UFRJ

INSTITUTO DE MATEMÁTICA
Universidade Federal do Rio de Janeiro

CLASSIFICAÇÃO DE POSTAGENS DO TWITTER EM PORTUGUÊS ASSOCIADAS A SENTIMENTOS NEGATIVOS

FELIPE ROCHA OSHIRO

SOLANGE JACOMO DA SILVA LOPES

Rio de Janeiro

2022

FELIPE ROCHA OSHIRO

SOLANGE JACOMO DA SILVA LOPES

CLASSIFICAÇÃO DE POSTAGENS DO TWITTER EM PORTUGUÊS ASSOCIADAS A SENTIMENTOS NEGATIVOS

Trabalho de Conclusão apresentado ao Curso de Especialização em Ciência de Dados da Universidade Federal do Rio de Janeiro.

Orientador: João Batista de Morais Pereira

Oshiro. Felipe Rocha

Lopes. Solange Jacomo da Silva

Classificação de postagens do Twitter em português associadas a sentimentos negativos / Felipe Rocha Oshiro e Solange Jacomo da Silva Lopes. – Rio de Janeiro, 2022.

53f.

Inclui referências.

Orientador: Prof. João Batista de Morais Pereira.

Monografia (Especialização em Ciência de Dados) – Instituto de Matemática da Universidade Federal do Rio de Janeiro.

FELIPE ROCHA OSHIRO

SOLANGE JACOMO DA SILVA LOPES

CLASSIFICAÇÃO DE POSTAGENS DO TWITTER EM PORTUGUÊS ASSOCIADAS A SENTIMENTOS NEGATIVOS

Trabalho de Conclusão apresentado ao Curso de Especialização em Ciência de Dados da Universidade Federal do Rio de Janeiro.

Aprovada em 24 de setembro de 2022.

BANCA EXAMINADORA

_____

Prof. João Batista de Morais Pereira

IM/UFRJ

_____

Prof. XXX

IM/UFRJ

RESUMO

A depressão é um transtorno mental que afeta a população em geral, interferindo nas atividades do cotidiano do indivíduo e nas suas relações interpessoais, podendo levar ao suicídio nos casos mais graves.

O diagnóstico da depressão é feito por um profissional da área da saúde através de análise do histórico do paciente. Uma das ferramentas que auxiliam os profissionais da área da saúde no diagnóstico dos pacientes é o *Patient Health Questionnaire* (PHQ-9), que é um questionário de rápida aplicação que avalia nove sintomas associados à depressão.

As redes sociais podem ser importante aliadas para identificar indivíduos que sofram de depressão, pois são um ambiente livre ao qual os usuários se sentem confortáveis para compartilhar seus pensamentos e sentimentos.

Este trabalho teve como objetivo analisar postagens de usuários na rede social Twitter, que continham em sua mensagem um ou mais termos associados a algum dos nove sintomas presentes na PHQ-9, somados a um décimo léxico sobre o uso de medicamentos associados a depressão, classificando se a mensagem estava associada a sentimentos positivos ou negativos.

Foram testados os modelos de classificação de Regressão Logística, *Support Vector Machine* (SVM) e *Random Forest*. O classificador SVM apresentou métricas de qualidade do ajuste mais estáveis e satisfatórias, alcançando 78% de acurácia.

Palavras-chave: Twitter, depressão, PHQ-9, aprendizado supervisionado, classificadores

ABSTRACT

Depression is a mental disorder that affects the general population, interfering with the individual's daily activities and interpersonal relationships, and may lead to suicide in the most severe cases.

The diagnosis of depression is made by a healthcare professional through analysis of the patient's history. One of the tools that help healthcare professionals diagnose patients is the Patient Health Questionnaire (PHQ-9), which is a quick-to-apply questionnaire that assesses nine symptoms associated with depression.

Social media can be important allies to identify individuals who suffer from depression, as they are a free environment in which users feel comfortable to share their thoughts and feelings.

This study had as goal analyze user posts on the social media Twitter, which contained in their message one or more terms associated with any of the nine symptoms present in PHQ-9, added to a tenth lexicon on the use of medications associated with depression, classifying whether the message was associated with positive or negative feelings.

The Logistic Regression, Support Vector Machine (SVM) and Random Forest classification models were tested. The SVM classifier presented a more stable and satisfactory quality of fit metrics, reaching 78% accuracy.

Keywords: Twitter, depression, PHQ-9, supervised learning, classifier model

SUMÁRIO

1 INTRODUÇÃO

De acordo com a Organização Mundial da Saúde (OMS), a depressão é um transtorno mental altamente prevalente na população em geral, podendo ser isolada ou associada a algum transtorno físico. A depressão situa-se em quarto lugar entre as principais causas de ônus, respondendo por 4,4% dos ônus acarretados por todas as doenças durante a vida.

O aparecimento de sintomas associados à depressão pode ocorrer em qualquer idade, sendo mais comum ao final da terceira década da vida, atingindo ambos os sexos, porém sendo mais frequente entre as mulheres.

A depressão interfere ativamente na rotina do indivíduo, afetando a sua performance em atividades comuns, como em sua capacidade de estudar, trabalhar, dormir ou se alimentar e em suas relações pessoais. Em casos graves, a depressão pode levar ao suicídio, sendo essa, de acordo com a Organização Mundial da Saúde (OMS), a quarta principal causa de morte entre jovens de 15 a 29 anos.

O Ministério da Saúde considera três principais causas para a depressão: genética, bioquímica cerebral e eventos vitais. O diagnóstico da depressão é clínico, feito por um profissional da saúde mental após coleta de histórico do paciente e realização de um exame do estado mental. Não existem exames laboratoriais específicos para realizar o diagnóstico da depressão.

No Brasil, segundo dados da Pesquisa Nacional de Saúde (PNS), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), em 2019, 10,2% (16,3 milhões) da

população adulta declarou ter recebido diagnóstico de depressão por profissional de saúde mental.

Um dos métodos que auxiliam os profissionais da área da saúde a diagnosticar pacientes que sofrem deste transtorno é o *Patient Health Questionnaire* (PHQ-9), que é uma ferramenta criada em 2001 pela farmacêutica Pfizer, composta por um questionário rápido que avalia nove léxicos sobre sintomas associados à depressão.

Além dos nove sintomas presentes na PHQ-9, MENDES, PASSADOR e CASELI (2021) sugerem o uso de um décimo léxico sobre o uso de medicamentos associados à depressão para auxiliar na identificação de pessoas que sofrem deste transtorno.

Os dez léxicos associados à depressão estão descritos na tabela 1.1.

Tabela 1.1 - Sintomas da PHQ-9 e décimo léxico

Léxico	Descrição
(1) Falta de interesse	Perda de interesse ou prazer
(2) Humor depressivo	Sentimentos de tristeza, inutilidade ou culpa
(3) Desordem do sono	Insônia ou hipersonia
(4) Falta de energia	Fadiga ou perda de energia
(5) Desordem alimentar	Diminuição ou aumento do peso ou apetite
(6) Baixa auto-estima	Falta de confiança em si mesmo
(7) Problemas de concentração	Concentração diminuída ou indecisão
(8) Hiperatividade ou baixa-atividade	Agitação ou retardo psicomotor
(9) Pensamentos de suicídio	Pensamentos recorrentes de morte
(10) Medicamentos	Medicamentos relativos à depressão

Identificar pessoas com perfis possivelmente depressivos possibilita a intervenção e o acompanhamento por um profissional da área da saúde. Neste sentido, as redes sociais

onlines podem ser importantes aliadas, uma vez que proporcionam aos usuários um ambiente aberto para que compartilhem proativamente seus sentimentos e pensamentos.

Segundo MENDES, PASSADOR e CASELI (2021), as redes sociais onlines são ambientes naturais para a identificação de pessoas com perfis possivelmente depressivos quando comparados com instrumentos comumente usados na área da saúde. MENDES, PASSADOR e CASELI (2021) investigaram a prevalência, ao longo do tempo, de sintomas associados à depressão nos Estados Unidos a partir de uma base de dados do Twitter.

O Twitter é uma rede social e um serviço de microblog, criado em 2006, que promove a conversa pública entre seus usuários através de postagens de até 280 caracteres, chamadas de *tweets*, em um ambiente livre e seguro. As atualizações são exibidas no perfil do usuário em tempo real e o acesso a rede social é gratuito, dependendo apenas da existência de conexão com a *internet*.

De acordo com a empresa Statista, em 2022, o Twitter informou possuir mais de 217 milhões de usuários ativos no mundo, sendo o Brasil o quarto país com maior penetração na rede social, com cerca de 19 milhões de usuários ativos.

Apresentado neste primeiro capítulo a importância que envolve a temática da depressão devido a alta prevalência do diagnóstico na população, juntamente com seu potencial de causar incapacidade aos indivíduos que sofrem deste transtorno, e também a importância das redes sociais onlines como aliadas na identificação de indivíduos com perfis depressivos, este trabalho irá realizar uma análise textual de postagens feitas por usuários do Twitter, com o objetivo de identificar as postagens que possam conter sentimentos negativos

associados à depressão, através de palavras associadas a algum dos sintomas da PHQ-9 ou ao décimo léxico sobre o uso de medicamentos.

