

UNIVERSIDADE FEDERAL DE OURO PRETO  
DEPARTAMENTO DE COMPUTAÇÃO

Lucas de Rocha Castro

**UM ESTUDO EMPÍRICO SOBRE TÉCNICAS  
PARA DETECÇÃO DE DISCURSOS DE ÓDIO  
EM POSTAGENS PÚBLICAS ESCRITAS EM  
PORTUGUÊS**

Ouro Preto, MG  
2019

Lucas de Rocha Castro

UNIVERSIDADE FEDERAL DE OURO PRETO  
DEPARTAMENTO DE COMPUTAÇÃO

Monografia II apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

**Orientador:** Amanda Sávio Nascimento e Silva

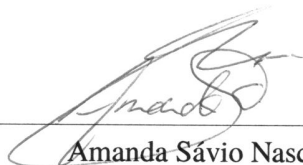
Ouro Preto, MG  
2019

Lucas de Rocha Castro

**UM ESTUDO EMPÍRICO SOBRE TÉCNICAS PARA DETECÇÃO DE  
DISCURSOS DE ÓDIO EM POSTAGENS PÚBLICAS ESCRITAS EM  
PORTUGUÊS**

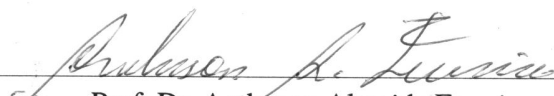
Monografia II apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau em Bacharel em Ciência da Computação.

Aprovada em Ouro Preto, 17 de julho de 2019.



---

Amanda Sávio Nascimento e Silva  
Universidade Federal de Ouro Preto  
Orientador



---

Prof. Dr. Anderson Almeida Ferreira  
Universidade Federal de Ouro Preto- UFOP  
Examinador



---

Prof. Dra. Dayanne Gouveia Coelho  
Universidade Federal de Ouro Preto - UFOP  
Examinador

# Resumo

É cada vez maior o número de usuários conectados a redes sociais, ambientes propícios para a liberdade de expressão. Contudo, também é crescente o número de discursos de ódios ou ofensivos nessas redes, podendo inclusive serem enquadrados em códigos penais e levar a punições diversas. Nesse contexto, há muitos autores que estudam e propõe meios para detecção de discursos de ódios e mensagens ofensivas em redes sociais, sendo recorrentes soluções que são modeladas utilizando métodos de aprendizado de máquina. Contudo, ainda são poucos trabalhos que abordam a detecção de ódio em postagens escritas em português. Nesse contexto, este trabalho é um estudo comparativo que demonstra a eficiência da combinação de técnicas de extração de radical, *data augmentation and pseudo labelling*, *undersampling* e *feature selection* para a detecção de discurso de ódio em redes sociais brasileiras utilizando os classificadores de Naive Bayes e Support Vector Machine. Segundo a métrica de *f-measure*, os resultados finais demonstraram que *data augmentation* não é uma técnica efetiva, enquanto que, *feature selection* se comporta bem com qualquer outra técnica, resultando em uma *f-measure* de até 91% quando utilizada com *undersampling*. A extração de radical foi bom em alguns cenários mas pouco eficiente ou neutra em outros cenários. Para trabalhos futuros, pretende-se realizar testes utilizando outros classificadores, técnicas e implementar algoritmos para categorização dos discursos de ódio como machismo, homofobia, racismo, dentre outras formas de discriminações.

**Palavras-chave:** discurso de ódio. língua portuguesa. aprendizado de máquina. *Naive Bayes*, *Support Vector Machine*.

# Abstract

*The number of users connected to social networks, environments favorable to freedom of expression, is increasing. However, the number of hating or offensive discourses in these networks is also increasing, and may even be framed in penal codes and lead to various punishments. In this context, there are many authors who study and propose means for detecting hate speech and offensive messages in social networks, being recurrent solutions that are modeled using machine learning methods. However, there are still few studies that deal with hate detection in written Portuguese posts. In this context, this work is a comparative study that demonstrates the efficiency of the combination of radical extraction techniques, data augmentation and pseudo-labeling, undersampling and feature selection for detection of hate speech in Brazilian social networks using the Naive Bayes and Support Vector Machine classifiers. According to the f-measure metric, the final results demonstrated that data augmentation is not an effective technique, whereas feature selection behaves well with any other technique, resulting in a f-measure of up to 91% when used with undersampling. Radical extraction was good in some scenarios but poorly efficient or neutral in other scenarios. For future work, we intend to perform tests using other classifiers, techniques and implement algorithms for categorizing hate speech such as machism, homophobia, racism, among other forms of discrimination*

**Keywords:** *hate speech, Portuguese language, machine learning. Naive Bayes, Support Vector Machine.*

# Lista de ilustrações

Figura 2.1 – Objetos separados por hiperplanos, perante suas classificações. Fonte: (DRAKOS, 2018) . . . . .	14
Figura 2.2 – Hiperplano ótimo. Fonte: (DRAKOS, 2018) . . . . .	15
Figura 2.3 – K-fold Cross Validation. Fonte: (ASHFAQUE; IQBAL, 2019) . . . . .	16
Figura 3.1 – Arquitetura do método apresentado . . . . .	20
Figura 3.2 – Gráfico dos resultados alcançados pelo baseline . . . . .	23
Figura 3.3 – Resultados do presente trabalho utilizando a base de dados OffComBR2 e diferentes combinações das técnicas utilizadas . . . . .	25
Figura 3.4 – Resultados do presente trabalho utilizando a base de dados OffComBR3 . . . . .	25
Figura A.1 – Fonte: (RANA; SINGHAL et al., 2015) . . . . .	32

# Lista de abreviaturas e siglas

DM	<i>Data Mining</i>
FN	<i>False Negative</i>
FP	<i>False Positive</i>
IDF	<i>Inverse Document Frequency</i>
LSTM	<i>Long Short Term Memory networks</i>
ME	<i>Maximum Entropy</i>
ML	<i>Machine Learning</i>
NB	<i>Naive Bayes</i>
STAI	<i>Shared Task on Aggression Identificatio</i>
SVM	<i>Support Vector Machine</i>
TF	<i>Term Frequency</i>
TF-IDF	<i>Term Frequency-Inverse Document Frequency</i>
TN	<i>True Negative</i>
TP	<i>True Positive</i>

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	Definição do Problema e Justificativa	2
1.2	Objetivo Geral e Objetivos Específicos	2
1.2.1	Objetivos Específicos	2
1.2.2	Estudo: visão geral	3
1.2.3	Organização do Trabalho	3
<b>2</b>	<b>Revisão de Literatura</b>	<b>4</b>
2.1	Fundamentação Teórica	4
2.1.1	Liberdade de Expressão e Discurso de Ódio	4
2.1.2	<i>Data Science</i>	4
2.1.3	Pré-processamento	5
2.1.3.1	N-Gramas	5
2.1.3.2	<i>Dataset</i> desbalanceado	6
2.1.3.2.1	<i>Data Augmentation and Pseudo Labelling</i>	6
2.1.3.3	Extração de Radicais	7
2.1.4	<i>Feature Selection</i>	7
2.1.4.1	Teste de Chi Quadrado	7
2.1.5	Classificação	10
2.1.5.1	Algoritmo de Naive Bayes	10
2.1.5.1.1	Probabilidade normal	10
2.1.5.1.2	Probabilidade Condicional	10
2.1.5.1.3	Teorema de Bayes	11
2.1.5.1.4	Aplicação do algoritmo	12
2.1.5.2	Support Vector Machine (SVM)	14
2.1.6	Métricas de Avaliação dos Resultados	15
2.1.7	<i>K-Fold Cross Validation</i>	16
2.2	Trabalhos Relacionados	17
<b>3</b>	<b>Desenvolvimento</b>	<b>20</b>
3.1	Método	20
3.1.1	Dataset	20
3.1.2	Pré-processamento	21
3.1.3	Extração de Características	21
3.1.4	Experimentação	22
3.2	Resultados	23
3.2.1	Discussão dos Resultados	26
<b>4</b>	<b>Considerações Finais</b>	<b>27</b>



4.1	Conclusão . . . . .	27
4.2	Trabalhos Futuros . . . . .	27
	<b>Referências . . . . .</b>	<b>28</b>
	<b>Anexos</b>	<b>31</b>
	<b>ANEXO A Porcentagens da distribuição de chi-quadrado . . . . .</b>	<b>32</b>

# 1 Introdução

Segundo (STATISTICA, 2018),<sup>c</sup> aproximadamente 2 bilhões de pessoas estão usando redes sociais no mundo todo, número que tende a crescer com a popularização de acesso a essas redes via dispositivos móveis. Redes sociais permitem que usuários se conectem com amigos ou pessoas do mundo todo. Nessas plataformas, as pessoas podem se expressar mediante postagem ou compartilhando algum tipo de conteúdo. Essa liberdade de expressão pode ser um combustível para discursos de ódio nessas redes, polarizando diversos grupos.

Segundo (NOCKLEBY, 2000), discurso de ódio é qualquer comunicação que deprecie uma pessoa ou um grupo a partir, por exemplo, de menções ofensivas às características referentes à diversidade étnico-racial, de gênero, de orientação sexual, de nacionalidade, de classe sócio-econômica, de religião, entre outras.

Durante três meses, a agência nova/sb<sup>1</sup> monitorou dez tipos de intolerância nas redes sociais, em relação às seguintes situações: aparência das pessoas, classes sociais, inúmeras deficiências, homofobia, misoginia, política, idade/geração, racismo, religião e xenofobia, coletando e analisando um total de 393.284 conteúdos sobre esses temas no Facebook, Twitter, Instagram, blogs e comentários em *sites* para fins diversos. Desses conteúdos analisados, mais de 84% continham expressões negativas, caracterizando discurso de ódio. Frequentemente pessoas famosas também são alvo deste tipo de discurso em redes sociais, como o caso da Fernanda Gentil, apresentadora da Rede Globo, que sofreu ataques homofóbicos no Twitter<sup>2</sup> em 2016 e da Taís Araújo, atriz e apresentadora da Rede Globo, que foi duramente atacada por comentários racistas<sup>3</sup> no ano de 2015.

Nesse contexto, é crescente o número de autores que estudam e propõem meios de detecção de discursos de ódios e mensagens ofensivas em redes sociais. Em (AGARWAL; SUREKA, 2015; BARTLETT et al., 2014; TING et al., 2013), os autores utilizaram para esse propósito algoritmos de *machine learning* para detectar tais discursos. Esses algoritmos automatizam o processo de classificação (se é um discurso de ódio ou não) de uma dada entrada (comentários ou frases). Já (GITARI et al., 2015), além de *machine learning*, utilizam da análise léxica baseada em dicionários, que contêm palavras que remetem o discurso de ódio, para detectar usuários odiosos. (MONDAL et al., 2018) categorizam os discursos de ódio em diversas categorias, a saber: racismo, orientação sexual, gênero, etnia, religião, dentre outros. Esses mesmos autores correlacionam o ambiente *online* (redes sociais) com o *offline* (ambiente real) comparando as categorias do seu estudo com crimes de ódio incidentes nessas categorias nos EUA.

