando expressões diferentes para caracterizar problemas iguais. Levando em consideração as distintas formas que múltiplos autores podem se referir a esse problema, o que pode ocasionar confusões e até contradições. O autor apresenta uma tipologia que sintetiza as diferentes sub-tarefas presentes em detecção de discurso de ódio.

Em [Davidson et al. 2017], os autores abordaram o problema da detecção de discurso de ódio pela perspectiva de criar uma distinção, na caracterização de *tweets*, entre *tweets* com instâncias de linguagem ofensiva e *tweets* com a presença de discuso de ódio. Os textos foram anotados em três categorias: (i) se os *tweets* continham discurso de ódio; (ii) se continham linguagem ofensiva; e (iii) se não apresentavam nenhuma das duas características.

Semelhante ao trabalho de [Davidson et al. 2017], em [Badjatiya et al. 2017] os autores tiveram um foco em discriminar "tipos de ofensas" para o contexto de discurso de ódio. Utilizando o conjunto de dados provido por [Davidson et al. 2017], o estudo emprega o uso de classificação supervisionada, juntamente com um conjunto de atributos que incluem *n-gram*[6], *skip-gram* e representação de palavra baseada em agrupamento. Chegando à conclusão de que distinguir entre "discurso de ódio direcionado a um indivíduo ou grupo" e "palavrões ou termos de baixo calão" não é uma tarefa trivial, e que tal tarefa pode requerer atributos que tragam mais informações sobre o contexto.

[Watanabe et al. 2018] tem como objetivo detectar expressões de ódio no Twitter. O estudo faz o uso de unigramas, atributos baseados em semântica, sentimentos e padrões, que foram automaticamente coletados do conjunto de dados de 2010 *tweets*. No estudo são apresentados resultados de diferentes combinações de atributos em dois tipos de classificação (binária e ternária). Com o método de classificação proposto, alcançaram 87% de acurácia na classificação binária e 78% de acurácia na classificação ternária.

O estudo de [Zampieri et al. 2019] caracteriza o tipo e o alvo de uma mensagem ofensiva em redes sociais. Para tal propósito os autores compilaram o *Offensive Language Identification Dataset* (OLID), um novo conjunto de dados anotado com: (i) se o *tweet* possui traços de linguagem ofensiva ou não; (ii) No caso de possuir linguagem ofensiva, os anotadores também classificam se o texto possui um alvo ou não; (iii) Possuindo um alvo, também são anotados o tipo de alvo — indivíduo, grupo ou outros. Os autores ainda

[6]Termos compostos por n palavras. Um exemplo de 2-grams é "bom dia".

comparam diferentes modelos de classificação com aprendizado de máquina para inferir as classes anotadas.

A Tabela 1 resume os trabalhos mencionados de acordo com suas caraterísticas, ou seja, o tipo de fonte (corpus) que foi utilizado para obtenção dos textos nos experimentos, os atributos que foram considerados, os rótulos utilizados para cada classe e os algoritmos de classificação, respectivamente.

Tabela 1. Resumo dos trabalhos relacionados

Trabalho	Fonte Textual	Atributos	Rótulos	Algoritmos
[Djuric et al. 2015]	Yahoo Finance	BOW (tf); BOW (tf-idf); paragraph2vec	Não especificado	Não especificado
[Waseem and Hovy 2016]	Twitter	unigrams; bigrams; trigrams; fourgrams; word n-gram; gender; location	Offensive; Not offensive	Logistic regression
[Davidson et al. 2017]	Twitter	*Unigram; Bigram; Trigram*	*Hate; Offensive; Neither*	Logistic Regression; Naive Bayes; Decision Trees; Random Forests; Linear SVM
[Badjatiya et al. 2017]	Twitter	Char n-gram; Logistic Regression; TF-IDF; SVM; GBDT; BoWV; Random Embedding	Racist; Sexist; Neither	Fast Text; Convolutional Neural Networks; Long Short-term Memory Networks
[Watanabe et al. 2018]	Twitter	Unigram; Patterns; Sentiment-based; Semantic	Offensive or Not; Hateful, Offensive or Clean	J48graft
[Zampieri et al. 2019]	Twitter	Word uni-gram; FastText Embeddings; Updatable Embeddings	Offensive; Not offensive; Targeted Insult; Untargeted; Individual; Group; Other	SVM; BiLSTM; CNN

Como evidenciado pelos trabalhos relacionados apresentados anteriormente não existe na literatura um estudo comparativo em um conjunto de dados único. Busca-se esse trabalho para preencher tal lacuna.