A análise dos dados se dará através de modelos de classificação para dados binários, classificando as mensagens das postagens no Twitter em sentimentos positivos (não associados à depressão) ou sentimentos negativos (associados à depressão). Serão ajustados os modelos de classificação Regressão Logística, *Support Vector Machine* (SVM) e *Random Forest*, com o objetivo de analisar qual classificador se adequa melhor aos dados e qual o tamanho mínimo de amostra necessário para se obter resultados estimados robustos.

Estes três modelos foram selecionados a partir da revisão bibliográfica, onde foi conferido que possuem aplicabilidade ao tema, e por serem modelos ajustáveis para o aprendizado supervisionado, em que o treinamento do modelo é realizado a partir da base de dados com variável resposta já classificada. No capítulo 2, estão abordadas as técnicas já exploradas que estão na literatura para o tema de análise de textos com sentimentos associados à depressão através de dados do Twitter.

A base de dados a ser estudada foi obtida através da plataforma Twitter Developer e é composta por *tweets* em português postados na rede social Twitter por usuários que possuíam conta ativa e não privada no momento da extração dos dados. No capítulo 3, estão descritas as características gerais sobre a extração da base de dados para a análise, assim como as informações sobre a mineração e vetorização dos textos que compõem a base de dados a ser modelada.

A metodologia do trabalho está descrita no capítulo 4, composto pelo conteúdo teórico dos três modelos de classificação que tiveram seus resultados de ajustes comparados, assim como as métricas de qualidade dos ajustes considerados nessa comparação.

A aplicação dos métodos na base de dados estudada está presente no capítulo 5, assim como os resultados obtidos para posterior comparação e qualidade dos ajustes.

Por fim, no capítulo 6 está a conclusão do estudo, onde estão apresentados as considerações finais e possíveis temas para trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

DUQUE, RAYMUNDO e NETO (2018) analisaram sentimentos em publicações da rede social Twitter usando métodos de inteligência artificial, através do processamento de linguagem natural para identificar se um texto continha uma mensagem depressiva. A amostra era composta por 1.320 *tweets* em português, classificados manualmente pelos autores. Para mineração dos dados, foi realizada a remoção de *stopwords*, ou seja, palavras que não possuíam peso para a análise, e a extração dos radicais das palavras. O algoritmo utilizado para a análise foi o classificador de Naive Bayes, separando a base em 1.200 observações para a amostra de treinamento e 120 observações para a amostra de teste. Na apuração dos resultados da amostra de teste, alcançou-se o índice de 75% de classificações corretas.

ZANCHINI (2019) construiu um modelo de predição usando técnicas de processamento de linguagem natural em conjunto com o algoritmo de classificação supervisionada *Support Vector Machine* (SVM), aplicado a uma base de *tweets* rotulados como depressivos ou não-depressivos. Para obtenção da amostra, as mensagens que continham os termos "ansiedade", "depressão" ou "saúde mental" foram classificadas como depressivas e as mensagens atreladas a sentimentos positivos foram classificadas como não-depressivas. As mensagens neutras foram excluídas da análise. A amostra final estudada era composta por 7.148 *tweets* balanceados entre as duas classes. 70% da amostra foi utilizada para treinamento do modelo SVM e 30% para testes. O resultado do classificador SVM na amostra de teste foi de 97% de classificações corretas.

NETTO, GOMES e demais autores (2021) obtiveram uma base de 283 usuários do Twitter, com 46.600 *tweets* em português postados por esses usuários. Para a análise, a base de dados foi separada em três grupos: usuários que relataram em seus perfis terem sido diagnosticados com depressão; usuários que postaram *tweets* com pelo menos três palavras vinculadas à depressão; e usuários aos quais não foram encontrados nenhum *tweet* com palavras vinculadas à depressão. Os grupos foram nominados como depressivos, sintomáticos e assintomáticos, respectivamente. Para analisar os perfis dos usuários, foram testados os modelos *Random Forest*, K-vizinhos mais próximos, Naive Bayes, Regressão Logística e *Support Vector Machine* (SVM) com o objetivo de classificar em qual dos três grupos cada usuário estava presente. Entre os modelos analisados, o SVM apresentou a maior acurácia na classificação dos grupos de usuários, com 89% de classificações corretas. Os demais modelos tiveram acurácias próximas, entre 75% e 77% de classificações corretas.

Com a ajuda de um profissional de psicologia, YAZDAVARD, AL-OLIMAT e demais autores (2017) produziram, em inglês, uma lista de termos relacionados aos nove sintomas de depressão descritos na PHQ-9, além de termos relacionados ao uso de medicamentos associados à depressão, o décimo léxico. MENDES, PASSADOR e CASELI (2021) traduziram a lista produzida originalmente em inglês por YAZDAVARD, AL-OLIMAT e demais autores (2017) para português. A lista final em português é composta por 1.213 termos. A tabela 2.1 apresenta alguns exemplos de palavras e expressões associadas a cada um dos dez léxicos estudados.

Tabela 2.1 - Exemplos de termos associados aos dez léxicos

Léxico	Exemplos de termos
(1) Falta de interesse	Falta de ânimo, falta de prazer, entediado
(2) Humor depressivo	Tristonho, deprimente, abatido
(3) Desordem do sono	Insônia, sonolento, sem dormir
(4) Falta de energia	Preguiça, cansado, prostrado
(5) Desordem alimentar	Anorexia, bulimia, fazer dieta
(6) Baixa auto-estima	Não mereço, desprezível, fracasso
(7) Problemas de concentração	Disperso, distraído, desorientado
(8) Hiperatividade ou baixa-atividade	Ansiedade, agitado, inquieto
(9) Pensamentos de suicídio	Me matar, mereço morrer, melhor estar morto
(10) Medicamentos	Lítio, alprazolam, citalopram

3 MATERIAIS

A base de dados é composta por *tweets* publicados na rede social Twitter, entre os meses de junho e julho de 2022, por usuários que possuíam um perfil ativo e aberto na rede social, e cuja mensagem publicada, em português, possuía um ou mais termos que estivessem presentes na listagem de 1.213 termos associados a sintomas da PHQ-9 ou ao décimo léxico sobre medicamentos, construída em inglês por YAZDAVARD, AL-OLIMAT e demais autores (2017) e traduzida para o português por MENDES, PASSADOR e CASELI (2021).

Os *tweets* foram extraídos através do pacote *tweepy* disponível no *software* Python e a chave de acesso para utilização da API (*Application Programming Interface*) foi disponibilizada pelo Twitter através de requisição feita pelos autores na plataforma Twitter Developer.

Para a construção dos modelos estatísticos, selecionou-se uma amostra de 3.000 *tweets* e foi classificado pelos autores se a mensagem do texto estava associada a sentimentos positivos ou negativos. Ressalta-se que para reaplicação dos métodos usados neste trabalho, recomenda-se que um profissional da saúde mental auxilie na classificação dos sentimentos presentes na mensagem publicada pelos usuários no Twitter.

A tabela 3.1 contém alguns exemplos de *tweets* que compõem a base de dados.

Tabela 3.1 - Exemplos de *tweets* e suas classificações

Classificação de sentimentos	Exemplos de *tweets*
Positivo	É tão gostoso **ficar sozinha**, arrumei a casa ouvindo música.
	Ansiedade e medo à flor da pele para lançar a minha marca de jaquetas customizadas.
	A **cicatriz** da minha cesária é tão discreta que nem parece que fiz nada, a minha obstetra arrasou!
Negativo	Eu fico **triste** sem motivo nenhum.
	Sem ânimo para nada.
	Eu só **quero morrer**.

Observa-se na tabela 3.1 que apesar dos *tweets* conterem os termos "ficar sozinha", "ansiedade" e "cicatriz", estes estão associados a mensagens com sentimentos positivos, portanto foram classificados como positivos. Nos demais *tweets*, os termos "triste", "sem ânimo" e "quero morrer" estão associados a mensagens com sentimentos negativos e foram classificados como negativos.

Tratou-se do texto dos *tweets* para padronização das palavras, remoção de pontuações, caracteres especiais e palavras vazias ou *stopwords*, que são palavras que não possuem significado relevante para a análise. Foram extraídos os radicais das palavras, que são os elementos que contêm o significado básico da palavra, e foi realizada a vetorização das mensagens, que consiste na representação do texto em forma de vetor de palavras. Para a mineração dos textos foi utilizado o *software* R.

A preparação dos dados para a análise e a vetorização das mensagens foi realizada com o auxílio do pacote UDPipe no *software* R, que disponibiliza modelos pré-treinados construídos em *treebanks* de dependências universais em diversos idiomas, para esse estudo foi usado o português. O método UDPipe consiste em vetorizar o texto considerando a

tokenização e a marcação de partes, ou seja, a separação do texto em palavras e a posição em que a palavra se encontra dentro do texto. A tabela 3.2 apresenta exemplos da mineração dos textos dos *tweets*.