<sup>1</sup> <https://www.novasb.com.br/>, Accessed: 2018-11-27

<sup>2</sup> <https://catracalivre.com.br/cidadania/fernanda-gentil-e-alvo-de-homofobia-nas-redes-sociais/>, Accessed: 2018-11-27

<sup>3</sup> <http://g1.globo.com/rio-de-janeiro/noticia/2015/11/atriz-tais-araujo-e-alvo-de-comentarios-racistas-em-rede-social.html>, Accessed: 2018-11-27

## 1.1 Definição do Problema e Justificativa

Conforme mencionado, a comunicação em redes sociais, via postagens ou compartilhamento de conteúdos específicos, muitas vezes, favorece a propagação de discursos de ódio. Esses discursos afetam negativamente diversos grupos da sociedade, podendo, inclusive, ser enquadrados como crimes, Lei 7716/89<sup>4</sup>. Embora existam trabalhos voltados para a identificação desses discursos nas redes sociais (MONDAL et al., 2018; TING et al., 2013; AGARWAL; SUREKA, 2015; PELLE; MOREIRA, 2017; ALMEIDA; NAKAMURA; NAKAMURA, 2017), ainda são poucos os trabalhos (PELLE; MOREIRA, 2017) focados na análise de discursos de ódios expressos em português, mais precisamente, em postagens realizadas em redes sociais no Brasil. Ainda que existam estudos comparando as diversas técnicas para classificação de discurso de ódio em postagens públicas escritas em inglês (DJURIC et al., 2015; DAVIDSON et al., 2017), há uma lacuna por estudos comparando a precisão dessas técnicas ao classificar textos em português. Nesse contexto, cabe destacar que português e inglês tem estruturas diferentes no que tange, por exemplo, construções gramaticais, gêneros (feminino ou masculino dos adjetivos e substantivos), uso dos verbos e ordem das palavras (e.g., posição dos advérbios). Existindo uma lacuna por esses trabalhos com foco na língua portuguesa e de trabalhos com esse foco que demonstrem com clareza os resultados obtidos, o presente trabalho visa fornecer resultados de técnicas de pré-processamento, seleção de características e classificação, que são posteriormente comparados com resultados do *baseline* utilizado (PELLE; MOREIRA, 2017).

## 1.2 Objetivo Geral e Objetivos Específicos

Este trabalho tem como principal objetivo realizar um estudo comparativo das técnicas de pré-processamento e classificação de postagens públicas, escritas em português, visando identificar discursos de ódio. Mais precisamente, serão contempladas no estudo a comparação das seguintes técnicas: extração de radicais, *feature selection*, *undersampling* e *oversampling* aplicadas aos classificadores de *Naive Bayes* e *Support Vector Machine*.

### 1.2.1 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Analisar combinações de diferentes técnicas de pré-processamento e classificação de postagens públicas para a identificação de discursos de ódio.
- Identificar, dentre as técnicas de pré-processamento e classificação de postagens públicas utilizadas, aquelas que apresentam maiores medidas de *f-measure* na identificação de discursos de ódio.

<sup>4</sup> [presrepublica.jusbrasil.com.br/legislacao/111031/lei-do-crime-racial-lei-7716-89](http://presrepublica.jusbrasil.com.br/legislacao/111031/lei-do-crime-racial-lei-7716-89)

- Comparar os resultados obtidos neste trabalho com os resultados existentes na literatura de trabalhos que classificam discursos de ódios em postagens escritas em português.

### 1.2.2 Estudo: visão geral

O trabalho proposto é composto por três componentes: pré-processamento, extração de características e classificação. O proposto trabalho visa demonstrar resultados utilizando diferentes técnicas de pré-processamento e de extração de características se comparado ao baseline (PELLE; MOREIRA, 2017)

### 1.2.3 Organização do Trabalho

Este trabalho está organizado em 4 capítulos. No capítulo 2, são apresentados conceitos teóricos, técnicas e termos importantes para o entendimento deste trabalho e trabalhos relacionados à classificação de texto para a análise de discurso de ódio e polaridade presentes em postagens públicas. No Capítulo 3 é apresentado o desenvolvimento, descrevendo as técnicas utilizadas em cada parte do processo, contemplando desde o *dataset* utilizado, até os resultados obtidos para cada técnica apresentada. No Capítulo 4, são apresentadas as considerações finais e possíveis trabalhos futuros.

## 2 Revisão de Literatura

Neste capítulo, serão apresentados os conceitos fundamentais para o entendimento desta pesquisa (Seção 2.1) e os trabalhos relacionados (Seção 2.2).

### 2.1 Fundamentação Teórica

Nesta seção, são apresentados os principais conceitos referentes à liberdade de expressão e discurso de ódio (Seção 2.1.1) e Data Science (KELLEHER; TIERNEY, 2018) (Seção 2.1.2). São também apresentadas, em linhas gerais: conceitos, termos e técnicas de pré-processamento (Seção 2.1.3), extração de características (Seção 2.1.4) e conceitos e aplicação dos classificadores (Seção 2.1.5).

#### 2.1.1 Liberdade de Expressão e Discurso de Ódio

A Lei Constitucional n.º 1/92 de 25 de NOVEMBRO, artigo 37.º, nos assegura o direito à manifestação livre do pensamento, à busca e propagação de informações, sem impedimentos ou discriminações. De maneira geral, temos direito de nos expressar livremente sem que essa liberdade cause qualquer tipo de dano à alguma pessoa ou grupo. Uma das formas de abuso dessa liberdade é o discurso de ódio, que acontece quando um indivíduo manifesta de modo a inferiorizar e discriminar outra pessoa ou grupo baseado em suas características, como sexo, etnia, orientação sexual, religião, entre outras (SILVA, 2014). Atualmente a Lei 7.716/89<sup>1</sup> prevê punição para discriminação devido à raça, cor, etnia, religião ou procedência nacional. Encontra-se em tramitação na Câmara dos Deputados, o projeto de Lei 7582/2014, de responsabilidade da Deputada Maria do Rosário, que amplia a definição de crimes de ódio para abranger “*Toda pessoa, independentemente de classe e origem social, condição de migrante, refugiado ou deslocado interno, orientação sexual, identidade e expressão de gênero, idade, religião, situação de rua e deficiência goza dos direitos fundamentais inerentes à pessoa humana, sendo-lhe asseguradas as oportunidades para viver sem violência, preservar sua saúde física e mental e seu aperfeiçoamento moral, intelectual e social.* (ROSARIO, 2014)”

#### 2.1.2 Data Science

Data Science é a ciência que estuda o conjunto de princípios, definições de problema, algoritmos e processos para extrair comportamentos não explícitos e úteis de um grande *data set*<sup>2</sup>. É importante introduzir os conceitos de *Machine Learning* (ML) e *Data Mining* (DM) que

<sup>1</sup> [http://www.planalto.gov.br/ccivil\\_03/leis/l7716.htm](http://www.planalto.gov.br/ccivil_03/leis/l7716.htm)

<sup>2</sup> data set é um conjunto de dados relacionados a um conjunto de instâncias, tal que cada instância é composta por um conjunto de atributos.

são conceitos que possuem o mesmo foco: melhorar a tomada de decisão através da análise de dados. *Machine learning* concentra-se no *design* e avaliação de algoritmos para extrair padrões de dados, enquanto *data mining* geralmente concentra-se na análise de dados estruturados, isto é, dados que podem ser armazenados em tabelas. *Data science* engloba esses dois conceitos e ainda lida com a captura e refinamento de dados não estruturados de páginas da internet.

No conceito de *Data Science*, a classificação de dados se dá após a extração das características que o identifique. Por exemplo, no comentário que contém a frase "eu odeio gay" o comportamento extraído pode ser classificado como um discurso de ódio e ainda com um atributo mais específico: homofobia. A extração de características de um dado, na tarefa de classificação de texto, é a última parte do pré-processamento desses dados. O pré-processamento de dados é necessário para remover ruídos, inconsistência de dados e para representar os dados por meio de suas características de tal forma que o classificador consiga entendê-las e crie regras que definam cada classe. As técnicas de pré-processamento, extração de características e a os classificadores são descritas nas subseções 2.1.3, 2.1.4, 2.1.5, respectivamente.

É importante ressaltar que a utilização de *Data Science* é voltada para problemas não óbvios: “*If a human expert can easily create a pattern in his or her own mind, it is generally not worth the time and effort of using data science to “discover” it.* (KELLEHER; TIERNEY, 2018)”.

### 2.1.3 Pré-processamento

Na presente Seção serão apresentadas as técnicas de pré-processamento que foram utilizadas para tratar as frases que compõem o *dataset*. É descrito a estruturação das frases em n-gramas (Subseção 2.1.3.1), conceito de *dataset* desbalanceado, as técnicas de *oversampling* e *undersampling* (Subseção 2.1.3.2) e a técnica de balanceamento *Data Augmentation and Pseudo Labelling* (Subseção 2.1.3.3). As técnicas de pré-processamento:

- Transformação das palavras para minúsculo
- Remoção de *stopwords*

são utilizadas em todos os cenários, não sendo essas as técnicas que se deseja avaliar o desempenho. *Stopwords* são dados que não possuem relevância para o modelo (e.g., 'o', 'os', 'para' e 'de').

#### 2.1.3.1 N-Gramas

N-Gramas são estruturas de dados, utilizadas em processamento de linguagem natural e mineração de texto, nas quais textos são representados como sequência de  $n$  palavras, letras, sílabas ou fonemas dependendo da abordagem (BRODER et al., 1997). Dado a seguinte frase: "Eu odeio muito Ciclano", para gerar o n-grama para  $n = 1$  (unigrama) basta separar a frase em palavras nomeadas *tokens*: "Eu", "odeio", "muito", "Clicano". Para gerar o bigrama ( $n = 2$ ) o

raciocínio é similar, porém, cada *token* é representado por duas palavras, a palavra atual e a seguinte (caso haja), como elucidado o exemplo: "Eu odeio", "odeio muito", "muito Ciclano". A geração de n-gramas para  $n > 1$  segue o mesmo modelo do bigrama.

### 2.1.3.2 Dataset desbalanceado

Um dataset é desbalanceado se a quantidade de amostras de determinada classe é muito diferente da quantidade de amostras de outra classe. A partir do momento que uma dada classe seja representada por poucas amostras, o classificador também terá menos regras que a definem (SUN; WONG; KAMEL, 2009). O desbalanceamento é visto em diferentes áreas de atuações: identificação de falsas chamadas telefônicas (EZAWA; SINGH; NORTON, 1996), classificação de texto (MLADENIC; GROBELNIK, 1999) e no reconhecimento de doenças (WOODS et al., 1993). As técnicas utilizadas no atual trabalho foram *Undersampling* e *Oversampling*. Ambas as técnicas visam igualar a quantidade de amostras de uma classe à quantidade de amostras da outra classe. Enquanto que *Undersampling* reduz a quantidade de amostras da classe mais representada para a quantidade de amostras da classe menos representada, *Oversampling* visa aumentar o número de amostras da classe menos representadas até se igualar à quantidade de amostras da outra classe. Esse aumento da classe menos representada é feito com as próprias amostras da classe de maneira aleatória (SUN; WONG; KAMEL, 2009).