3. Linguagem Ofensiva em Redes Sociais

A tarefa de identificar e caracterizar linguagem ofensiva em redes sociais apresenta desafios particulares devido à fonte textual e, aos métodos utilizados para tal obtenção. Alguns desses desafios são:

Tamanho do texto: Um *tweet* contém um tamanho máximo de 280 caracteres e não contem um tamanho mínimo. Sendo assim, os *tweets* recolhidos apresentam uma grande variação em seus tamanhos[7].

Esparsidade dos dados: *Tweets* apresentam uma grande quantidade de variações na maneira em que as palavras são escritas. Esse fenômeno causa uma esparsidade dos dados e tem um impacto sobre o desempenho global da identificação de discurso de ódio. A principal razão para a esparsidade de dados é o fato de que uma grande porcentagem dos termos que aparecem nos *tweets* ocorrem menos de 10 vezes [Saif et al. 2012] em todo o conjunto de dados.

Símbolos especiais (Special tokens): Emojis e URLs são exemplos de símbolos utilizados em *tweets*. Tais símbolos podem gerar dificuldade na classificação de textos pela inabilidade de se julgar a ideia passada por uma imagem ou o significado de um emoji usado em uma situação específica.

Variação de Tópicos: Os tópicos abordados nos *tweets* não são limitados por nenhum fator. Podendo causar problemas na hora da classificação dos *tweets* onde palavras iguais ditas em situações diferentes podem ter significados distintos.

[7] `https://about.twitter.com/`

Quantidade de Dados: Mesmo levando-se em consideração a limitação de 280 caracteres imposta pela rede, a quantidade de *tweets* postada a cada instante é extremamente grande. Em 2016, a rede social afirmou[8] que, durante a transmissão televisiva do filme "*Laputa: Castelo no Céu*" no Japão, os usuários criaram um novo recorde de 143.199 *tweets* por segundo. Também foi informado pelo Twitter que em média 6 mil mensagens são postadas por segundo. Em um ambiente como este, a filtragem de algo especifico se torna muito complicada e poluída pela presença de várias informações ou textos "indesejáveis".

Estilo de linguagem: Devido à grande diversidade de usuários do Twitter, o estilo de escrita e linguagem também é muito variável. As mensagens variam de estilos de escrita, como por exemplo de textos jornalísticos a textos completamente informais com gírias e expressões idiomáticas. Além disso, o vocabulário usado pode mudar rapidamente. Tudo isso pode conduzir a problemas de identificação de discurso de ódio por meio de recursos léxicos ou em lidar com dados de treinamento que foram anotados com a classe previamente.

Contexto Multi-lingual: Usuários do Twitter podem usar em seus textos, além de palavras de sua língua nativa, palavras de outros idiomas, até em frases que estão escritas majoritariamente em seu idioma nativo.

Tokenização: Em *tweets*, palavras podem nem sempre estar separadas por espaços, podendo apresentar múltiplas palavras sem uma divisão clara entre elas. Tal fato dificulta a tarefa de tokenização, a qual normalmente é feita separando os termos através de espaços em branco.