Tabela 3.2 - Exemplos do tratamento e da vetorização dos textos dos *tweets*

Texto original	**Padronização do texto e limpeza de caracteres especiais**	**Remoção de *stopwords***	**Extração dos radicais das palavras**	**Vetorização das palavras**
É tão gostoso ficar sozinha, arrumei a casa ouvindo música.	e tao gostoso ficar sozinha arrumei a casa ouvindo musica	tao gostoso ficar sozinha arrumei casa ouvindo musica	tao gostoso ficar sozinho arrumar casa ouvir musica	[tao, gostoso, ficar, sozinho, arrumar, casa, ouvir, musica]
Ansiedade e medo à flor da pele para lançar a minha marca de jaquetas customizadas.	ansiedade e medo a flor da pele para lancar a minha marca de jaquetas customizadas	ansiedade medo flor pele lancar marca jaquetas customizadas	ansiedade medo flor pele lancar marca jaqueta customizar	[ansiedade, medo, flor, pele, lancar, marca, jaqueta, customizar]
A cicatriz da minha cesária é tão discreta que nem parece que fiz nada, a minha obstetra arrasou!	a cicatriz da minha cesaria e tao discreta que nem parece que fiz nada a minha obstetra arrasou	cicatriz cesaria tao discreta parece fiz nada obstetra arrasou	cicatriz cesaria tao discreto parecer fazer nada obstetra arrasar	[cicatriz, cesaria, tao, discreto, parecer, fazer, nada, obstetra, arrasar]
Eu fico triste sem motivo nenhum.	eu fico triste sem motivo nenhum	fico triste motivo nenhum	ficar triste motivo nenhum	[ficar, triste, motivo, nenhum]
Sem ânimo para nada.	sem animo para nada	animo nada	animo nada	[animo, nada]
Eu só quero morrer.	eu so quero morrer	quero morrer	querer morrer	[querer, morrer]

Os modelos classificadores a serem ajustados neste trabalho (Regressão Logística, SVM e *Random Forest*) utilizam como *inputs* uma matriz de desenho com elementos numéricos. Sendo assim, as palavras finais do vetor de palavras são codificadas para um espaço numérico.

4 METODOLOGIA

4.1 REGRESSÃO LOGÍSTICA

O modelo de Regressão Logística é um modelo estatístico usado para a análise de variáveis dicotômicas. Os métodos descritos neste capítulo estão detalhados em DOBSON e BARNETT (2008).

Considere uma amostra com n variáveis aleatórias independentes, $Y_1, \ldots, Y_n$, em que Y_i segue uma distribuição de $\mathrm{Bernoulli}$ e assume os valores $Y_i = 1$, com probabilidade π_i, ou $Y_i = 0$, com probabilidade $(1 - \pi_i)$, onde $i = 1, \ldots, n$ e n é o tamanho da amostra.

A função de probabilidade marginal de Y_i é dada por:

$$\mathrm{P}(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad \text{com} \quad y_i = 0, 1,$$

e temos que:

$$\mathrm{E}(Y_i) = \pi_i \quad \text{e} \quad \mathrm{Var}(Y_i) = \pi_i(1-\pi_i)\,.$$

A função de probabilidade conjunta é dada por:

$$f(y\,;\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\,,$$

onde y é o vetor $y = (y_1, \ldots, y_n)$ e π é o vetor $\pi = (\pi_1, \ldots, \pi_n)$.

Se as probabilidades de sucesso de todas as variáveis aleatórias forem iguais, pode-se definir:

$$Z = \sum_{i=1}^{n} Y_i\,,$$

onde Z é o número de sucessos em n observações. Assim, $Z \sim \mathrm{Binomial}\,(n\,, \pi)$.

Considerando que a amostra pode ser repartida em m subgrupos com a mesma probabilidade de sucesso dentro de cada subgrupo. Então $Z_j \sim \text{Binomial}\,(n_j\,, \pi_j)$, para $j = 1,\ldots, m$, onde m é a quantidade de subgrupos e n_j é número de replicações de Bernoulli dentro do subgrupo j. A tabela 4.1.1 mostra o desenho de uma Regressão Logística considerando as probabilidades associadas ao sucesso e ao fracasso dos m subgrupos.

Tabela 4.1.1 - Desenho de uma Regressão Logística

Subgrupo	**1**	**2**	**...**	m
Sucessos	Z_1	Z_2	...	Z_m
Fracassos	$n_1 - Z_1$	$n_2 - Z_2$	...	$n_m - Z_m$
Total	n_1	n_2	**...**	n_m

O modelo de Regressão Logística tem como objetivo descrever a proporção de sucessos em cada subgrupo em termos das variáveis explicativas que caracterizam os subgrupos.

A proporção de sucessos em cada subgrupo é dada por:

$$p_j = \frac{Z_j}{n_j}\ ,$$

e temos que:

$$\text{E}(p_j) = \frac{1}{n_j} \quad \text{e} \quad \text{E}(Z_j) = \pi_j\,.$$

Então, as probabilidades π_j são modeladas como:

$$g(\pi_j) = x_j^T \beta\ ,$$

onde x_j é o vetor de variáveis explicativas do subgrupo *j*, tal que $x_j = (1, x_1, \ldots, x_p)^T$, β é o vetor de coeficientes do modelo, tal que $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, onde p é o número de variáveis explicativas do modelo e g é a função de ligação.

O modelo de Regressão Logística pode assumir como função de ligação a função Logit, que é dada por:

$$\operatorname{logit}(\pi_j) = \ln\left(\frac{\pi_j}{1-\pi_j}\right),$$

e a função que relaciona β e π_j é:

$$\ln\left(\frac{\pi_j}{1-\pi_j}\right) = x_j^T\beta\,.$$

O vetor β é o vetor de parâmetros desconhecidos. A partir da estimação do vetor β é possível estimar π_j, que é a probabilidade de sucesso para cada observação pertencente ao subgrupo *j*. Defina $\hat{\beta}$ como o estimador pontual de β, então o estimador de π_j é dado por:

$$\hat{\pi}_j = \frac{e^{x_j^T\hat{\beta}}}{1+e^{x_j^T\hat{\beta}}}\,.$$

O princípio de máxima verossimilhança é um procedimento utilizado na obtenção de estimadores. O método consiste em encontrar o valor do parâmetro ou do vetor paramétrico que torna a amostra observada mais verossímil, que é o valor que maximiza a função de verossimilhança. Neste caso, isto pode ser feito através da derivação da função de verossimilhança para obter o seu ponto de máximo, podendo se utilizar, de forma equivalente, a função de log-verossimilhança.

A função de probabilidade da variável $Z_j \sim \text{Binomial}\,(n_j\,,\pi_j)$ é dada por:

$$\mathrm{P}(Z_j = z_j) = \binom{n_j}{z_j}\pi_j^{z_j}(1-\pi_j)^{n_j - z_j}$$, com $z_j = 0, \ldots, n_j$,

então, a função de verossimilhança é dada por:

$$f(\pi\,; z) = \prod_{j=1}^{m}\binom{n_j}{z_j}\pi_j^{z_j}(1-\pi_j)^{n_j - z_j}\ ,$$

e a função de log-verossimilhança é:

$$l(\pi\,;\, z) = \sum_{j=1}^{m} z_j \ln\left(\frac{\pi_j}{1-\pi_j}\right) + n_j \ln\left(1-\pi_j\right) + \ln\binom{n_j}{z_j}\ ,$$

onde z é o vetor $z = (z_1,\ldots, z_m)$ e π é o vetor $\pi = (\pi_1,\ldots, \pi_m)$.

O estimador de máxima verossimilhança não possui uma forma fechada, sendo complexo obtê-lo, então se utilizam algoritmos iterativos para maximizar a função, como o algoritmo de Newton Raphson.

Os estimadores de máxima verossimilhança possuem as seguintes propriedades:

- São assintoticamente não viciados;
- São estimadores consistentes;
- Apresentam distribuição assintótica Normal;
- São eficientes, ou seja, apresentam variância mínima dentro da classe dos estimadores assintoticamente não viciados.

4.2 *SUPPORT VECTOR MACHINES*

O método SVM (*Support Vector Machines*, em português, Máquina de Vetores Suporte) é um algoritmo de aprendizado de máquina supervisionado usado para classificação de dados, que se baseia em uma generalização do Classificador de Margem Máxima.

A metodologia descrita neste capítulo está detalhada em JAMES, WITTEN e demais autores (2013).

O algoritmo SVM se aproveita da geometria, onde cada dado é representado em um espaço p-dimensional com os valores de suas coordenadas e o algoritmo tem como objetivo encontrar um hiperplano ideal que seja capaz de separar os dados de forma a melhor diferenciar as classes, maximizando a assertividade da classificação.

O hiperplano é um subespaço com $(p - 1)$ dimensões, onde p é a quantidade de variáveis explicativas disponíveis para a análise do modelo. Uma observação é estimada como pertencente a uma classe dependendo de qual lado do hiperplano se encontre.

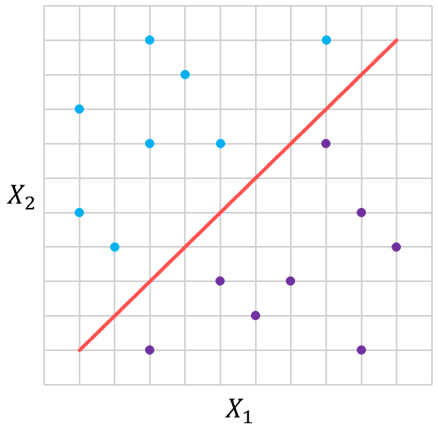
Este trabalho tem interesse em explorar a classificação de um problema binário, entretanto, este método pode ser aplicado em contextos de classificação com mais de duas categorias.

No problema de classificação binária, ou seja, quando $Y_i \in \{-1, 1\}$, seja a matrix X composta por n observações e p variáveis explicativas:

$$X = \begin{pmatrix} x_{11} & \dots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \dots & x_{np} \end{pmatrix}$$

e seja y o vetor de variáveis respostas, tal que $y_1,\ldots, y_n \in \{-1, 1\}$. Se tivermos $p = 2$, o modelo será em um espaço bidimensional.