#### 2.1.3.2.1 Data Augmentation and Pseudo Labelling

"Given a classification task, one may apply transformations to generate additional data and let the learning algorithm infer the transformation invariance (SIMARD et al., 2003)". Utilizando redes neurais convolucionais na tarefa de classificação de documentos visuais (SIMARD et al., 2003) propõem uma técnica de reamostragem do seu conjunto de treinamento utilizando *Oversampling*, entretanto, as imagens acrescidas, passaram por uma distorção. A distorção nas imagens foi feita movendo cada pixel da imagem original para uma nova posição, e assim, gerando uma nova imagem semelhante à imagem original. Essa distorção é uma operação invariante à classe, pois por mais que uma imagem se torne diferente, ela continua possuindo a mesma classificação de antes. Os autores constaram que a performance da sua rede melhorou. Baseada na técnica anterior, (KUMAR et al., 2018b) adequaram essas operações invariantes para a tarefa de classificação de texto. A partir de um conjunto de dados anotados na língua inglesa, os autores reamostraram o seu conjunto de dados utilizando o próprio conjunto após ser traduzido para quatro línguas intermediárias<sup>3</sup> e depois retraduzido para o inglês. Dessa forma, o conjunto de dados se tornou cinco vezes maior do que o conjunto inicial. Essa técnica é utilizada no proposto trabalho para avaliar se sua utilização também impacta positivamente, na avaliação do classificador, para a língua portuguesa.

<sup>3</sup> as línguas intermediárias foram: Hindi, Espanhol, Francês e Hindi

### 2.1.3.3 Extração de Radicais

A extração de radicais das palavras em classificação de texto é utilizada para reconhecer palavras cujas bases são as mesmas mas estão escritas de maneira distintas, como: "editaram" e "editou", que tem como radical a palavra "edit". Acredita-se que essa técnica melhorará nos resultados dos classificadores

### 2.1.4 Feature Selection

O problema de extração de características se resume a quais características de determinados dados são importantes para que se possa utilizá-las para treinar uma máquina eficientemente. Modelos de classificação de texto usualmente utilizam n-gramas (Subseção 2.1.5) para representar as características de frases presentes em um texto (PELLE; MOREIRA, 2017; SOUZA et al., 2016; KIILU GEORGE OKEYO, 2018). Algumas dessas características extraídas são irrelevantes ou podem desempenhar um papel negativo na classificação final de uma base de dados. Técnicas de *feature selection* visam remover essas características baseadas em alguns critérios, de maneira que, o classificador se torne mais preciso. (SHARMA; DEY, 2012) realizam um estudo comparativo de técnicas de *feature selection* aplicadas às técnicas de aprendizado de máquina NB, SVM e Maximum Entropy (ME) para o problema de análise de sentimento. Foi observado que SVM obteve uma melhor performance na análise de sentimento, entretanto, NB demonstra uma melhor performance quando utilizado com menos *features*.

#### 2.1.4.1 Teste de Chi Quadrado

O teste de chi quadrado, ou teste de chi quadrado de Pearson (PEARSON, 1900), é um teste estatístico que mede quão independente as variáveis são perante grupos distintos, basicamente se mede a diferença dos valores observados e dos valores esperados para cada variável. Primeiramente é necessário introduzir o conceito de tabela de contingência. Uma tabela de contingência pode ser descrita como uma matriz de tamanho  $x * y$ , de tal forma que cada elemento dessa matriz corresponde à frequência de  $x_i$  quando ocorrido simultaneamente com  $y_i$ . Um exemplo dessa tabela para o problema de análise de discurso de ódio pode ser visto abaixo (Esse exemplo não reflete as frequências reais do *dataset* utilizado):

Tabela 2.1 – Tabela de contingência para o problema de analisar discurso de ódio. Sim representa uma palavra classificada como discurso de ódio e Não o contrário. Essa tabela é composta por valores observados

	Sim	Não	Total
mãe políticos	6	2	8
Total	10	8	18

A frequência que a palavra “mãe” apareceu e a classificação atribuída foi Sim é de 6, 2



caso a classificação seja Não. O raciocínio anterior se aplica para todos os atributos (palavras). Calculando a frequência que a palavra “mãe” apareceu dentre as duas palavras é:

$$\frac{8}{18} = 0,44 \quad (2.1)$$

Agora pressupondo que as classificações sejam independentes, ou seja, que a frequência que a palavra “mãe” aparece independe de sua classificação, esperaria-se que a frequência que a palavra “mãe” apareça seja igual tanto para a classificação Sim quanto para Não. Abaixo, temos uma tabela de valores esperados, ainda não calculados para melhor visualização do problema:

Tabela 2.2 – Tabela 2.1 com os valores observados trocados pelas variáveis  $e_1 \dots e_4$  para se calcular os valores esperados

	Sim	Não	Total
mãe	$e_1$	$e_2$	8
políticos	$e_3$	$e_4$	10
Total	10	8	18

O cálculo dos valores esperados para classificações independentes da palavra “mãe” é descrito pela equação 2.2 abaixo:

$$\frac{e_1}{10} = \frac{e_2}{8} = \frac{8}{18} = 0,44 \quad (2.2)$$

da equação 2.2 temos que:

$$e_1 = \frac{(10 * 8)}{18} = 4,44 \quad (2.3)$$

A partir de  $e_1$  é possível descobrir o restante dos valores:

$$\begin{aligned} e_1 + e_2 &= 8 \Rightarrow e_2 = 3,56 \\ e_1 + e_3 &= 10 \Rightarrow e_3 = 5,56 \\ e_3 + e_4 &= 10 \Rightarrow e_4 = 4,44 \end{aligned} \quad (2.4)$$

Com os valores esperados conhecidos, é possível substituir as variáveis da Tabela 2.3 por seus respectivos valores:

Tabela 2.3 – Tabela 2.2 após a substituição das variáveis pelos valores esperados

	Sim	Não	Total
mãe	4,44	3,56	8
políticos	5,56	4,44	10
Total	10	8	18

Analisando a Tabela 2.1, de valores observados, e a Tabela 2.3, de valores esperados, é notável a diferença entre seus valores. A hipótese nula, como nomeia PEARSON, aplicada ao problema da análise de discurso de ódio, diz que as palavras “mãe” e “políticos” são independentes de suas classificações. Para confirmar essa hipótese é necessário calcular o teste de chi-square ( $X^2$ ), que baseia-se na distribuição de chi-quadrado calculada da seguinte forma:

$$X^2 = \sum_{i=1}^n \frac{(O^i + E^i)^2}{E^i} \quad (2.5)$$

onde n é a quantidade de elementos da tabela,  $O^i$  é o valor observado na posição  $i$  e  $E^i$  é o valor esperado na posição  $i$ . Para melhor entendimento a tabela abaixo demonstra os cálculos do teste de chi-quadrado.

Tabela 2.4 – Tabela dos cálculos de teste chi-quadrado

O	E	(O - E)	$(O - E)^2$	$\frac{(O-E)^2}{E}$
6	4,44	1,56	2,4336	0,5481
2	3,56	-1,56	2,4336	0,6835
4	5,56	-1,56	2,4336	0,4376
6	4,44	1,56	2,4336	0,5481
				$X^2 = 2,2173$

Para interpretar o valor de  $X^2$  é necessário, primeiramente, calcular o grau de liberdade da tabela de contingência. O cálculo do grau de liberdade (gl) é definido a partir da quantidade L (linhas) e da quantidade C (colunas) da tabela, excluindo a linha e a coluna do total de frequências, como exemplificado abaixo:

$$gl = (L - 1) * (C - 1) \quad (2.6)$$

Dessa maneira, tem-se que o grau de liberdade das tabelas 2.2, 2.3 e 2.4 é de 1, pois,  $(2 - 1) * (2 - 1) = 1$ . Para atestar ou não a hipótese nula é necessário verificar o valor do teste na tabela de porcentagens da distribuição de chi-square presente no Anexo A. Para grau de liberdade 1 e  $X^2 = 2,2173$ , a probabilidade das palavras "mãe" e "políticos" serem independentes é de aproximadamente 10%. Antes de atestar o teste, é preciso definir um grau de confiança que representa o nível de probabilidade que será aceito. Se o valor do teste for maior que o grau de confiança a hipótese é refutada, se for menor, ela será aceita. Para um grau de confiança de 5%, a hipótese nula se confirma, ou seja, as palavras "mãe" e "políticos" são dependentes das suas respectivas classificações.

O teste de chi-quadrado foi utilizado para a seleção de características do presente trabalho. Após o cálculo da frequência para cada palavra, sua frequência condicionada às classes e a quantidade de palavras, é calculado o valor de  $X^2$  para cada palavra. Após todas as palavras terem sido calculadas, a lista de resultadas é ordenada em ordem crescente e as melhores palavras são selecionadas para se realizar a extração de características. As melhores palavras são aquelas

cujo valor de  $X^2$  é o maior, pois são essas as palavras que possuem maior dependência perante às classificações. As quantidades de palavras na estrutura unigrama e bigrama para os dois *datasets* foram as mesmas utilizadas no baseline (PELLE; MOREIRA, 2017). As quantidades são: 250 e 426 para as estruturas unigrama e bigrama respectivamente quando utilizado a base de dados "OffComBR2" e 122 e 103 para as estruturas unigrama e bigrama respectivamente para a base "OffComBR3".

## 2.1.5 Classificação

A seguir serão descritos o funcionamento dos dois classificadores utilizados no presente trabalho: Naive Bayes (Subseção 2.1.5.1) e Support Vector Machine (Subseção 2.1.5.2)

### 2.1.5.1 Algoritmo de Naive Bayes

Naive Bayes é um algoritmo probabilístico baseado no Teorema de Bayes, logo, para compreender o algoritmo é necessário introduzir o conceito de probabilidade normal(priori), probabilidade condicional(posteriori) e Teorema de Bayes, respectivamente descritos nas subseções 2.1.5.1.1, 2.1.5.1.2, 2.1.5.1.3 e a aplicação do algoritmo na Subseção 2.1.5.1.4.

#### 2.1.5.1.1 Probabilidade normal

Probabilidade normal é a probabilidade de eventos acontecerem antes de outro evento ocorrido, por exemplo: Suponha uma caixa com duas bolas, uma preta e uma branca, a probabilidade de selecionarmos aleatoriamente uma bola na caixa é dado por  $P(B) = 1/2$ , ou seja, 50%. O mesmo raciocínio se mantém caso a caixa possua 3 bolas,  $P(B) = 1/3$ , ou seja, 33%.