3.1. Automatizando o processo

Após uma análise criteriosa dos diversos trabalhos sobre identificação de discurso de ódio é possível observar que, em sua maioria, tais métodos seguem uma sequência de passos comuns à mineração de textos (ver Figura 1). O primeiro passo diz respeito à seleção de um conjunto de *tweets* de interesse. Esse processo se dá por meio de uma busca por tópicos de pesquisa, por *hashtags*, *emoticons*, um período de pesquisa, entre outras formas. Estes *tweets* devem ser anotados, sendo que na maioria dos casos este processo é realizado manualmente, com o auxílio de especialistas de domínio, onde os anotadores são orientados com um conjunto de regras as quais devem ser utilizadas para se avaliar o que é desejado.

Ainda na Figura 1, no passo 3 os *tweets* anotados são organizados em uma lista com seus respectivos rótulos. O passo 4 diz respeito ao pré-processamento, que deve ser feito buscando a padronização do texto, removendo características que são consideradas irrelevantes ou sem utilidade para a classificação. Para a padronização, utiliza-se de ferramentas como *Stemming* e *Lemmatization* que fazem o que é chamado de normalização de texto (ou normalização de palavras), onde palavras com diferentes inflexões são alteradas para sua palavra "raiz". A alteração de caraterísticas indesejáveis como a remoção de palavras como *Stop Words*, que são palavras com baixo poder de discriminação (por exemplo, "*of*", "*out*", "*as*", etc.), remoção de múltiplos espaços em branco e dígitos e alteração de emojis e URL's, podem ajudar na classificação de um texto ou no treino de um classificador.

[8] `https://www.dsayce.com/social-media/tweets-day/`

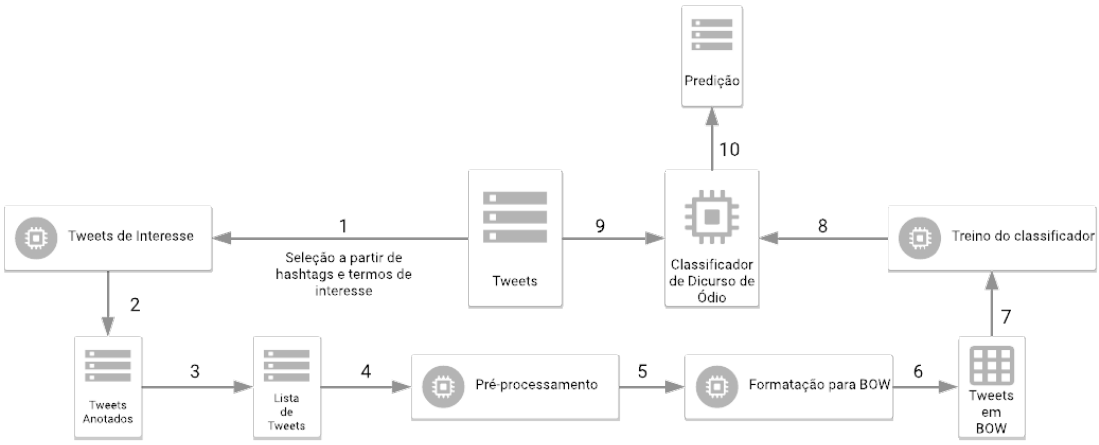

Figura 1. Metodologia Utilizada

Com o texto pré-processado, é necessário representá-lo de forma estruturada para que este seja dado como entrada para o classificador em questão (Passos 6 e 7 da Figura 1). Um método comumente utilizado é a transformação do texto para o formato *Bag of Words*, onde os *tweets* são representados por meio das linhas de uma tabela, como ilustrado na Tabela 2, na qual as colunas representam as palavras ou termos nas mensagens, e os valores associados às colunas representam a frequência ou presença desses termos [Silva 2016].

Tabela 2. Representação dos *tweets* em um *Bag of Words*

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

Após a formatação do texto, o treino de um classificador escolhido é feito. A *Bag of Words* é fornecida como entrada para o classificador escolhido (Passos 7 da Figura 1). Vários classificadores podem ser utilizados, entre eles estão, Árvore de decisão, *Random Forest*, *Neural Net*, *Naive Bayes*, *QDA*, etc.