Figura 4.2.1 - Representação geométrica bidimensional de um modelo SVM



No exemplo da figura 4.2.1, existe um hiperplano (uma reta) capaz de separar perfeitamente as observações, classificando-as corretamente.

Em um espaço é possível se obter infinitos hiperplanos. Para identificar o hiperplano que melhor diferencia as classes, é calculado a assertividade de cada hiperplano e também a margem de cada hiperplano, que é a distância perpendicular entre os pontos de dados mais próximos ao hiperplano. O melhor hiperplano será o que possuir maior assertividade e maior margem.

O Classificador de Margem Máxima assume o hiperplano de maior margem, ou seja, aquele que possui a maior distância perpendicular ao hiperplano, para que para novas observações exista uma margem maior no espaço que permita que a observação seja classificada corretamente.

A construção do hiperplano de margem máxima é a solução de um problema de otimização, dado por:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,M} \quad M$$

sujeito a:

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

e

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \quad \forall\, i = 1, \ldots, n \ ,$$

onde $\beta_0, \beta_1, \ldots, \beta_p$, são os coeficientes do hiperplano de margem máxima e M é a largura da margem.

Todas as observações estarão do lado certo do hiperplano, havendo uma distância perpendicular ao hiperplano $M > 0$. Se as classes não foram linearmente separáveis, o problema de otimização não terá solução com $M > 0$.

Para os casos em que não há solução capaz de separar linearmente as classes, é aceitável um erro de classificação. Esse classificador é chamado de Classificador de Margem Suave ou Classificador de Vetor Suporte, que é uma generalização do Classificador de Margem Máxima.

Neste caso, o problema de otimização é dado por:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M} \quad M$$

sujeito a:

$$\sum_{j=1}^{p} \beta_j^2 = 1 \ ,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \quad \forall \ i = 1, \ldots, n \ ,$$

$$\epsilon_i \geq 0 \ \text{ e } \ \sum_{i=1}^{n} \epsilon_i \leq C \, ,$$

onde ϵ_i é o erro de classificação associado a observação i e C é um parâmetro de ajuste do máximo de violações aceitável, sendo $C \geq 0$.

Seja Y_i a i-ésima observação. Se $\epsilon_i = 0$, a observação foi classificada corretamente no hiperplano. Se $\epsilon_i > 1$, a observação foi classificada erroneamente no hiperplano. E, se $0 < \epsilon_i < 1$, a observação foi classificada corretamente no hiperplano, porém do lado errado da margem.

O parâmetro de ajuste C é escolhido por meio de validação cruzada, controlando o equilíbrio entre viés e variância. Quando C é pequeno, as margens são estreitas e pouco violadas, isto equivale a um classificador altamente ajustado aos dados, que pode ter baixo viés, mas alta variância. Por outro lado, quando C é maior, a margem é maior, permitindo mais violações, isso retorna um classificador mais tendencioso, porém com menor variância.

O Classificador de Vetor Suporte é afetado apenas pelas observações que estão próximas da margem, que ao serem alteradas de posição são capazes de alterar a distância perpendicular ao hiperplano. Essas observações, próximas à distância perpendicular ao hiperplano, são chamadas de vetores suporte. Isso significa que o Classificador de Vetor Suporte é um classificador robusto para observações que estejam distantes do hiperplano.

Em vez de ajustar o classificador apenas com p variáveis, $X_1, X_2, \ldots, X_p$, pode acrescentar-se $2p$ variáveis, $X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2$, obtendo-se assim uma versão não linear do Classificados de Vetor Suporte.

Então, o problema de otimização é dado por:

$$\max \quad M$$

$$\beta_0, \beta_{11}, \beta_{12}, \ldots, \beta_{p1}, \beta_{p2}, \epsilon_1, \ldots, \epsilon_n, M$$

sujeito a:

$$\sum_{j=1}^{p} \sum_{k=1}^{2} \beta_{jk}^2 = 1 \ ,$$

$$y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} x_{ij}^2 \right) \geq M(1 - \epsilon_i) \quad \forall \ i = 1, \ldots, n \ ,$$

$$\epsilon_i \geq 0 \ \text{ e } \ \sum_{i=1}^{n} \epsilon_i \leq C \, ,$$

O modelo Máquina de Vetores Suporte (SVM) é uma extensão do Classificador de Vetor Suporte para classificações lineares ou não-lineares. Nestes casos, amplia-se o espaço de características usando funções polinomiais quadráticas, cúbicas ou de outras ordens.

Através da Máquina de Vetores Suporte, é possível ampliar o espaço das variáveis em um espaço de dimensão infinita com cálculos computacionais eficientes, sem aumentar a quantidade de parâmetros a serem estimados.

O produto interno de duas observações x_i e $x_{i'}$ é dada por:

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} \, .$$

A fronteira de decisão do classificador suporte é dada por:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \ \ ,$$

onde $x_1, \ldots, x_n$ são as observações de treinamento do modelo e existem n parâmetros α_i.

Para estimar $\beta_0, \alpha_1, \ldots, \alpha_n$ precisamos dos $\binom{n}{2}$ produtos internos $\langle x_i, x_{i'} \rangle$ entre todos os pares de observações.

Os únicos $\alpha_i \neq 0$ são os referentes aos vetores suporte. Então, seja *S* a coleção de índices dos vetores suporte, a equação $f(x)$ é reescrita como:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle \, .$$

Para calcular e estimar f apenas os produtos internos são necessários. O cálculo do produto interno pode ser substituído por uma generalização do produto interno, na forma:

$$K(x_i, x_{i'}) \, .$$

onde K é uma função *kernel* que quantifica a similaridade de duas observações.

Seja o *kernel* polinomial de grau d, com d sendo um polinômio positivo inteiro:

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j} \right)^d \, .$$

Outra possível escolha é o *kernel* radial, com γ constante positivo:

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right) \, .$$

A equação $f(x)$ pode ser reescrita como:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \, .$$

4.3 *RANDOM FOREST*

O classificador *Random Forest* é um conjunto de classificadores *Decision Tree*, ao qual um grande número de árvores de classificação individuais são modeladas de forma descorrelacionada, e cada árvore irá retornar uma estimativa de classificação para cada observação na amostra. O *Random Forest* consiste em combinar a estimativa de todas as árvores, e, ao final, produzir uma única estimativa, em consenso, para cada observação da amostra.

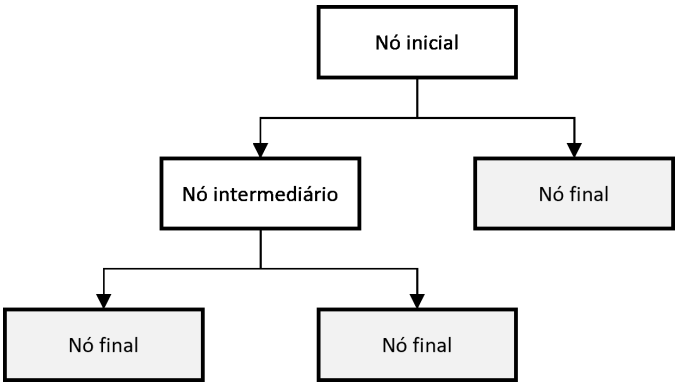
Os métodos descritos neste capítulo estão detalhados em TACONELI (2008), DIAS, LAGE e demais autores (2008) e JAMES, WITTEN e demais autores (2013).

O algoritmo *Decision Tree* cria uma árvore de classificação a partir de partições binárias recursivas dos dados, que transformam o problema inicial em diversos problemas menores e menos complexos. O modelo é representado por uma estrutura de segmentação hierárquica, como em uma árvore invertida. As divisões binárias dos dados em grupos são feitas de forma que cada vez os grupos sejam mais heterogêneos entre si, e que cada grupo seja homogêneo em relação a variável resposta.

Cada divisão é chamada de nó, que podem ser representados por retângulos. O primeiro nó é chamado de nó inicial ou nó raiz e é formado por toda a base de dados. A partir dele são criadas as regras de divisão. Os últimos nós são os nós finais e todos os nós do meio são nós intermediários, sendo cada nó formado por um subconjunto dos dados da amostra. O conjunto de nós da mesma linha é chamado de geração.

A representação de uma Árvore de Classificação está na figura 4.3.1.

Figura 4.3.1 - Representação de uma Árvore de Classificação



Para encontrar a melhor partição do nó, procura-se minimizar a impureza dos nós resultantes, que pode ser avaliada através do índice de impureza de Gini. A impureza de um nó é o grau de heterogeneidade dos nós resultantes da divisão.

O processo de divisão considera todas as combinações de variáveis explicativas e seleciona aquela que possuir menor grau de impureza. É utilizado um algoritmo para verificar as divisões possíveis e encontrar qual maximiza a diferença entre as proporções de sucesso em cada grupo e reduz a heterogeneidade dentro de cada grupo.

Para variáveis explicativas quantitativas, a divisão do grupo é feita no ponto de corte que maximiza a diferença entre os grupos resultantes. Quando a variável explicativa possui k categorias, existem $(2^{k-1} - 1)$ possíveis divisões. Uma mesma variável pode ser utilizada diversas vezes ao longo do processo de partição dos dados.