#### 2.1.5.1.2 Probabilidade Condicional

Já na probabilidade condicional deseja-se saber a probabilidade de um evento acontecer dado que outro evento já tenha acontecido, por exemplo: Seja uma caixa A com duas bolas de cores distintas, 2 vermelhas(BV) e 3 amarelas(BA), a probabilidade de selecionar uma bola vermelha na caixa corresponde a:

$$P(BV/A) = \frac{\text{Vermelhas}}{\text{Total}} = \frac{2}{5} = 40\% \quad (2.7)$$

onde Vermelhas é a quantidade total de bolas vermelhas na caixa e Total é a quantidade total de bolas na caixa. Supondo agora que deseja-se calcular a probabilidade de escolher outra bola. A probabilidade desse exemplo é representada por  $P(B|BV)$  (lê-se probabilidade de B ocorrer dado que BV ocorreu), onde B é o evento correspondente a "pegar uma bola qualquer" e BV é o evento correspondente a "retirar uma bola vermelha". Nesse instante a quantidade de bolas total na caixa é 4, logo:

$$P(B|BV) = \frac{\text{Bola}}{\text{Total}} = \frac{1}{4} = 25\% \quad (2.8)$$

de modo que "Bola" representa uma bola aleatória e "Total" é a quantidade de bolas atual. Nesse exemplo é possível perceber que o segundo evento é dependente do primeiro.

### 2.1.5.1.3 Teorema de Bayes

O Teorema de Bayes relaciona probabilidade normal com probabilidade condicional para determinar o grau de credibilidade de uma determinada hipótese como define (REVISTABW, 2015), que por sua vez é representado pela seguinte equação:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \quad (2.9)$$

onde  $P(B)$  pode ser representado por:

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \quad (2.10)$$

e dessa maneira temos que:

$$P(A|B) = P(A) \frac{P(B|A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \quad (2.11)$$

e  $P(\bar{A})$  é a negação de  $P(A)$ . Assim sendo, os dados e o evento de B são considerados como sucessores de A, quando interpretado como uma regra para indução.  $P(A)$  pode ser chamada de probabilidade a priori, que será modificada pela experiência.  $P(A/B)$  é a probabilidade posterior, ou o nível de crença após a realização do experimento, isto é, revisada após a evidência obtida. Dado o seguinte cenário enunciado na (REVISTABW, 2015): Em uma eleição uma pessoa será selecionada para uma pesquisa de intenção de voto, supõe que a quantidades de mulheres e de homens é igual, portanto, 50% de probabilidade de escolher um dos dois. Ao iniciar a pesquisa o entrevistado cita ser desempregado. Dado que 85% dos homens estavam desempregados e 20% das mulheres estavam desempregadas deseja-se calcular a probabilidade de escolher uma pessoa do sexo masculino e que está desempregada.

- $P(A)$  = probabilidade da pessoa selecionada ser homem = 0.5
- $P(\bar{A})$  = probabilidade da pessoa selecionada ser mulher = 0.5
- $P(B|A)$  = 0.85 = probabilidade da pessoa estar desempregada dado que é um homem.
- $P(B|\bar{A})$  = 0.2 = probabilidade da pessoa estar desempregada dado que é uma mulher.

Aplicando os dados na equação temos que a probabilidade da pessoa ser homem e estar desempregada é:

$$P(A|B) = \frac{(0.85 * 0.5)}{(0.85 * 0.5) + (0.2 * 0.5)} = 80,95\% \quad (2.12)$$

### 2.1.5.1.4 Aplicação do algoritmo

As definições a seguir foram extraídas de (ZHANG, 2004). Assuma um documento  $X = x_1, x_2 \dots x_{10}$ , composto por dez *features*, considerados dados de treinamento, (Seção 3.2.2) e suas respectivas classes (duas no total), como mostra a Tabela 2.1. Considere agora classificar

Feature	contêm("carne")	contêm("anta")	contêm("políticos")	contêm("bosta")	Classe
x1	Verdadeiro	Falso	Falso	Falso	Não
x2	Verdadeiro	Falso	Verdadeiro	Falso	Não
x3	Falso	Verdadeiro	Falso	Falso	Sim
x4	Falso	Verdadeiro	Falso	Verdadeiro	Sim
x5	Falso	Falso	Verdadeiro	Verdadeiro	Não
x6	Verdadeiro	Falso	Falso	Falso	Não
x7	Verdadeiro	Falso	Falso	Verdadeiro	Sim
x8	Falso	Verdadeiro	Verdadeiro	Falso	Não
x9	Falso	Falso	Falso	Verdadeiro	Sim
x10	Falso	Falso	Verdadeiro	Falso	Sim

Tabela 2.5 – Tabela representando um documento X qualquer onde o atributo contêm("anta") significa a presença da palavra "anta" no texto original (antes da extração de características), "Sim" representa uma *feature* considerada como discurso de ódio e "Não" o oposto.

uma nova feature  $x_n$ , como discurso de ódio ou não, cujos atributos são:

("contêm(carne)": Verdadeiro); ("contêm(anta)": Falso); ("contêm(políticos)": Verdadeiro); ("contêm(bosta)": Verdadeiro)

O algoritmo de Naive Bayes busca prever se  $x_n$  é um discurso de ódio ou não através do cálculo da probabilidade a posteriori para cada classe pertencente ao conjunto de classes, a classificação de  $x_n$  será referente à classe que obtiver a maior probabilidade a posteriori. As probabilidades a posteriori para essa instância são representadas por:

$$P(\text{Sim}|x_n) = P(\text{Sim}) \frac{P(x_n|\text{Sim})}{P(x_n)} \quad (2.13)$$

$$P(\text{Não}|x_n) = P(\text{Não}) \frac{P(x_n|\text{Não})}{P(x_n)} \quad (2.14)$$

Entretanto,  $P(x_n)$  por ser uma constante pode ser descartada de ambas as equações (2.6) e (2.7):

$$P(\text{Sim}|x_n) = P(\text{Sim})P(x_n|\text{Sim}) \quad (2.15)$$

$$P(\text{Não}|x_n) = P(\text{Não})P(x_n|\text{Não}) \quad (2.16)$$

Logo, para calcular a probabilidade a posteriori, precisamos conhecer as probabilidades a priori:  $P(\text{Sim})$ ,  $P(\text{Não})$  e as probabilidades condicionais:  $P(x_n|\text{Sim})$  e  $P(x_n|\text{Não})$ . As probabilidades a

priori  $P(\text{Sim})$  e  $P(\text{Não})$  são triviais de se calcular:

$$P(\text{Sim}) = \frac{5}{10} = 0,50 = 50\% \quad (2.17)$$

$$P(\text{Não}) = \frac{5}{10} = 0,50 = 50\% \quad (2.18)$$

O classificador de Naive Bayes como sugerido pelo próprio nome, assume que os atributos das instâncias são independentes e, com isso, o cálculo de  $P(x_n|\text{Sim})$  e  $P(x_n|\text{Não})$ , representado na forma de atributos, pode ser escrito como o produto de seus atributos individuais:

$$P(\text{Sim}|a_1, \dots, a_4) = P(a_1|\text{Sim})P(a_2|\text{Sim})P(a_3|\text{Sim})P(a_4|\text{Sim}) \quad (2.19)$$

$$P(\text{Não}|a_1, \dots, a_4) = P(a_1|\text{Não})P(a_2|\text{Não})P(a_3|\text{Não})P(a_4|\text{Não}) \quad (2.20)$$

A representação de  $P(\text{Sim}|a_1, \dots, a_4)$ , ao substituir as variáveis  $a_1, \dots, a_4$  por seus respectivos valores, é:

$$\begin{aligned} &P(\text{contêm("carne") : Verdadeiro}|\text{Sim}) * P(\text{contêm("anta") : Falso}|\text{Sim}) * \\ &P(\text{contêm("políticos") : Verdadeiro}|\text{Sim}) * P(\text{contêm("bosta") : Verdadeiro}) \end{aligned} \quad (2.21)$$

Calculando as probabilidades condicionais:

$$\begin{aligned} P(\text{contêm("carne") : Verdadeiro}|\text{Sim}) &= \frac{1}{5} = 0,2 \\ P(\text{contêm("anta") : Falso}|\text{Sim}) &= \frac{3}{5} = 0,6 \\ P(\text{contêm("políticos") : Verdadeiro}|\text{Sim}) &= \frac{1}{5} = 0,2 \\ P(\text{contêm("bosta") : Verdadeiro}|\text{Sim}) &= \frac{3}{5} = 0,6 \end{aligned} \quad (2.22)$$

ou seja,  $P(a_1, \dots, a_4|\text{Sim}) = 0,2 * 0,6 * 0,2 * 0,6 = 0,014$ . O cálculo para a classe negativa segue o mesmo raciocínio, logo,  $P(a_1, \dots, a_4|\text{Não}) = 0,6 * 0,8 * 0,6 * 0,2 = 0,057$ . Agora que se tem as probabilidades a priori e condicionais, é possível calcular as probabilidades a posteriori:

$$P(\text{Sim}|x_n) = 0,5 * 0,014 = 0,008 \quad (2.23)$$

$$P(\text{Não}|x_n) = 0,5 * 0,057 = 0,029 \quad (2.24)$$

A classificação de determinada instância se dá a partir da maior probabilidade a posteriori dentre o conjunto de classes, e com isso, conclui-se que:  $x_n$  não é um discurso de ódio, pois  $P(\text{Não}|x_n) > P(\text{Sim}|x_n)$ , que é calculado como explicitado na Subseção 2.1.3.3.

### 2.1.5.2 Support Vector Machine (SVM)

Support Vector Machine objetiva separar um conjunto de dados, classificados como X ou Y, por um hiperplano em um espaço dimensional de tamanho  $n$  (BURGES, 1998), onde  $n$  é a quantidade total de *features*. O presente trabalho focará a discussão apenas no SVM linear, o caso mais simples, que demonstra ótimos resultados na classificação de texto (YANG; LIU et al., 1999).

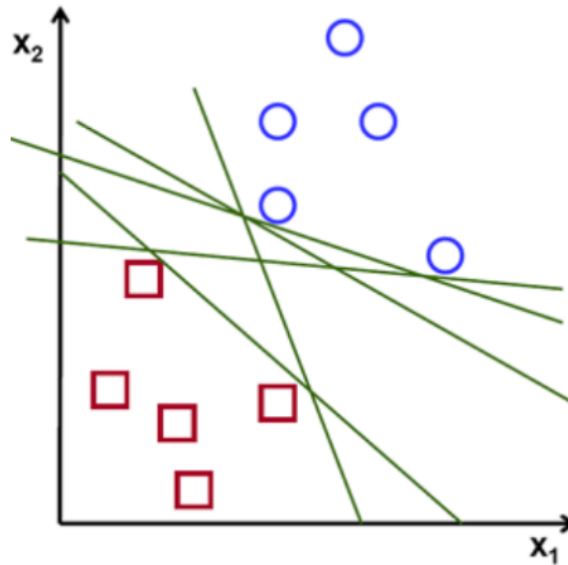


Figura 2.1 – Objetos separados por hiperplanos, perante suas classificações. Fonte: (DRAKOS, 2018)

A Figura 2.1 representa objetos distintos (círculos e quadrados) separados por hiperplanos (no caso do plano 2D um hiperplano é uma linha) de forma aleatória. SVM visa maximizar o tamanho da margem entre um separador de hiperplano e outro. A Figura 2.2 descreve como seria a separação por um hiperplano com maior margem, onde os objetos que estão coloridos são os separadores do hiperplano.