Com o treino do classificador realizado, obtém-se um modelo de classificação de Discurso de ódio o qual pode ser aplicado em novas mensagens. Novos *tweets* agora precisam ser pré-processados e submetidos ao modelo de classificação obtido. Tal modelo é capaz de inferir a qual classe o texto entregue pertence (Passos 8, 9 e 10 da Figura 1).

4. Experimentos Realizados

4.1. Conjunto de Dados

Para os experimentos realizados neste trabalho utilizamos o conjunto de dados disponibilizado por [Zampieri et al. 2019] e denominado "OLID" *Offensive Language Identification*

Dataset. Composto de 13.240 *tweets* anotados usando *crowdsourcing*[9]. O conjunto de dados é composto de *tweets*, selecionados a partir de *hashtags* e termos de interesse. Tal conjunto contém as anotações necessárias que classificam o *tweet* obtido em (1) ofensivo ou não; (2) classificam se o *tweet* tem ou não um alvo específico e (3) caso o *tweet* tenha um alvo, se o alvo é um individuo, um grupo ou outros, como uma organização.

Cada linha do arquivo é composta por quatro seções: a primeira contendo texto presente no *tweet* e as outras três indicando cada uma das sub-tarefas passadas pelo SemEval, indicando respectivamente a presença de linguagem ofensiva, a categorização da linguem ofensiva, quando presente, em relação ao alvo (um grupo ou indivíduo ou sem um alvo específico) e por último a identificação desse alvo (um indivíduo, um grupo ou outros) assim como apresentado na Tabela 3.

Tabela 3. Quatro *tweets* do conjunto de dados, cada um com suas respectivas anotações para cada classe

Tweet	SubTask A	SubTask B	SubTask C
@USER Good riddance.	NOT	—	—
@USER Yeah we need some more made up bullshit protestors and antifa lol time for an epic beatdown	OFF	UNT	—
Shouldn't pussy grabbing @USER be the one wearing the gloves while handling food? #MAGA URL	OFF	TIN	IND
@USER Why? Why are liberals so trashy?	OFF	TIN	GRP

Dos 13.240 *tweets* disponibilizados, como visto na Figura 2, na tarefa 'A', 4.400 *tweets* são categorizados como Ofensivos ("OFF") e 8.840 são caracterizados como Não Ofensivos ("NOT"). Na tarefa 'B' dentre *tweets* considerados ofensivos 3.876 são insultos e ameaças a um alvo específico ("TIN") e 524 não tem um alvo específico ("UNT"). Na tarefa 'C', os insultos são classificados por alvo, onde 2.407 são direcionados a um indivíduo ("IND"), 1.074 são direcionados a um grupo de pessoas ("GRP") e 395 são direcionados a outros ("OTH") (eventos, organizações ou situações).

(a) Tarefa A (b) Tarefa B (c) Tarefa C

Figura 2. Numero de *tweets* em cada classe das tarefa

4.2. Preprocessamento

O objetivo de aumentar a previsibilidade do texto motivou as seguintes técnicas de pré processamento a serem utilizadas. As palavras consideradas *stopwords* pelo *corpus* provido pelo NLTK (*"Natural Language Toolkit"*)[Bird et al. 2009] foram removidas. As-

[9]https://pt.wikipedia.org/wiki/Crowdsourcing

sim como, foram removidos dígitos e pontuações. Citações, URL's e emoji's foram substituídos por "@USER", "URL" e "EMOJI" respectivamente.

Por último o *Porter Stemmer* do NLTK foi utilizado com o objetivo de padronizar diferentes flexões de palavras relacionadas em um palavra base ou raíz e o texto foi divido em um *Bag of Words* com *n-grams*, com n variando de 1 a 3.

4.3. Classificadores

Experimentos foram realizados com vários classificadores sem alterações em como o pré-processamento foi feito. Todos os experimentos foram realizados com validação cruzada com 10 partições e os resultados representam a média dos valores obtidos na validação cruzada. Os classificadores utilizados foram: Random Forest, SVM com kernel linear, rbf e sigmoid, Naive Bayes multinomial e Árvore de decisão.