O índice de impureza de Gini define a impureza de um nó que possui variável resposta binária como:

$$\mathrm{I} = 1 - p_1^2 - p_0^2 \,,$$

onde p_1 é a probabilidade de sucesso da variável no nó e p_0 é a probabilidade de fracasso da variável no nó, sendo $p_1 = 1 - p_0$.

Se dois registros são selecionados aleatoriamente com reposição em um nó, a probabilidade de que ambos sejam sucesso é p_1^2 e a probabilidade de que ambos sejam fracasso é p_0^2. Então, $\mathrm{I} = 1 - p_1^2 - p_0^2$ pode ser interpretado com a probabilidade de que dois registros selecionados aleatoriamente com reposição sejam diferentes. Deste modo, o índice contabiliza a proporção de observações em cada classe da variável resposta no nó raiz.

O valor máximo assumido pelo índice de Gini é de 0,5, que ocorre quando as duas classes estão igualmente representadas. Um nó puro tem índice de Gini igual a 0, que ocorre quando o nó possui observações pertencentes a uma única classe.

Na construção da árvore alguns critérios são utilizados para determinar quando parar a divisão dos grupos. Um grupo não será dividido se:

- Todas as unidades amostrais dentro do grupo possuírem o mesmo valor para todas as variáveis preditoras;
- O grupo se tornar puro, ou seja, todas as unidades amostrais apresentam a mesma resposta para a variável resposta;
- O número de unidades na divisão seguinte for menor que um número mínimo pré definido. É importante definir um número mínimo para o tamanho de um grupo, pois grupos com poucas observações podem causar instabilidade e prejudicar a capacidade de predição do modelo;
- A árvore alcançar seu tamanho pré-definido.

Para a construção do classificador *Random Forest,* em cada árvore, é selecionada uma amostra aleatória com m variáveis preditoras candidatas a divisão dos nós. Essa seleção acontece para que as árvores não sejam semelhantes entre si, sendo assim descorrelacionadas.

A probabilidade de sucesso é estimada individualmente em todas as árvores de classificação e é agregada para se obter apenas uma estimativa, sendo a estimativa final da probabilidade de sucesso, a média da probabilidade estimada por cada árvore.

4.4 QUALIDADE DO AJUSTE

A sensibilidade e a especificidade são métricas de avaliação do desempenho de modelos de classificação, calculadas a partir da matriz de confusão. A sensibilidade do modelo é a probabilidade do teste de diagnóstico produzir um resultado positivo, quando de fato o resultado é positivo, e a especificidade do modelo é a probabilidade do teste de diagnóstico apresentar um resultado negativo, quando de fato o resultado é negativo.

Seja $\hat{\pi}_i$ a probabilidade estimada de sucesso para cada observação i. Um valor p é definido, tal que, se $\hat{\pi}_i \geq p$, a observação é classificada como sucesso, e se $\hat{\pi}_i < p$, a observação é classificada como fracasso.

Tabela 4.4.1 - Representação de um teste de diagnóstico

Resultado do modelo	**Padrão real**	
	Positivos	Negativos
Positivos	Verdadeiros positivos (VP)	Falsos positivos (FP)
Negativos	Falsos negativos (FN)	Verdadeiros negativos (VN)
Desempenho	$\text{Sensibilidade} = \text{S}_\text{E} = \frac{\text{VP}}{\text{VP} + \text{FN}}$	$\text{Especificidade} = \text{E}_\text{S} = \frac{\text{VN}}{\text{FP} + \text{VN}}$

A tabela 4.4.1 mostra a representação de um teste de diagnóstico, também chamado de matriz de confusão, onde os verdadeiros positivos e verdadeiros negativos são os acertos do modelo, e os falsos positivos e falsos negativos são os erros do modelo. A partir dessa representação pode-se estimar a sensibilidade e a especificidade do modelo. O resultado ideal do teste de diagnóstico é que tanto a sensibilidade quanto a especificidade sejam iguais a 1, ou seja, que a quantidade de falsos negativos e falsos positivos sejam iguais a 0.

A partir das medidas do teste de diagnóstico, pode-se calcular a acurácia, que é a média global de classificações corretas do modelo:

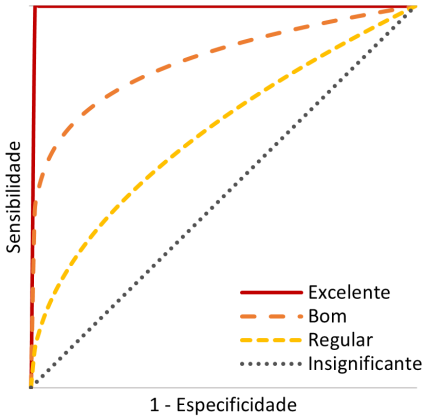
$$\mathrm{A_C} = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{FP} + \mathrm{VN} + \mathrm{FN}} .$$

Um outro método de avaliação do desempenho de um modelo de classificação que possua variável resposta binária é a curva ROC (*Receiver Operating Characteristic*), que a partir do teste de diagnóstico, verifica a probabilidade de acerto do modelo através das medidas de sensibilidade e especificidade do modelo.

A curva ROC é uma função contínua de $\mathrm{S_E}$ *versus* $1 - \mathrm{E_S}$ construída a partir do resultado de testes para diversos valores de p. A curva que representa um teste de diagnóstico com perfeita discriminação está localizada da origem até o canto superior esquerdo. Já um teste incapaz de discriminar apresenta uma curva linear entre a origem e o canto superior direito.

Para analisar a distância entre a curva empírica e a curva teórica sem poder de discrimização, observa-se a área sob a curva ROC. Quanto melhor a capacidade de discriminação de um modelo, a curva se aproxima da curva ideal e a área sob a curva se aproxima de 1. Em geral, modelos que apresentem área sob a curva maior ou igual a 0,5 são considerados adequados. A figura 4.4.1 apresenta exemplos de curvas e suas respectivas qualidades para o ajuste.

Figura 4.4.1 - Exemplos de Curvas ROC e qualidade do ajuste



Para melhor entendimento sugere-se a leitura de MARTINEZ, NETO e PEREIRA (2003).

5 ANÁLISES E RESULTADOS

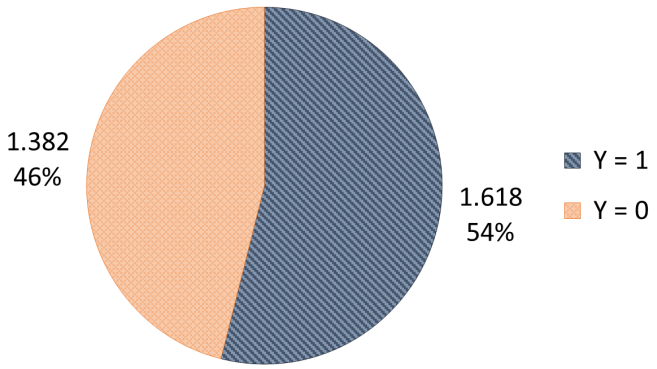
A amostra dos dados analisada é composta por 3.000 *tweets*, dos quais 1.618 foram classificados como mensagens associadas a sentimentos negativos, sendo 54% da amostra; e 1.382 foram classificados como mensagens associadas a sentimentos positivos, sendo 46% da amostra.

A variável resposta Y foi definida como:

$$Y = \begin{cases} 0 & \text{, se a mensagem está associada a sentimentos positivos} \\ 1 & \text{, se a mensagem está associada a sentimentos negativos} \end{cases}$$

A distribuição da variável resposta Y está apresentada na figura 5.1.

Figura 5.1 - Distribuição da variável resposta *Y*

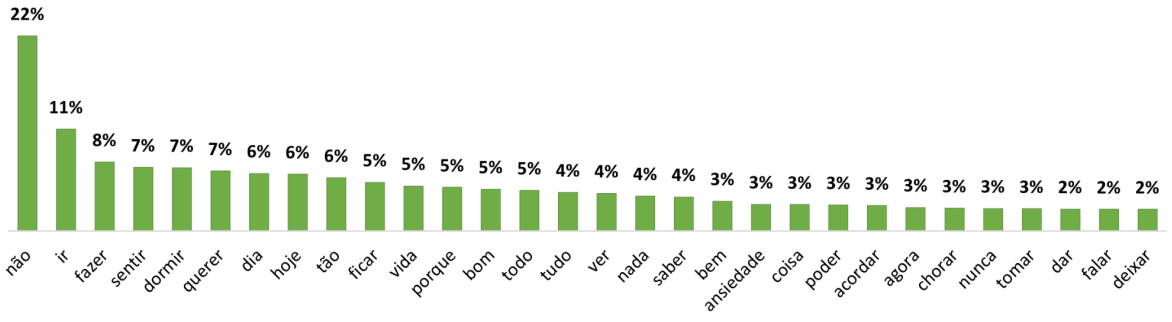


A classe $Y = 1$ é a mais presente na amostra de dados, porém não compromete o balanceamento da amostra.

Após a mineração do texto para limpeza de caracteres especiais, extração dos radicais das palavras e remoção de *stopwords*, ou seja, palavras que sozinhas não possuem peso para a análise, a frequência de palavras é apresentada nas figuras 5.2 e 5.3.