Suponha um conjunto de dados de treinamento  $\{x_1 \dots x_n\}$  e as respectivas classificações  $\{y_1 \dots y_n\} \in \{-1, 1\}$ . O problema de encontrar a maior margem dentre os separadores do hiperplano é um problema de otimização formulado como (LORENA; CARVALHO, 2007):

$$\text{Maximizar } \sum_{n=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (2.25)$$

$$\text{Restrições : } \begin{cases} \alpha_i \geq 0, & \forall i = 1, \dots, n \\ \sum_{n=1}^n \alpha_i y_i = 0 \end{cases} \quad (2.26)$$

sendo que  $\alpha_i$  é um parâmetro que varia de  $i$  até  $n$ .

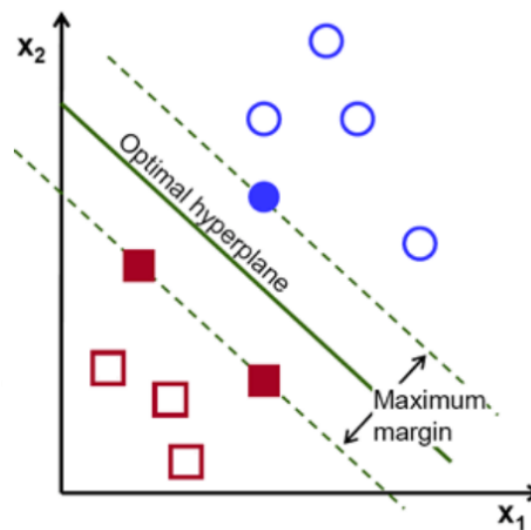


Figura 2.2 – Hiperplano ótimo. Fonte: (DRAKOS, 2018)

### 2.1.6 Métricas de Avaliação dos Resultados

No processo de classificação de discurso de ódio foram utilizadas quatro métricas para avaliar a precisão dos classificadores utilizados, sendo essas: *precision*, *recall* e *f-measure*. Essas métricas dependem de quatro variáveis para realizarem a medição como demonstra (SANTANA, 2017):

- *True Positive* (TP) = representa uma classificação correta da classe positiva, e.g., um classificador classifica uma frase, predefinida positiva, como positiva.
- *True Negative* (TN) = representa uma classificação correta da classe negativa, e.g., um classificador classifica uma frase, predefinida negativa, como negativa.
- *False Positive* (FP) = representa uma classificação errada da classe positiva, e.g., um classificador classifica uma frase, predefinida positiva, como negativa.
- *False Negative* (FN) = representa uma classificação errada da classe negativa, e.g., um classificador define uma frase, predefinida negativa, como positiva.

sendo que, a classe positiva representa um texto classificado como discurso de ódio e a classe negativa o oposto. *Precision* é o número de resultados corretos dividido pelo pela soma de verdadeiros positivos e falsos positivos, ou seja, pelo número total de elementos identificados como positivos. A equação pode ser vista a seguir:

$$\frac{TP}{TP + FP} \quad (2.27)$$

Já *recall*, que é o número de resultados corretos dividido pela soma de verdadeiros positivos e falsos negativos, ou seja, a quantidade de elementos identificadas como uma classe que realmente



pertencem a essa classe. A equação que representa *recall* é:

$$\frac{TP}{TP + FN} \tag{2.28}$$

Dado uma base de dados fictícia, na qual 7 frases são classificadas como discurso de ódio e 5 não. Caso o classificador identifique apenas 6 das frases como discurso de ódio e, das 6, apenas 2 são realmente discurso de ódio, então 2/6 é o valor de *precision* e 2/12 o valor de *recall* para a classe positiva. O raciocínio se mantém para a classe negativa. *f-measure* é a média harmônica de *precision* e *recall* representada por:

$$2 \times \frac{(precision \times recall)}{(precision + recall)} \tag{2.29}$$

É importante citar que alguns trabalhos como (PANG; LEE; VAITHYANATHAN, 2002; SOUZA et al., 2016) utilizam a métrica acurácia para avaliar seus modelos. Essa métrica é definida como:

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{2.30}$$

Por mais que esses trabalhos utilizem essa métrica, no presente trabalho a acurácia não foi utilizada pois o baseline avalia o modelo somente com a métrica *f-measure*.

### 2.1.7 K-Fold Cross Validation

*K-fold cross validation* é um método de reamostragem de dados utilizado para avaliar um modelo de *machine learning* perante sua capacidade de generalização, quando se tem uma base de dados limitada. O processo de avaliação se consiste em dividir o *dataset* de treinamento em K partições (BROWNLEE, 2018). Em cada iteração, K - 1 partições (*folds*), são utilizados para treinamento e uma partição é utilizado para avaliação, como demonstra a Figura 2.1:

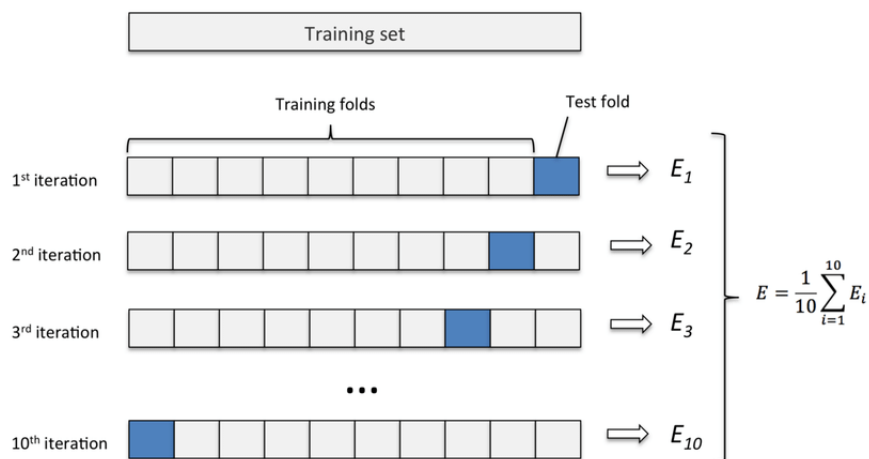


Figura 2.3 – K-fold Cross Validation. Fonte: (ASHFAQUE; IQBAL, 2019)

A cada iteração o modelo é avaliado segundo as métricas definidas, no presente trabalho é utilizado a métrica *f-measure* (descrita na Subseção 2.1.6), e os resultados armazenados. Ao

final das  $K$  iterações é calculado uma média aritmética dos resultados anotados. Esse método é comumente encontrado em trabalhos de classificação de texto (PELLE; MOREIRA, 2017; PANG; LEE; VAITHYANATHAN, 2002; SOUZA et al., 2016) e, na sua maioria, utilizam  $K = 10$  como parâmetro para a avaliação.

## 2.2 Trabalhos Relacionados

O trabalho (PANG; LEE; VAITHYANATHAN, 2002) realiza um estudo de análise de sentimento utilizando três classificadores diferentes, *Naive Bayes* (NB), *Support Vector Machines* (SVM) e *Maximum Entropy* (ME). Nesse trabalho, os dados são representados por segmentos de sentenças do tipo unigrama ou bigrama. O autor conclui que os três algoritmos estão muito próximos em acurácia, entretanto, SVM tende a ser melhor e a abordagem por unigrama obteve o melhor resultado de 82,9% de acurácia. Por outro lado, os autores não conseguem obter precisão no problema de classificação de sentimentos comparável àqueles relatados para categorização baseada em tópicos.

(SOUZA et al., 2016) utilizam os mesmos algoritmos de (PANG; LEE; VAITHYANATHAN, 2002) mas utilizando o conceito de *bag-of-words* somado com *Term Frequency* (TF) e *Inverse Document Frequency* (IDF) representados em conjunto como TF-IDF. *Bag-of-words* é a representação de dados composta por listas de frequências de *tokens* presentes em cada texto. *Term frequency* refere-se a medida estatística que define quão importante é uma palavra através de sua frequência em um determinado texto. Já *Inverse document frequency* atribui pesos maiores para palavras que são raramente utilizadas. TF-IDF é a união desses dois últimos conceitos, que atribui pesos maiores para palavras que aparecem raramente nos documentos (SILGE, 2018). Os autores conseguiram uma acurácia de 93% utilizando a abordagem de *bag-of-words* com TF-IDF quando utilizado com o classificador SVM para o problema de análise de polaridade nos tweets, que consiste em determinar uma sentença como positiva ou negativa, referentes à abertura do processo de impeachment da presidente do Brasil.

(KIILU GEORGE OKEYO, 2018) para a tarefa de analisar sentimentos no Tweeter, pré-processam tweets utilizando uma técnica adicional chamada de *Pos Tag*, que se consiste em atribuir etiquetas morfológicas aos *tokens* dos n-gramas, por exemplo, após aplicar essa técnica no unigrama:

["Eu", "realmente", "odeio", "o", "Ciclano"]

obtem-se:

[('Eu', 'PRP'), ('realmente', 'RB'), ('odeio', 'JJ'), ('o', 'DT'), ('Ciclano', 'NNP')]

onde

- PRP representa um pronome.

- RB representa um advérbio.
- JJ representa um adjetivo.
- DT representa um determinador.
- NNP representa um nome próprio.

Os autores obtiveram acurácias de 64.47% utilizando unigramas e 70% com bigramas, para o classificador NB.

Com o foco na língua portuguesa, (PELLE; MOREIRA, 2017) analisam discurso de ódio presente no portal G1<sup>4</sup>. O autor apresenta uma base de dados anotada com comentários classificados como ofensivos ou não. Os autores ainda fornecem resultados dos algoritmos de *Support Vector Machine* e *Naive Bayes* com os dados estruturados em unigramas, bigramas e trigramas, aplicados à essa base de dados mencionada. (PELLE; MOREIRA, 2017) disponibilizam um *dataset* de comentários na língua portuguesa. Esses comentários estão devidamente rotulados como discurso de ódio ou não. O *dataset* criado por PELLE; MOREIRA é utilizado como baseline do presente trabalho, entretanto, utilizando outras técnicas de pré-processamento e seleção de características.

(MONDAL et al., 2018) classificam discursos de ódio em duas redes sociais (Twitter e Whisper) com o auxílio do Hatebase<sup>5</sup>, um repositório *crowdsourced online* de conteúdo estruturado, multilíngue e baseado em palavras de ódio. Para tal, utiliza a estratégia de procurar sentenças da forma "I<intensity><userintent><hatetarget>", onde "I" é o sujeito que escreve, "intensity" representa uma amplificação de suas expressões, "userintent" é a componente que representa o ódio e "hatetarget" é o grupo vítima do discurso de ódio. O autor caracteriza o discurso de ódio em quatro dimensões: os alvos principais dos discursos, o conteúdo expresso, correlação com o anonimato e a geografia dos atacantes. O autor ainda encontra uma relação de expressões de ódio online com dados registrados sobre crimes de ódio nas várias regiões dos Estados Unidos. Foi implementada uma tentativa de classificar as sentenças escritas em português utilizando essa técnica, contudo, não foram obtidos bons resultados já que nas postagens analisadas observou-se que raramente as frases escritas em português são assim estruturadas dessa forma tão direta (sujeito, advérbio, verbo, objeto).