4.4. Avaliação

Diferentes métricas foram utilizadas para a avaliação dos resultados das tarefas A, B e C. Para melhor entendimento, apresentamos as seguintes definições:

- *Verdadeiro Positivo* (VP): Diz respeito à correta classificação como "ofensivo". Por exemplo, a classe real é "ofensivo" e o classificador prediz o rótulo como "ofensivo".
- *Verdadeiro Negativo* (VN): Significa a correta classificação como "não ofensivo". Por exemplo, a classe real é "não ofensivo" e o classificador prediz o rótulo como "não ofensivo".
- *Falso Positivo* (FP): Significa a classificação errada como "ofensivo". Por exemplo, a classe real é "não ofensivo" e o classificador prediz o rótulo como "ofensivo".
- *Falso negativo* (FN): Significa a classificação errada como "não ofensivo". Por exemplo, a classe real é "ofensivo" e o classificador prediz o rótulo como "não ofensivo".

Os classificadores são avaliados de acordo com as seguintes métricas: acurácia, precisão, *recall* e F1-Score. As equações abaixo apresentam como os cálculos são feitos.

$$\textit{Acurácia} = \frac{VP + VN}{VP + FN + VP + VN} \tag{1}$$

$$\textit{Precisão} = \frac{VP}{VP + FP} \tag{2}$$

$$\textit{Recall} = \frac{VP}{VP + FN} \tag{3}$$

$$\textit{F1-Score} = \frac{2 * Precisão * Recall}{Precisão + Recall} \tag{4}$$

4.5. Nossa abordagem

Para alcançar os resultados presentes no artigo foi utilizada uma metodologia experimental (veja a Figura 1). Após a obtenção do conjunto de dados, o pré-processamento dos dados foi realizado, retirando as palavras consideradas irrelevantes para o entendimento do texto, as *"stopwords"*. Além disso, a remoção de pontuações e a padronização de palavras com diferentes escritas (mas com o mesmo significado, em uma mesma palavra) foram realizados. Em seguida, o texto foi convertido para o formato de *bag of words*.

Diferentes experimentos foram realizados alterando-se os classificadores e os parâmetros utilizados em cada classificador, sem alterar a maneira como o pré-processamento dos dados foi realizado, com o intuito de avaliar a qualidade de predição de cada um dos classificadores testados. Os classificadores utilizados incluem Árvore de decisão, *Random Forest*, *Naive Bayes*, *SVM Linear* com *tuning*[10], e *SVM RBF* com *tuning*. Nos classificadores, onde o *tuning* foi aplicado, os valores testados dos parâmetros foram as combinações dos valores de C = 0.001, 0.10, 0.1, 10, 25, 50, 100, 1000 e de Gamma = 1e-2, 1e-3, 1e-4, 1e-5.

Na tarefa B, os mesmos experimentos foram realizados. Os *tweets* já classificados como ofensivos foram utilizados com o objetivo de classificar se eles continham alvo específico "TIN" ou não "UNT". O mesmo aconteceu na tarefa C, com os *tweets* classificados como ofensivos e com um alvo específico, agora com o intuito de se avaliar se o alvo em questão era um individuo "IND", um grupo "GRP" ou outros "OTH" como empresas ou instituições.

5. Resultados

As Tabelas 4, 5 e 6 apresentam os valores de Precisão, *Recall*, *F1 score*, Acurácia e o tempo de treinamento do classificador em segundos para cada classe de cada tarefa dos classificadores considerados. Assim como os melhores valores obtidos pelo estudo de [Zampieri et al. 2019]. Os resultados foram calculados a partir da média de tais métricas obtidas com a validação cruzada com 10 partições (*10 fold cross validation*).

Como pode ser observado na Tabela 4, os classificadores SVM RBF e SVM Sigmoid (ambos com *tuning*) alcançaram o mesmo desempenho em termos de acurácia. Entretanto como o SVM Sigmoid é mais rápido, sugere-se ao usuário optar por ele no seu treinamento. Considerando o desempenho em relação às classes individualmente, em específico a classe "OFF", é possível observar que o SVM RBF (com *tuning*) obteve um *F1-Score* de 0.6.