Figura 5.2 - Nuvem das 100 palavras mais frequentes

Figura 5.3 - Distribuição das 30 palavras mais frequentes



A frequência de palavras para $Y = 0$ e $Y = 1$ é apresentada nas figuras 5.4 e 5.5, respectivamente:

Figura 5.4 - Distribuição das 30 palavras mais frequentes para *Y* = 0

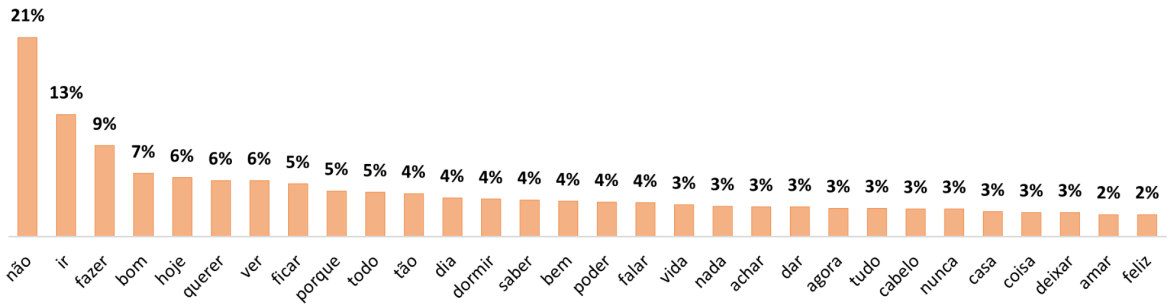
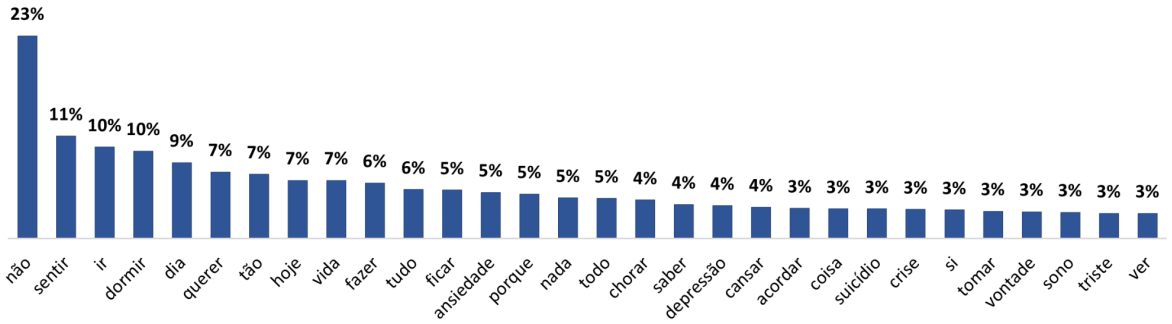


Figura 5.5 - Distribuição das 30 palavras mais frequentes para *Y* = 1



Observa-se nas figuras 5.4 e 5.5 que a palavra "não" é a mais presente em ambas as classes de *tweets*, estando presente em 21% dos *tweets* classificados como $Y = 1$ e em 23% dos *tweets* classificados como $Y = 0$. A palavra "não" é capaz de inverter o sentido de uma sentença, sendo assim sua presença bastante relevante na análise.

Para $Y = 1$, na figura 5.4, é possível ver a presença das palavras "bom", "bem", "amar" e "feliz" entre as mais frequentes nos *tweets*. Já para $Y = 0$, na figura 5.5, é possível ver a presença das palavras "ansiedade", "chorar", "depressão", "suicídio" e "triste" entre as mais frequentes.

Para a modelagem dos dados, separou-se a base de dados composta por 3.000 *tweets* em dez amostras de tamanhos diferentes, com o objetivo de investigar a performance do modelo para diferentes tamanhos de amostra. Cada uma das dez amostras foi separada em duas amostras, sendo 70% da amostra usada para o treinamento dos modelos e 30% da amostra usada para testes dos resultados, conforme apresentado na tabela 5.1.

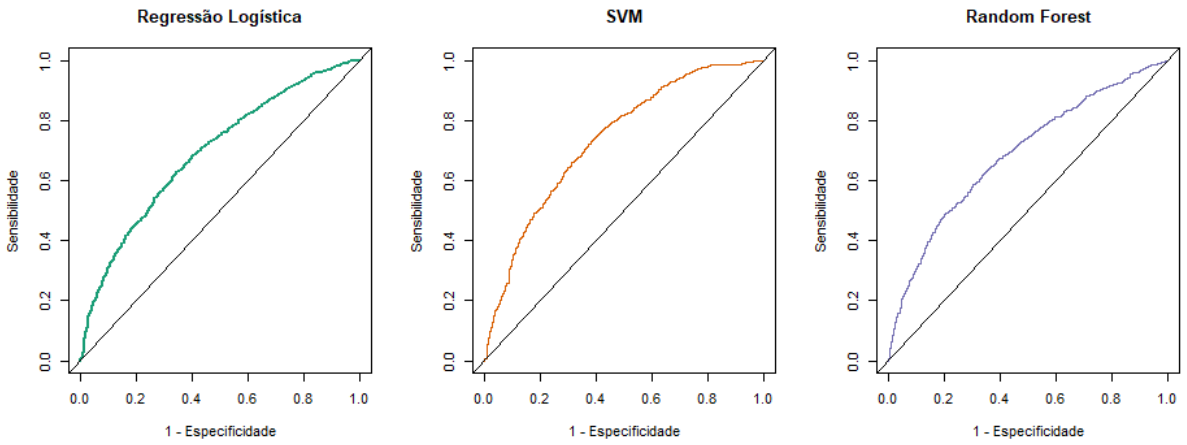
Tabela 5.1 - Tamanhos das amostras

Amostra	*Share* amostra total (*r*)	Tamanho amostra	Tamanho amostra de treino (70%)	Tamanho amostra de teste (30%)
1	10%	300	210	90
2	20%	600	420	180
3	30%	900	630	270
4	40%	1.200	840	360
5	50%	1.500	1.050	450
6	60%	1.800	1.260	540
7	70%	2.100	1.470	630
8	80%	2.400	1.680	720
9	90%	2.700	1.890	810
10	100%	3.000	2.100	900

Para cada uma das dez amostras foram treinados os modelos de Regressão Logística, *Support Vector Machine* e *Random Forest*, utilizando o *software* R e os pacotes stats, e1071 e randomForest, respectivamente.

Utilizou-se a curva ROC para avaliar o desempenho dos modelos de classificação. A figura 5.6 apresenta as curvas ROC para os modelos de Regressão Logística, SVM e *Random Forest* com amostragem de 100% da base total.

Figura 5.6 - Curvas ROC para os modelos de Regressão Logística, SVM e *Random Forest* com amostra de 100% da base total



As tabelas 5.2, 5.3 e 5.4 apresentam as áreas sob a curva ROC de cada modelo treinado, para as dez amostras e para os modelos de Regressão Logística, SVM e *Random Forest*, respectivamente, onde r é o percentual da amostra utilizado para a modelagem.

Tabela 5.2 - Área sob a curva ROC para os modelos de Regressão Logística

r = 10%	*r* = 20%	*r* = 30%	*r* = 40%	*r* = 50%	*r* = 60%	*r* = 70%	*r* = 80%	*r* = 90%	*r* = 100%
0,74	0,76	0,74	0,72	0,72	0,71	0,71	0,70	0,70	0,69

Tabela 5.3 - Área sob a curva ROC para os modelos SVM

r = 10%	*r* = 20%	*r* = 30%	*r* = 40%	*r* = 50%	*r* = 60%	*r* = 70%	*r* = 80%	*r* = 90%	*r* = 100%
0,83	0,83	0,81	0,78	0,77	0,76	0,75	0,75	0,74	0,74

Tabela 5.4 - Área sob a curva ROC para os modelos *Random Forest*

r = 10%	*r* = 20%	*r* = 30%	*r* = 40%	*r* = 50%	*r* = 60%	*r* = 70%	*r* = 80%	*r* = 90%	*r* = 100%
0,67	0,72	0,73	0,70	0,69	0,69	0,70	0,71	0,69	0,69

O modelo de Regressão Logística teve a maior área sob a curva para amostragem de 20% da base total, com área de 0,76. As áreas possuem variabilidade entre 0,69 e 0,76 entre os diferentes percentuais de amostragem.

O modelo SVM teve a maior área sob a curva ROC para os modelos que utilizam a amostragem de 10% ou 20% da base de dados, apresentando áreas de 0,83. As áreas para os diferentes percentuais de amostragem variam entre 0,74 e 0,83, sendo a performance do modelo SVM, para esse indicador, superior ao modelo de Regressão Logística.

O modelo *Random Forest* teve a maior área abaixo da curva ROC para a amostra que utiliza 30% da base de dados, com área de 0,73. A variabilidade das áreas para este modelo está no intervalo entre 0,67 e 0,73, ficando abaixo da performance dos modelos Regressão Logística e SVM, para este indicador.

Para calcular a acurácia dos modelos, selecionou-se diversos pontos de p, variando no intervalo entre 5% e 95%, tal que, se a probabilidade de sucesso estimada para a observação for maior ou igual a p, a observação é classificada como sucesso ($Y = 1$), e se a probabilidade de sucesso estimada para a observação for menor do que p, a observação é classificada como fracasso ($Y = 0$). O objetivo deste cálculo é entender o valor ideal de p para realizar as classificações das observações de forma a maximizar a acurácia dos modelos.