(RIBEIRO et al., 2018), com o auxílio de um *dataset*, de sua autoria, pré-classificado com usuários odiosos e não odiosos, modelam o problema de detectar discurso de ódio como um problema de grafos com aprendizado de máquina. Os autores procuram por usuários odiosos através de um grafo de retweets que cada usuário possui, através desse grafo ele descobre se um usuário, ainda não classificado, é um possível *hater*, caso, esse tenha retweetado um tweet

<sup>4</sup> <https://g1.globo.com/>, Accessed: 2018-08-15

<sup>5</sup> <http://www.hatebase.org/>, Accessed: 2018-10-15

de um usuário odioso. Uma acurácia de 90.9% e *f-measure* de 67% é obtida através do modelo "graphSage".

(KUMAR et al., 2018a) realizam um estudo comparativo entre diversos modelos, que participaram da *Shared Task on Aggression Identificatio* (STAI), para identificar agressões nas mídias sociais na língua inglesa e hindi. Foi observado que tanto modelos que utilizam redes neurais quanto abordagens lineares possuem pouca diferença de performance na tarefa de identificar textos de cunho odiosos. Dentre os quinze melhores modelos, apenas um utiliza da abordagem léxica, uma abordagem que visa comparar cada palavra, que possui seu próprio peso no texto, com palavras registradas em um dicionário de palavras negativas. (AROYEHUN; GELBUKH, 2018) conseguiram a melhor performance na STAI com o modelo *Long Short Term Memory networks* (LSTM) de redes neurais, que são utilizadas em *machine learning*, o autor decodifica *emoji* em sua representação textual e utiliza a técnica de *pseudo labelling* (Subseção 2.1.8), que se baseia na tradução de palavras para outro idioma e depois na retradução dessa palavra traduzida para o idioma original (inglês), para aumentar o tamanho do *dataset*. O autor obteve melhores resultados com a utilização da técnica de *pseudo labelling* para todos os modelos testados em seu estudo.

A solução proposta no presente trabalho visa produzir resultados comparativos, com base nos resultados obtidos (PELLE; MOREIRA, 2017) e, para tal, utiliza do mesmo *dataset* utilizado e criado pelo autor, além de outras técnicas de pré-processamento e classificação.

## 3 Desenvolvimento

Neste capítulo será apresentado o método utilizado no presente trabalho (Seção 3.1), os *datasets* utilizados (Subseção 3.1.1), a descrição de como foi feito o pré-processamento (Subseção 3.1.2), os passos para se extrair as características das palavras (Subseção 3.1.3), os experimentos realizados (Subseção 3.1.4) e os resultados alcançados (Seção 3.2)

### 3.1 Método

A arquitetura do presente método consiste no fluxograma representado na Figura 3.1. Cada uma das atividades do fluxograma são descritas em mais detalhes nas subseções seguintes.

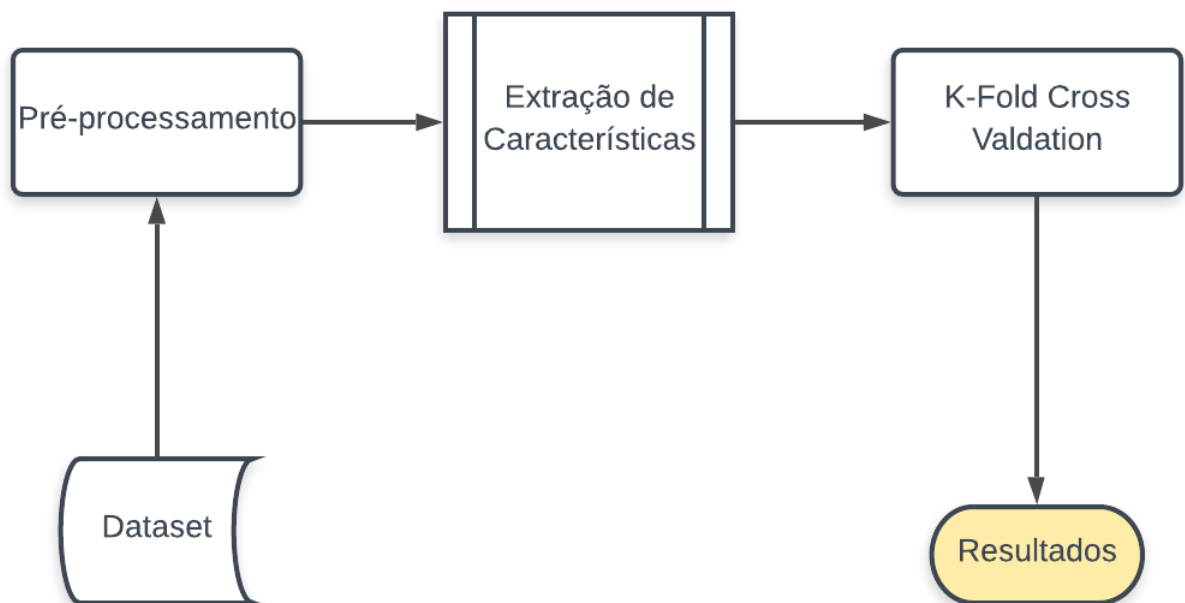


Figura 3.1 – Arquitetura do método apresentado

#### 3.1.1 Dataset

No presente trabalho foi utilizado os *datasets* 'OFFCOMBR-2' e 'OFFCOMBR-3'<sup>1</sup> criados por (PELLE; MOREIRA, 2017), que são compostos por comentários retirados do G1<sup>2</sup>. Os comentários foram analisados por 3 juízes que classificaram cada comentário como sendo ofensivo ou não. No OFFCOMBR-2, 419 de 1.250 comentários foram classificados como ofensivo

<sup>1</sup> Link para as bases <http://inf.ufrgs.br/rpelle/hatedetector/>

<sup>2</sup> <https://g1.globo.com/>

se e só se 2, de 3 juízes, concordassem, já no OFFCOMBR-3, 202 de 1.033 comentários foram classificados como ofensivos se e só se os 3 juízes concordassem. Foram analisados 1.250 comentários no primeiro *data set* dos quais 419 foram especificados como discurso de ódio. Um exemplo de como os dados são dispostos no *data set* pode ser vista a seguir:

- sim, Fica ai Janete traste inutil Destruiu o brasil vaca;
- não, o Brasil ta quebrado e temer esta emprestando bilhoes ao FMI;
- não, Nao a pizza anunciada de anistia do caixa Sim a divulgacao da lista da Odebrecht;
- não, NAO ESTAO NAO SENAO RECRUTARIAM A TUA MAE SEU VERME;

de maneira que sim é a classificação atribuída a um comentário ofensivo e não o oposto.

### 3.1.2 Pré-processamento

O pré-processamento é composto pelos seguintes passos:

- remoção de *stopwords*;
- transformação das palavras para minúsculo;
- extração ou não dos radicais;
- transformação das palavras em ngramas;
- balanceamento ou não, utilizando; *undersampling*

É importante salientar que, como o proposto trabalho visa analisar o impacto que a extração dos radicais ou o balanceamento por *undersampling* tem na tarefa de classificação de discurso de ódio, essas técnicas podem ser utilizadas ou não. As palavras foram estruturadas em unigramas ou unigramas + bigramas.

### 3.1.3 Extração de Características

A extração de características é uma forma de fazer com que o classificador entenda cada frase como um conjunto de características particulares das frases perante o conjunto de frases. Após a fase de pré-processamento, é criado um dicionário para todas as frases estruturadas como unigramas e um dicionário para todas as frases estruturas como unigramas + bigramas. A partir desse dicionário são criadas *features* para cada ngrama. O conjunto de *features* de uma frase possui uma *feature* para cada ngrama do dicionário que está presente na frase em questão, seguido por sua classificação. Por exemplo.: Dado duas frases estruturadas em unigramas seguidas por suas classificações: [(hoje), (eu), (odeio), (muito), (fulano), sim] e [(ta), (muito), (quente), (hoje),

não]. Um dicionário para esse conjunto ngramas é composto por todos os ngramas sem repetição: [eu, odeio, muito, fulano, ta, quente, hoje]. A partir desse dicionário, as *features* das frases serão representadas como:

Frase 1: [{"hoje": Verdadeiro, "eu": Verdadeiro, "odeio": Verdadeiro, "muito": Verdadeiro: "fulano": Verdadeiro, "ta": Falso, "quente": Falso}, sim]

Frase 2: [{"hoje": Verdadeiro, "eu": Falso, "odeio": Falso, "muito": Verdadeiro: "fulano": Falso, "ta": Verdadeiro, "quente": Verdadeiro}, não]

A extração de características é feita tanto na estrutura que utiliza somente unigramas quanto na estrutura que utiliza unigramas + bigramas, naturalmente, essa última estrutura possuirá um dicionário maior de palavras, por conter tanto unigramas quanto bigramas.

Quando é utilizada a seleção de características, sua aplicação ocorre antes da extração de características. Os melhores ngramas são selecionados utilizando o teste chi-quadrado (Subsubseção 2.1.9.1) e um novo dicionário é criado a partir desses ngramas. Após a criação do dicionário a extração de características ocorrerá assim como ocorre no exemplo acima.

### 3.1.4 Experimentação

Após a extração de características de cada dado pré-processado do *dataset*, as *features* totais são particionadas seguindo o processo de *k-fold cross validation* (Seção 2.1.7) para  $k = 10$ . Dessa forma os dados de treinamento e de teste terão tamanho de 9/10 e 1/10 do tamanho total do *dataset* respectivamente, para cada iteração. Ainda que o tamanho dos dados de teste e de treino não variem, para cada iteração da validação cruzada o conjunto de treino e de teste alteram seu conteúdo. Para cada iteração  $k$  do processo, os dados de treinamento passaram por um processo de *Oversampling com Data Augmentation e Pseudo Labelling* (Subseções 2.1.3.2 e 2.1.3.3, respectivamente), ou seja, os dados de treinamento são aumentados com os próprios dados após traduzidos para a língua alemã e depois retraduzidos para a língua portuguesa. Foi utilizado traduções só para uma língua quando o *dataset* utilizado fora o OffComBR2 porque o mesmo possui um desbalanceamento de 2:1 de frases positivas para negativas. Já no OffComBR3 fora utilizado traduções para três línguas diferentes, sendo elas: Alemão, Inglês e Espanhol, pois nesse caso, o desbalanceamento possui uma proporção de 4:1 de frases positivas para negativas. Algumas frases traduzidas e retraduzidas são idênticas, outras introduzem novos sinônimos para algumas palavras e outras, após a retradução, retornam com alguma palavra não traduzida. Desse modo é importante notar que essa última técnica deve ser implementada durante o *k-fold cross validation*, nas *features* de treino, para que não seja possível um dado de treino ser exatamente igual a um dado de teste durante as iterações, o que prejudica na avaliação do classificador.