Na Tabela 5 observa-se que os classificadores *Random Forest* e SVM RBF com *tuning* obtiveram performances semelhantes em relação à acurácia. O mesmo ocorre ao se observar o desempenho da classe "TIN", onde ambos os classificadores obtiveram um *F1 Score* de 0.93. Levando em consideração tais desempenhos, recomenda-se o uso do classificador *Random Forest* por ter sido mais rápido que o SVM RBF neste caso.

Como pode ser observado na Tabela 6, por se tratar de um problema multi-classe ("GRP", "IND", "OTH") houve um aumento na complexidade da classificação. Desta forma, verifica-se que o espaço de atributos escolhidos (n-gramas, com n=1,2,3) e/ou os

[10]Expressão inglesa traduzida como afinação ou otimização, em que se avaliam vários valores para os hiperparâmetros dos algoritmos em questão com o objetivo de aumentar o poder preditivo.

Tabela 4. Valores obtidos na tarefa A – Os maiores valores de cada métrica estão em negrito.

Classificador	Classes	F1-Score	Precisão	Recall	Acurácia	Tempo (em seg.)
DecisionTree	NOT	0.80	0.78	0.83	0.73	212.8
	OFF	0.57	0.62	0.53		
	Média	0.68	0.7	0.68		
RandomForest	NOT	0.83	0.73	0.96	0.74	39.0
	OFF	0.43	0.81	0.28		
	Média	0.63	**0.77**	0.62		
NaiveBayes	NOT	0.76	0.80	0.73	0.7	0.3
	OFF	0.58	0.54	0.63		
	Média	0.67	0.67	0.68		
SVM Linear c=10	NOT	0.82	0.78	0.87	0.75	812.7
	OFF	0.58	0.68	0.51		
	Média	0.7	0.73	0.69		
SVM RBF c=1000 gamma=0.001	NOT	0.82	0.79	0.87	**0.76**	939.8
	OFF	0.6	0.67	0.54		
	Média	**0.71**	0.73	**0.7**		
SVM Sigmoid c=50 gamma=0.01	NOT	0.83	0.78	0.89	**0.76**	898.1
	OFF	0.59	0.7	0.51		
	Média	**0.71**	0.74	**0.7**		
[Zampieri et al. 2019] CNN	NOT	0.90	0.87	0.93	–	–
	OFF	0.70	0.78	0.63		
	Média	0.80	0.82	0.78		

Tabela 5. Valores obtidos na tarefa B – Os maiores valores de cada métrica estão em negrito.

Classificador	Classes	F1-Score	Precisão	Recall	Acurácia	Tempo (em seg.)
DecisionTree	TIN	0.91	0.88	0.94	0.84	25.4
	UNT	0.14	0.22	0.11		
	Média	0.53	0.50	0.52		
RandomForest	TIN	0.93	0.88	0.99	**0.87**	5.9
	UNT	0.06	0.41	0.03		
	Média	0.5	**0.65**	0.51		
NaiveBayes	TIN	0.87	0.88	0.86	0.78	0.19
	UNT	0.17	0.15	0.2		
	Média	0.52	0.52	0.53		
SVM Linear c=25	TIN	0.9	0.89	0.92	0.83	25.4
	UNT	0.17	0.22	0.15		
	Média	0.54	0.56	**0.54**		
SVM RBF c=1000 gamma=0.01	TIN	0.93	0.88	0.97	**0.87**	39.54
	UNT	0.15	0.39	0.09		
	Média	**0.54**	0.64	0.53		
SVM Sigmoid c=1000 gamma=0.01	TIN	0.92	0.88	0.95	0.85	33.2
	UNT	0.16	0.25	0.11		
	Média	**0.54**	0.57	0.53		
[Zampieri et al. 2019] CNN	TIN	0.92	0.94	0.90	–	–
	UNT	0.33	0.32	0.63		
	Média	0.62	0.63	0.76		

classificadores treinados não foram adequados, tendo em vista os valores de Precisão, *Recall* e *F1-Score* iguais a zero para a classe "OTH". Entretanto, observa-se o mesmo comportamento no trabalho de [Zampieri et al. 2019].