Tabela 5.5 - Acurácia dos modelos de Regressão Logística para as amostras de teste

	***r* = 10%**	***r* = 20%**	***r* = 30%**	***r* = 40%**	***r* = 50%**	***r* = 60%**	***r* = 70%**	***r* = 80%**	***r* = 90%**	***r*=100%**
***p* = 5%**	76%	64%	64%	59%	64%	64%	62%	60%	65%	66%
***p* = 10%**	76%	65%	65%	59%	64%	64%	61%	61%	65%	68%
***p* = 15%**	76%	64%	65%	59%	64%	64%	62%	61%	65%	68%
***p* = 20%**	79%	64%	66%	59%	64%	64%	62%	62%	66%	68%
***p* = 25%**	78%	66%	65%	60%	63%	64%	62%	62%	66%	68%
***p* = 30%**	80%	67%	65%	61%	62%	63%	61%	61%	64%	67%
***p* = 35%**	77%	66%	65%	61%	63%	63%	60%	60%	63%	66%
***p* = 40%**	78%	67%	64%	62%	62%	63%	59%	59%	62%	66%
***p* = 45%**	76%	67%	63%	62%	62%	61%	59%	58%	61%	64%
***p* = 50%**	76%	68%	61%	63%	62%	59%	57%	57%	60%	63%
***p* = 55%**	74%	68%	60%	63%	61%	59%	57%	58%	60%	62%
***p* = 60%**	73%	69%	60%	62%	62%	59%	57%	57%	59%	62%
***p* = 65%**	74%	68%	57%	61%	62%	59%	56%	56%	59%	62%
***p* = 70%**	72%	68%	58%	61%	61%	59%	58%	57%	60%	61%
***p* = 75%**	71%	68%	58%	61%	60%	57%	57%	58%	60%	61%
***p* = 80%**	69%	67%	57%	61%	59%	56%	57%	58%	60%	59%
***p* = 85%**	68%	67%	57%	61%	59%	54%	56%	57%	58%	57%
***p* = 90%**	66%	64%	55%	58%	58%	53%	54%	56%	57%	56%
***p* = 95%**	62%	62%	53%	56%	57%	52%	54%	56%	57%	54%

A tabela 5.5 apresenta as acurácias dos modelos de Regressão Logística para as dez amostras e para os cortes de classificação das observações comparando as probabilidades estimadas com os valores de p. A maior acurácia obtida para esse modelo é de 80%, quando foi utilizada 10% da base de dados para o treinamento do modelo e com o ponto de corte de probabilidade fixado em 30%, ou seja, se a probabilidade estimada pelo modelo para uma observação for maior ou igual a 30%, a observação é classificada como sucesso, caso

contrário, a observação é classificada como fracasso. Para este modelo, a pior acurácia obtida foi de 52% quando usado 60% da base total como amostra e com o ponto de corte de probabilidade fixado em 95%.

Tabela 5.6 - Acurácia dos modelos SVM para as amostras de teste

	***r* = 10%**	***r* = 20%**	***r* = 30%**	***r* = 40%**	***r* = 50%**	***r* = 60%**	***r* = 70%**	***r* = 80%**	***r* = 90%**	***r*=100%**
***p* = 5%**	59%	54%	57%	58%	57%	59%	57%	54%	56%	57%
***p* = 10%**	60%	56%	57%	58%	58%	59%	59%	55%	57%	58%
***p* = 15%**	62%	57%	57%	57%	58%	59%	60%	56%	57%	58%
***p* = 20%**	64%	60%	57%	57%	60%	61%	60%	57%	58%	59%
***p* = 25%**	66%	61%	57%	57%	61%	62%	61%	57%	58%	61%
***p* = 30%**	70%	62%	60%	56%	60%	61%	61%	58%	60%	61%
***p* = 35%**	74%	63%	60%	58%	61%	61%	61%	58%	62%	63%
***p* = 40%**	74%	64%	60%	58%	60%	61%	60%	60%	64%	64%
***p* = 45%**	76%	67%	59%	58%	58%	62%	61%	60%	64%	67%
***p* = 50%**	76%	64%	59%	59%	59%	62%	61%	62%	65%	67%
***p* = 55%**	78%	65%	60%	60%	61%	61%	61%	62%	66%	67%
***p* = 60%**	77%	68%	59%	61%	62%	63%	61%	61%	66%	66%
***p* = 65%**	74%	68%	59%	62%	63%	61%	59%	61%	65%	67%
***p* = 70%**	73%	65%	57%	62%	62%	60%	58%	60%	64%	66%
***p* = 75%**	72%	64%	60%	63%	61%	60%	58%	59%	64%	65%
***p* = 80%**	68%	62%	57%	61%	60%	58%	59%	58%	62%	63%
***p* = 85%**	62%	62%	57%	61%	58%	57%	58%	56%	61%	62%
***p* = 90%**	57%	60%	56%	58%	57%	55%	56%	57%	60%	61%
***p* = 95%**	60%	58%	51%	54%	56%	52%	54%	57%	59%	58%

A tabela 5.6 apresenta os resultados para os modelos SVM. Neste modelo, a maior acurácia obtida é de 78%, também quando utilizado 10% da base de dados para o treinamento do modelo, e com o ponto de corte de probabilidade fixado em 55%. Em

relação à pior acurácia, para o modelo SVM é de 51%, quando usado 30% da base total como amostra de treinamento e com o ponto de corte de probabilidade fixado também em 95%.

Tabela 5.7 - Acurácia dos modelos *Random Forest* para as amostras de teste

	***r* = 10%**	***r* = 20%**	***r* = 30%**	***r* = 40%**	***r* = 50%**	***r* = 60%**	***r* = 70%**	***r* = 80%**	***r* = 90%**	***r*=100%**
***p* = 5%**	53%	49%	58%	56%	55%	56%	53%	51%	53%	54%
***p* = 10%**	53%	49%	58%	56%	56%	56%	53%	51%	53%	54%
***p* = 15%**	56%	53%	59%	56%	56%	56%	53%	51%	53%	54%
***p* = 20%**	57%	54%	58%	56%	56%	57%	54%	52%	54%	55%
***p* = 25%**	60%	57%	58%	56%	57%	58%	56%	54%	56%	58%
***p* = 30%**	63%	58%	59%	56%	58%	60%	57%	56%	58%	60%
***p* = 35%**	63%	60%	61%	58%	61%	62%	59%	58%	60%	62%
***p* = 40%**	63%	62%	63%	58%	64%	66%	60%	59%	62%	64%
***p* = 45%**	66%	64%	60%	61%	66%	69%	62%	59%	64%	65%
***p* = 50%**	71%	65%	60%	61%	65%	66%	61%	62%	65%	65%
***p* = 55%**	73%	63%	59%	63%	65%	62%	59%	60%	64%	65%
***p* = 60%**	72%	63%	60%	63%	63%	58%	58%	59%	63%	64%
***p* = 65%**	69%	64%	55%	64%	61%	56%	58%	59%	61%	62%
***p* = 70%**	64%	62%	55%	61%	58%	54%	56%	57%	60%	60%
***p* = 75%**	61%	62%	54%	56%	57%	54%	55%	57%	58%	57%
***p* = 80%**	59%	60%	50%	52%	56%	52%	54%	56%	56%	55%
***p* = 85%**	54%	55%	47%	51%	52%	51%	52%	53%	52%	51%
***p* = 90%**	52%	53%	44%	48%	49%	47%	49%	51%	50%	49%
***p* = 95%**	49%	51%	42%	44%	45%	44%	47%	49%	48%	47%

A tabela 5.7 apresenta as acurácias dos modelos *Random Forest* para as dez amostras e para os cortes de classificação das observações comparando as probabilidades estimadas com os valores de p. A maior acurácia obtida para esse modelo é de 73%, quando utilizado 10% da base de dados para a modelagem dos dados, e com o ponto de corte de

probabilidade fixado em 55%. Para este modelo, a pior acurácia obtida foi de 42%, quando usado 30% da base total como amostra e com o ponto de corte de probabilidade fixado em 95%.

A tabela 5.8 apresenta os melhores modelos para cada classificador de acordo com a maior acurácia.

Tabela 5.8 - Melhores modelos para cada classificador

	Regressão Logística	**SVM**	***Random Forest***
r	10%	10%	10%
p	30%	55%	55%
Sensibilidade	85%	77%	71%
Especificidade	74%	79%	76%
Acurácia	80%	78%	73%
Curva ROC	0,74	0,83	0,67

Para os três classificadores a amostragem de 10% da base total apresenta os melhores resultados, sendo para a Regressão Logística e corte de probabilidade fixada em 30% e para os modelos SVM e *Random Forest* em 55%.

A maior acurácia é de 80% para o modelo de Regressão Logística, porém há um desbalanceamento entre as métricas de sensibilidade e especificidade, que são de 85% e 74% respectivamente, isso significa que o modelo é mais assertivo para a classe $Y = 1$ do que para a classe $Y = 0$.

O classificador SVM apresenta a segunda maior acurácia entre os três classificadores, em 78%, e sensibilidade e especificidade balanceadas em 77% e 79% respectivamente. Também possui a maior área abaixo da curva ROC de 0,83, confirmando o balanceamento

entre as métricas de sensibilidade e especificidade. Isso significa que o modelo acerta quase igualmente as classificações para $Y = 0$ ou $Y = 1$.