Em cada iteração  $k$  da validação cruzada, o classificador é treinado e a *f-measure* calculada para as classes positivas e negativas. Ao final da iteração, é feita uma média ponderada (o peso é o tamanho da classe) entre as *f-measures* das duas classes. Ao fim das  $k$  iterações, é realizado

uma média aritmética dos resultados armazenados. O resultado final é a média aritmética das médias ponderadas retornadas para cada validação cruzada. O modelo foi avaliado perante as seguintes técnicas:

- Utilização do radical das palavras na geração dos *tokens* (Subseção 2.1.3.4),
- Seleção de Características por chi quadrado (Subseção 2.1.4.1),
- *Undersampling* (Subseção 2.1.3.4),
- *Data augmentation and Pseudo Labelling* (Subseção 2.1.3.4.1).

Para a extração dos radicais foi utilizado a função “RSLPStemmer()” da nltk, uma função própria para extrair radicais da língua portuguesa. A função “BigramAssocMeasures.chi\_sq()” é responsável por retornar as melhores palavras do dicionário, a partir da frequência normal e condicional daquela palavra para determinada classe. A técnica de *Undersampling* foi utilizada para balancear o *dataset*, reduzindo a quantidade de palavras positivas para que se igualem à quantidade das negativas. As traduções foram feitas utilizando a API de tradução de textos da Microsoft Azure <sup>3</sup>

## 3.2 Resultados

Nesta seção, são mostrados os resultados das técnicas aplicadas ao modelo e seus resultados comparados com os resultados do baseline. A Figura 3.2 representa os resultados alcançados pelo baseline:

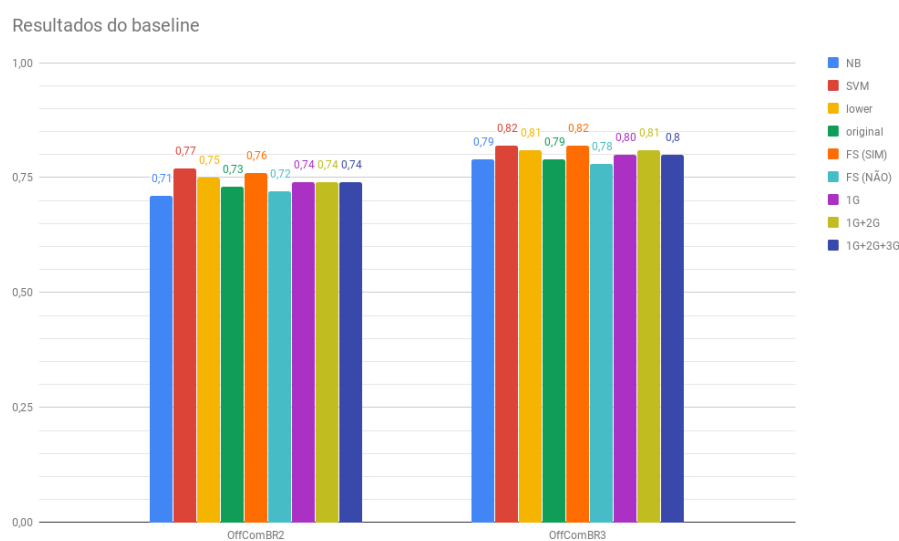


Figura 3.2 – Gráfico dos resultados alcançados pelo baseline

<sup>3</sup> <https://docs.microsoft.com/en-us/azure/cognitive-services/translator/reference/v3-0-reference>, Accessed: 2019-04-19



(PELLE; MOREIRA, 2017) avaliam os classificadores utilizando o *k-fold cross validation* com  $k = 10$ . Em cada iteração, os autores calculam uma medida *f-measure* para ambas as classes e armazena a média ponderada entre essas medidas, considerando o tamanho da classe como peso. Ao final das  $k$  iterações, o autor calcula a média aritmética de todas as médias ponderadas armazenadas. Os resultados do baseline foram obtidos a partir da aplicação das técnicas e da classificação utilizando a ferramenta weka<sup>4</sup>. Para a extração de características, os autores utilizam a abordagem de *bag-of-words*, que se trata de uma abordagem similar à utilizada nesse trabalho, a diferença é que ao invés de cada *feature* possuir um valor de Sim ou Não, indicando a presença da palavra, é utilizado a frequência de cada palavra dentro da frase.

As colunas da Figura 3.2 representam a média das *f-measures* através de todas as iterações da validação cruzada dado uma determinada técnica. Por exemplo, a coluna “lower” demonstra os resultados do classificador quando no pré-processamento dos dados as palavras são convertidas para minúsculo considerando ambos os classificadores (média dos resultados de ambos classificadores). As colunas NB e SVM representam os resultados dos classificadores sem a aplicação de nenhuma das técnicas de pré-processamento ou seleção de características e a coluna FS (SIM) representa os resultados da classificação após utilizar a seleção de características utilizando *Information Gain* implementada no weka para a classe sim e FS (NÃO) para a classe não. As colunas 1G, 1G+2G e 1G+2G+3G demonstram os resultados para a classificação quando os dados são estruturados como unigramas, unigramas + bigramas e unigramas + bigramas + trigramas respectivamente. Por falta de clareza de (PELLE; MOREIRA, 2017), não é possível saber com exatidão como ele encontrou tais resultados, por exemplo, na coluna “lower”, só é explicado por PELLE; MOREIRA que essa coluna representa os resultados de ambos os classificadores (NB e SVM), mas não deixa claro qual estrutura (n-grama) ele utilizou para os mesmos.

Ao contrário do baseline, os resultados de cada técnica do presente trabalho são mostrados individualmente para cada estruturação e escolha do classificador. Acredita-se que dessa maneira os resultados são mais claros. Entretanto, não é possível comparar com exatidão os resultados deste trabalho com os do baseline. A Figura 3.3 mostra os resultados obtidos ao se utilizar o *dataset* OffComBR2 e todas as combinações das técnicas de extração de radicais, *undersampling*, *data augmentation* e *feature selection*.

A remoção de *stopwords* e transformação das palavras para minúsculo acontecem em todos os cenários. A linha “NORMAL” representa os resultados sem a utilização de nenhuma técnica. As linhas “RA”, “FS”, “US” e “DA” representam a utilização da técnica de extração de radicais, *feature selection*, *undersampling* e *data augmentation* respectivamente. “RA+FS” representa a utilização da extração de radicais em conjunto com *feature selection*.

Dos resultados apresentados, para o *dataset* OffComBR2, a combinação das técnicas de seleção por características e *undersampling* mostraram as melhores *f-measures*, alcançando 91% quando o classificador utilizado fora o NB e a estrutura utilizada fora unigrama + bigrama.

<sup>4</sup> <https://www.cs.waikato.ac.nz/ml/weka/>

OffComBR2				
	NB		SVM	
	Unigrama	Unigrama+Bigrama	Unigrama	Unigrama+Bigrama
NORMAL	0,78	0,79	0,76	0,74
RA	0,78	0,79	0,77	0,77
FS	0,86	0,85	0,84	0,83
US	0,75	0,74	0,70	0,70
DA	0,73	0,74	0,72	0,71
RA+FS	0,84	0,85	0,81	0,81
RA+US	0,72	0,71	0,76	0,74
RA+DA	0,72	0,73	0,70	0,72
FS+US	<b>0,90</b>	<b>0,91</b>	<b>0,86</b>	<b>0,85</b>
FS+DA	0,83	0,80	0,76	0,75
RA+FS+DA	0,77	0,74	0,74	0,66
RA+FS+US	0,86	0,88	0,83	0,84

Figura 3.3 – Resultados do presente trabalho utilizando a base de dados OffComBR2 e diferentes combinações das técnicas utilizadas

Independente da estrutura ou técnica, o classificador de NB se saiu melhor do que o SVM, um comportamento oposto daquele apontado pelo baseline. Nota-se que as técnicas de balanceamento, quando utilizadas exclusivamente, possuem uma *f-measure* inferior à ausência de técnicas. Na maioria dos casos, quando houve extração de radicais, houve uma piora nos resultados.

A Figura 3.4, demonstra os resultados obtidos utilizando a base de dados OffComBR3:

OffComBR3				
	NB		SVM	
	Unigrama	Unigrama+Bigrama	Unigrama	Unigrama+Bigrama
NORMAL	0,80	0,80	0,79	0,79
RA	0,83	0,82	0,80	0,80
FS	0,82	0,82	0,79	0,81
US	0,67	0,69	0,66	0,64
DA	0,76	0,75	0,75	0,74
RA+FS	0,83	0,81	0,81	0,78
RA+US	0,69	0,68	0,69	0,70
RA+DA	0,77	0,76	0,78	0,77
FS+US	0,89	<b>0,89</b>	0,83	<b>0,84</b>
FS+DA	0,74	0,77	0,67	0,69
RA+FS+DA	0,68	0,75	0,64	0,66
RA+FS+US	<b>0,91</b>	0,85	<b>0,89</b>	0,83

Figura 3.4 – Resultados do presente trabalho utilizando a base de dados OffComBR3

Os resultados para o *dataset* OffComBR3 também escalaram como no baseline. Como o baseline utiliza as médias dos classificadores para anotar as medidas *f-measure*, não é possível saber o resultado de cada técnica aplicada a um classificador específico. Ao contrário dos resultados utilizando o OffComBR2, a extração de radical demonstrou resultados positivos para ambos

classificadores, estruturas e combinações de técnicas. A seleção de características, de maneira geral, utilizando a base OffComBR3, teve resultados um pouco melhores que o "NORMAL", entretanto não escalou como na OffComBR2. *Undersampling* e *data augmentation* continuaram entre as piores técnicas, com resultados inferiores ao *default*.

### 3.2.1 Discussão dos Resultados

Nesta Subseção são analisados e discutidos os resultados obtidos. Os resultados obtidos, em quase todos os cenários, foram superiores àqueles apresentados por PELLE; MOREIRA. A utilização exclusiva das técnicas de balanceamento resultaram em uma piora nos resultados, entretanto, a utilização do *undersampling* em conjunto com a extração de radicais e/ou *feature selection* demonstraram os melhores resultados, chegando a 91% de *f-measure*. Acredita-se que os resultados do presente trabalho foram superiores ao baseline porque (PELLE; MOREIRA, 2017) não remove *stopwords* e extrai as características utilizando *bag-of-words*, o que sugere uma redução na capacidade de classificação dos classificadores. No presente trabalho foi utilizado um SVM cuja implementação foi baseada na biblioteca liblinear (FAN et al., 2008), já o baseline utiliza um SVM baseado em (PLATT, 1998), dessa forma, é plausível dizer que a utilização de duas implementações diferentes de SVM e abordagens para extração de características distintas podem gerar resultados distintos para a análise de discurso de ódio.

Analisando os resultado do presente trabalho, percebe-se que os resultados da base "OffComBR2" são um pouco inferiores à base "OffComBR3", o que é visto também por PELLE; MOREIRA. A utilização de *undersampling* com seleção de características e a utilização das duas com extração de radical se mostraram as combinações mais promissoras, em ambas as bases. Não é conclusivo se a utilização de radicais ao invés da palavra inteira é uma técnica boa ou ruim, acredita-se que ao utilizar uma base de dados maior, o comportamento seja mais visível. A utilização de unigrama ou unigramas + bigramas demonstraram resultados muito parecidos, sendo assim, ambas as abordagens geram resultados bons.