Tabela 6. Valores obtidos na tarefa C – Os maiores valores de cada métrica estão em negrito.

Classificador	Classes	F1 Score	Precisão	Recall	Acurácia	Tempo (em seg.)
Decision Tree	GRP	0.52	0.52	0.53	0.65	23.8
	IND	0.77	0.74	0.79		
	OTH	0.09	0.13	0.07		
	Média	0.46	0.46	0.46		
Random Forest	GRP	0.43	0.59	0.34	0.67	4.2
	IND	0.78	0.68	0.92		
	OTH	0	0	0		
	Média	0.40	0.42	0.42		
Naive Bayes	GRP	0.55	0.50	0.64	0.66	0.18
	IND	0.77	0.77	0.78		
	OTH	0.04	0.13	0.02		
	Média	0.45	0.47	0.48		
SVM Linear c=10	GRP	0.58	0.59	0.56	0.70	46.26
	IND	0.81	0.76	0.85		
	OTH	0.08	0.22	0.05		
	Média	**0.49**	0.52	0.49		
SVM RBF c=1000 gamma=0.001	GRP	0.59	0.59	0.57	**0.71**	50
	IND	0.81	0.76	0.87		
	OTH	0.08	0.3	0.05		
	Média	**0.49**	**0.55**	**0.5**		
SVM Sigmoid c=1000 gamma=0.001	GRP	0.58	0.6	0.55	0.70	48
	IND	0.81	0.75	0.87		
	OTH	0.08	0.25	0.05		
	Média	**0.49**	0.53	0.49		
[Zampieri et al. 2019] CNN	GRP	0.67	0.75	0.60	–	–
	IND	0.75	0.63	0.94		
	OTH	0.00	0.00	0.00		
	Média	0.47	0.46	0.51		

6. Conclusão e trabalhos futuros

A presença do discurso de ódio é um problema existente em todas as redes sociais, e com o número de usuários e de postagens diárias a classificação manual de tudo que é publicado se torna inviável. A quantidade de dados postados nas redes sociais evidencia a necessidade de se automatizar a detecção, caracterização e possível remoção das mensagens contendo discurso de ódio.

Este trabalho demonstrou a possibilidade do uso de classificadores clássicos para a tarefa de classificação de discurso de ódio para a língua inglesa. Utilizando um conjunto de dados de 13.240 *tweets* anotados, testados com 5 classificadores diferentes (Árvore de decisão, *Random Forest*, *Naive Bayes*, *SVM Linear* e *SVM RBF*) em 3 tarefas distintas. Foi obtido uma resultado de 76%, 87% e 71% de acurácia nas tarefas A, B e C respectivamente.

Como trabalhos futuros, pretende-se estudar novos atributos, tais como léxicos [Razavi et al. 2010], [Njagi et al. 2015], analizadores de classes gramaticais (*Part-of-Speech Tagging*) [Derczynski et al. 2013], bem como avaliar outros classificadores, como o BiLSTM e CNN [Zampieri et al. 2019].

Referências

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. *CoRR*, abs/1706.00188.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 29–30, New York, NY, USA. ACM.

Njagi, D., Zuping, Z., Hanyurwimfura, D., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, AI'10, pages 16–27, Berlin, Heidelberg. Springer-Verlag.

Saif, H., He, Y., and Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. In *2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at the 21st International Conference on theWorld Wide Web (WWW'12)*, pages 2–9. CEUR Workshop Proceedings (CEUR-WS.org).

Silva, N. F. F. d. (2016). *Análise de sentimentos em textos curtos provenientes de redes sociais*. PhD thesis, Universidade de São Paulo.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*.

Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. *CoRR*, abs/1902.09666.