O classificador *Random Forest* apresentou os piores resultados entre os três classificadores estudados, com 73% de acurácia e área abaixo da curva ROC em 0,67. A sensibilidade do modelo é de 71%, sendo a menor entre os três classificadores, e a especificidade do modelo é de 76%, ficando acima do modelo de Regressão Logística porém abaixo do modelo SVM.

Tabela 5.9 - Exemplos de classificações para os melhores modelos de cada classificador

Classificação de sentimentos	**Exemplos de *tweets***	**Regressão Logística**	**SVM**	***Random Forest***
Positivo (*Y* = 0)	É tão gostoso **ficar sozinha**, arrumei a casa ouvindo música.	0	0	0
	Ansiedade e medo à flor da pele para lançar a minha marca de jaquetas customizadas.	0	0	0
	A **cicatriz** da minha cesária é tão discreta que nem parece que fiz nada, a minha obstetra arrasou!	0	0	1
Negativo (*Y* = 1)	Eu fico **triste** sem motivo nenhum.	1	1	1
	Sem ânimo para nada.	1	1	1
	Eu só **quero morrer**.	1	1	1

A tabela 5.9 apresenta as estimativas para Y dos três melhores classificadores apresentados na tabela 5.8, para os exemplos de *tweets* da tabela 3.1, no capítulo 3. Os classificadores de Regressão Logística e SVM classificaram corretamente todos os seis exemplos de *tweets* apresentados, já o classificador *Random Forest* classificou erroneamente um dos seis *tweets* exemplos.

6 CONCLUSÕES

Os resultados obtidos pelos três classificadores estudados neste trabalho, Regressão Logística, *Support Vector Machine* e *Random Forest*, foram satisfatórios, sendo possível a identificação de sentimentos negativos associados à depressão através de postagens no Twitter.

O classificador *Support Vector Machine* (SVM) teve métricas de qualidade mais estáveis do que os demais classificadores e com alta acurácia, de 78%. Também teve a maior especificidade, de 77%, e maior área abaixo da curva ROC, de 0,83, entre os três classificadores estudados.

O classificador de Regressão Logística teve uma alta acurácia, de 80%, porém teve um desbalanceamento entre as métricas de sensibilidade e especificidade, e apresentou a pior performance para a especificidade entre os classificadores estudados, de 74%, fazendo com que sua performance ficasse atrás da performance do classificador SVM, mesmo com maior valor de acurácia.

O classificador *Random Forest* apresenta resultados satisfatórios, porém com métricas de qualidade piores do que dos demais classificados, tendo apresentado os piores valores para a maioria das métricas de qualidade do ajuste do modelo.

Para os testes de tamanho da amostra, os três modelos apresentaram resultados satisfatórios para amostragem de 10% da base total, sendo utilizadas 210 observações para treinamento dos modelos e 90 observações para validação dos resultados. Este cenário ocorre pois conforme aumenta-se o tamanho da amostra de teste, aumenta-se também a

quantidade de palavras distintas presentes nas variáveis explicativas, incluindo palavras que aparecem com baixa frequência, dificultando assim a capacidade preditiva do modelo. Isto significa que, apesar de apresentar resultados satisfatórios, a amostragem de apenas 10% da base total não é necessariamente capaz de representar o total populacional, pois perde informações relevantes no momento do teste. No presente trabalho foi utilizado 70% da amostra para treinamento dos modelos e 30% da amostra para validação dos resultados. Para trabalhos futuros, recomenda-se uma investigação para diferentes porcentagens de tamanhos da amostra para teste e validação.

Neste cenário, recomenda-se o uso do classificador *Support Vector Machine* (SVM) para este problema e ponto de corte de probabilidade fixado em 55%, ou seja, se a probabilidade estimada pelo modelo for maior ou igual a 55%, o *tweet* deverá ser classificado como mensagem associada a sentimentos negativos, caso contrário, o *tweet* deverá ser classificado como mensagem associada a sentimentos positivos. O tamanho da amostra a ser utilizado para teste e validação dos modelos, sugere-se um valor baixo, pois o modelo é capaz de aprender e gerar resultados satisfatórios para amostras pequenas.

Estudado neste trabalho a vetorização de mensagens em palavras, como por exemplo na frase "eu fico triste sem motivo algum" (tabela 3.2, no capítulo 3), onde após os tratamentos, o vetor final de palavras a ser *input* dos modelos é [ficar, triste, motivo, nenhum], para trabalhos futuros, sugere-se re-treinar os modelos usando como variáveis explicativas as expressões, que podem ser compostas por uma ou mais palavras, ao invés de apenas as palavras de forma separada. Também sugere-se o re-treinamento dos modelos

com o uso de pesos diferentes para expressões que possam ser consideradas mais relevantes. No exemplo citado, a expressão "eu fico triste" pode ser testada com um peso diferente do que as demais expressões da frase.

REFERÊNCIAS BIBLIOGRÁFICAS

DIAS, Talita; LAGE, Ludmila; RIBEIRO, Raphael; ROCHA, Gustavo; RODRIGUES, Junio; SANTOS, Thiago; FRANCO, Glaura; LOSCHI, Rosangela & BRAGA, Mauro (2008). Cursos diurnos e noturnos: Fatores de aprovação no vestibular da UFMG. *Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, Cadernos de Pesquisa, v.38, n.133, p.127-146.*

DOBSON, Annette & BARNETT, Adrian (2008). An introduction to Generalized Linear Models. *Editora: Chapman e Hall/CRC.*

DUQUE, José; RAYMUNDO, Abner & NETO, Pedro (2018). Uma aplicação de Big Data para classificação de sentenças depressivas no Twitter. *Revista H-Tec Humanidades e Tecnologia, v.2, n.1, p.82-95.*

HACK, Augusto; NUNES, Luis; SILVA, Matheus & LIMA, Thiago (2013). Text Mining. *Universidade Federal de Santa Catarina.*

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor & TIBSHIRANI, Robert (2013). An Introduction to Statistical Learning: With Applications in R. *Editora: Springer*.

LIAW, Andy & WIENER, Matthew (2002). 'randomForest'. Classification and Regression by randomForest. *Disponível em: CRAN.R-project.org/doc/Rnews.*

MARTINEZ, Edson; LOUZADA, Francisco & PEREIRA, Basílio. (2003). A curva ROC para testes diagnósticos. *Caderno de* Saúde *Coletiva, Rio de Janeiro, n.11, p.7-31.*

MENDES, Augusto; PASSADOR, Rafael & CASELI, Helena. (2021). Identificando sintomas de depressão em postagens do Twitter em português do Brasil. *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana.*

MEYER, David; DIMITRIADOU, Evgenia; HORNIK, Kurt; WEINGESSEL, Andreas & LEISCH, Friedrich (2022). 'e1071'. _e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071). *Disponível em: CRAN.R-project.org/package=e1071*.

Ministério da Saúde. *Disponível em: www.gov.br.*

NETTO, Kayque; GOMES, Rogério; SANTOS, Bruno & SILVA, Edson (2021). Detecção de perfis sintomáticos de depressão no Twitter utilizando aprendizado de máquina. *Congresso Brasileiro de Inteligência Computacional.*

PNS, Pesquisa Nacional de Saúde (2020). Percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal. *Coordenação de Trabalho e Rendimento, IBGE, p.69. Disponível em: www.ibge.gov.br.*

Python Software Foundation. *Python Language Reference, v.2.7. Disponível em: www.python.org.*

R Core Team (2022). R: The R Project for Statistical Computing. *Disponível em: www.R-project.org.*

R Core Team (2022). 'stats'. The 'stats' package is part of R. *Disponível em: www.R-project.org.*

ROESSLEIN, Joshua (2020). 'tweepy'. An easy-to-use Python library for accessing the Twitter API. *Disponível em: www.tweepy.org.*

SOUZA, Paulo (2008). Uma análise em possíveis casos de patologias médicas utilizando a decisão médica em busca de melhor precisão de resposta na Web. *Artigo apresentado no 14º Congresso Internacional ABED de Educação a Distância 'Mapeando o Impacto da EAD na Cultura do Ensino-Aprendizagem'.*

Statista (2022). Twitter - Statistics & Facts . *Disponível em: www.statista.com.*

TACONELI, Cesar (2008). Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade de entropia. *Tese apresentada à Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo.*

Twitter. *Disponível em: about.twitter.com.*

Twitter Developer API. *Disponível em: developer.twitter.com.*

WHO, World Health Organization (em português: OMS, Organização Mundial da Saúde). *Disponível em: www.who.int.*

WIJFFELS, Jan (2022). 'UDPipe'. *_udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency. Disponível em: www.CRAN.R-project.org/package=udpipe.*

YAZDAVAR, Amir; AL-OLIMAT, Hussein; EBRAHIMI, Monireh; BAJAJ, Goonmeet; BANERJEE, Tanvi; THIRUNARAYAN, Krishnaprasad; PATHAK, Jyotishman & SHETH, Amit (2017). Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. *Proc IEEE ACM Int Conf Adv Soc Netw Anal Min. 2017 julho-agosto; 2017: 1191-1198.*

ZANCHINI, Vinícius (2019). Criação de um modelo de classificação de tweets depressivos utilizando máquina de vetores suporte. *Trabalho de conclusão de curso apresentado à Faculdade de Engenharia Elétrica da Universidade Federal de Uberlândia.*