Foi observado que palavras diferentes podem representar tanto um discurso de ódio como não, e.g. "você" e "não" foram palavras que apareceram tanto em uma classe quanto em outra. Algumas palavras foram perdidas durante o pré-processamento por serem jargões ou por não apresentarem acentuação, como na palavra "sapatao" que foi classificada como discurso de ódio mas foi considerada como *stopword* pela função implementada pela plataforma NLTK<sup>5</sup>.

<sup>5</sup> <https://www.nltk.org/>, Accessed: 2018-10-11

## 4 Considerações Finais

Nesse capítulo será apresentado a conclusão do trabalho (Seção 5.1) e os trabalhos futuros (Seção 5.2).

### 4.1 Conclusão

Com o aumento da quantidade de usuários ao longo dos anos, as interações por meio de comentários e postagens consequentemente também aumentaram. Essas postagens e comentários, quando em teor de ódio, podem ser classificados como discurso de ódio. Inúmeros trabalhos vem produzindo estudos na detecção desses discursos de forma automática (computador) com a utilização do aprendizado de máquina, entretanto, para a máquina aprender com os dados é necessário que esses já estejam pré-classificados. Preenchendo a lacuna de um *dataset* anotado em português, (PELLE; MOREIRA, 2017) fornece duas bases de dados anotadas, que foram utilizadas na implementação do método do presente trabalho.

Neste estudo foi utilizados combinações de técnicas de detecção de discursos de ódio existentes na literatura. De maneira geral, os resultados foram superiores aos demonstrados no *baseline*(PELLE; MOREIRA, 2017). Conclui-se que os resultados obtidos utilizando a técnica de seleção de característica, exclusivamente ou combinada, gera as melhores estimativas. A não remoção de *stopwords* pelo *baseline* e a utilização de *bag-of-words* na representação de características das frases sugerem que são abordagens piores. Não foi possível concluir se a utilização dos radicais das palavras foi relevante para a classificação, acredita-se que com uma base de dados maior será possível ver uma melhora ou piora real. A utilização do balanceamento por *undersampling* só se mostrou positiva quando utilizada em conjunto com outras técnicas, enquanto que, o balanceamento por *data augmentation and pseudo labelling* é ruim em todos os cenários. Conclui-se que o classificador de Naive Bayes possui resultados superiores àqueles utilizando o SVM.

### 4.2 Trabalhos Futuros

Como trabalhos futuros será realizado testes utilizando novas técnicas e classificadores para a categorização dos discursos de ódio como machismo, homofobia, racismo, dentre outras formas de discriminações.

# Referências

- AGARWAL, S.; SUREKA, A. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In: *Distributed Computing and Internet Technology*. Springer, 2015. ISBN 978-3-319-14976-9. Disponível em: <[http://dx.doi.org/10.1007/978-3-319-14977-6\\_47](http://dx.doi.org/10.1007/978-3-319-14977-6_47)>.
- ALMEIDA, T.; NAKAMURA, F.; NAKAMURA, E. Uma abordagem para identificar e monitorar haters em redes sociais online. In: *Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres*. [S.l.: s.n.], 2017. p. 41–46.
- AROYEHUN, S. T.; GELBUKH, A. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. [S.l.: s.n.], 2018. p. 90–97.
- ASHFAQUE, J.; IQBAL, A. Introduction to support vector machines and kernel methods. 04 2019.
- BARTLETT, J.; REFFIN, J.; RUMBALL, N.; WILLIAMSON, S. Anti-social media. *Demos*, p. 1–51, 2014.
- BRODER, A. Z.; GLASSMAN, S. C.; MANASSE, M. S.; ZWEIG, G. Syntactic clustering of the web. *Computer networks and ISDN systems*, Elsevier, v. 29, n. 8-13, p. 1157–1166, 1997.
- BROWNLEE, J. *A Gentle Introduction to k-fold Cross-Validation*. 2018. Disponível em: <<https://machinelearningmastery.com/k-fold-cross-validation/>>.
- BURGES, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, Springer, v. 2, n. 2, p. 121–167, 1998.
- DAVIDSON, T.; WARMSLEY, D.; MACY, M.; WEBER, I. Automated hate speech detection and the problem of offensive language. In: *Eleventh international aai conference on web and social media*. [S.l.: s.n.], 2017.
- DJURIC, N.; ZHOU, J.; MORRIS, R.; GRBOVIC, M.; RADOSAVLJEVIC, V.; BHAMIDIPATI, N. Hate speech detection with comment embeddings. In: *ACM. Proceedings of the 24th international conference on world wide web*. [S.l.], 2015. p. 29–30.
- DRAKOS, G. *Support Vector Machine vs Logistic Regression*. 2018. <<https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f>>. Accessed: 2019-07-2.
- EZAWA, K. J.; SINGH, M.; NORTON, S. W. Learning goal oriented bayesian networks for telecommunications risk management. In: *ICML*. [S.l.: s.n.], 1996. p. 139–147.
- FAN, R.-E.; CHANG, K.-W.; HSIEH, C.-J.; WANG, X.-R.; LIN, C.-J. Liblinear: A library for large linear classification. *Journal of machine learning research*, v. 9, n. Aug, p. 1871–1874, 2008.
- GITARI, N. D.; ZUPING, Z.; DAMIEN, H.; LONG, J. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, v. 10, n. 4, p. 215–230, 2015.

- KELLEHER, J. D.; TIERNEY, B. *Data Science*. [S.l.]: The MIT Press, 2018.
- KIILU GEORGE OKEYO, R. R. K. O. K. K. Using naïve bayes algorithm in detection of hate tweets. *International Journal of Scientific and Research Publications (IJSRP)*, v. 8, n. 3, 2018.
- KUMAR, R.; OJHA, A. K.; MALMASI, S.; ZAMPIERI, M. Benchmarking aggression identification in social media. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. [S.l.: s.n.], 2018. p. 1–11.
- KUMAR, R.; OJHA, A. K.; ZAMPIERI, M.; MALMASI, S. Proceedings of the first workshop on trolling, aggression and cyberbullying (trac-2018). In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. [S.l.: s.n.], 2018.
- LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007.
- MLADENIC, D.; GROBELNIK, M. Feature selection for unbalanced class distribution and naïve bayes. In: *ICML*. [S.l.: s.n.], 1999. v. 99, p. 258–267.
- MONDAL, M.; SILVA, L. A.; CORREA, D.; BENEVENUTO, F. Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, Taylor & Francis, p. 1–21, 2018.
- NOCKLEBY, J. T. Hate speech. *Encyclopedia of the American constitution*, Detroit: Macmillan Reference USA, v. 3, p. 1277–79, 2000.
- PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. [S.l.], 2002. p. 79–86.
- PEARSON, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor & Francis, v. 50, n. 302, p. 157–175, 1900.
- PELLE, R. P. de; MOREIRA, V. P. Offensive comments in the brazilian web: a dataset and baseline results. In: *6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. [S.l.: s.n.], 2017. To appear.
- PLATT, J. Fast training of support vector machines using sequential minimal optimization. In: SCHOELKOPF, B.; BURGESS, C.; SMOLA, A. (Ed.). *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998. Disponível em: <<http://research.microsoft.com/~jplatt/smo.html>>.
- RANA, R.; SINGHAL, R. et al. *Chi-square test and its application in hypothesis testing - Scientific Figure on ResearchGate*. 2015. <[https://www.researchgate.net/figure/Excerpts-from-the-Chi-square-distribution-table\\_tb11\\_277935900](https://www.researchgate.net/figure/Excerpts-from-the-Chi-square-distribution-table_tb11_277935900)>. Accessed: 2019-07-2.
- REVISTABW. *Teorema de Bayes*. 2015. <[www.revistabw.com.br/revistabw/teorema-de-bayes/](http://www.revistabw.com.br/revistabw/teorema-de-bayes/)>. Accessed: 2018-11-25.
- RIBEIRO, M. H.; CALAIS, P. H.; SANTOS, Y. A.; ALMEIDA, V. A.; JR, W. M. Characterizing and detecting hateful users on twitter. *arXiv preprint arXiv:1803.08977*, 2018.



ROSARIO, M. do. Projeto de lei n.º 7.582. 2014.

SANTANA, R. *Café Com Código 09: Entendendo Métricas de Avaliação de Modelos*. 2017. <<http://minerandodados.com.br/index.php/2017/10/10/caf%C3%A9-com-c%C3%B3digo-09-m%C3%A9tricas-de-avaliacao-de-modelos/>>. Accessed: 2018-11-27.

SHARMA, A.; DEY, S. A comparative study of feature selection and machine learning techniques for sentiment analysis. In: ACM. *Proceedings of the 2012 ACM research in applied computation symposium*. [S.l.], 2012. p. 1–7.

SILGE, D. R. J. *Term Frequency and Inverse Document Frequency (tf-idf) Using Tidy Data Principles*. 2018. <[https://cran.r-project.org/web/packages/tidytext/vignettes/tf\\_idf.html](https://cran.r-project.org/web/packages/tidytext/vignettes/tf_idf.html)>. Accessed: 2018-11-27.

SILVA, G. A. *A liberdade de expressão e o discurso de ódio*. 2014. <<https://gus91sp.jusbrasil.com.br/artigos/152277318/a-liberdade-de-expressao-e-o-discurso-de-odio>>. Accessed: 2018-10-10.

SIMARD, P. Y.; STEINKRAUS, D.; PLATT, J. C. et al. Best practices for convolutional neural networks applied to visual document analysis. In: *Icdar*. [S.l.: s.n.], 2003. v. 3, n. 2003.

SOUZA, B. A.; ALMEIDA, T. G.; MENEZES, A. A.; NAKAMURA, F. G.; FIGUEIREDO, C.; NAKAMURA, E. F. For or against?: Polarity analysis in tweets about impeachment process of brazil president. In: ACM. *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*. [S.l.], 2016. p. 335–338.

STATISTICA. *Most famous social network sites worldwide as of July 2018, ranked by number of active users (in millions)*. 2018. Disponível em: <<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>>.

SUN, Y.; WONG, A. K.; KAMEL, M. S. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, v. 23, n. 04, p. 687–719, 2009.

TING, I.-H.; CHI, H.-M.; WU, J.-S.; WANG, S.-L. An approach for hate groups detection in facebook. In: SPRINGER. *The 3rd International Workshop on Intelligent Data Analysis and Management*. [S.l.], 2013. p. 101–106.

WOODS, K. S.; DOSS, C. C.; BOWYER, K. W.; SOLKA, J. L.; PRIEBE, C. E.; JR, W. P. K. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, v. 7, n. 06, p. 1417–1436, 1993.

YANG, Y.; LIU, X. et al. A re-examination of text categorization methods. In: *Sigir*. [S.l.: s.n.], 1999. v. 99, n. 8, p. 99.

ZHANG, H. The optimality of naive bayes. *AA*, v. 1, n. 2, p. 3, 2004.

# **Anexos**



## ANEXO A – Porcentagens da distribuição de chi-quadrado

<b>df</b>	<b>Probability level (alpha)</b>					
	<b>0.5</b>	<b>0.10</b>	<b>0.05</b>	<b>0.02</b>	<b>0.01</b>	<b>0.001</b>
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

Figura A.1 – Fonte: (RANA; SINGHAL et al., 2015)