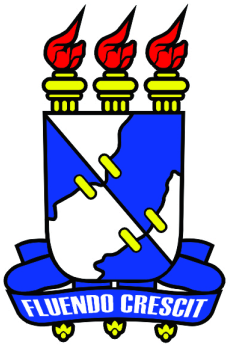

Universidade Federal de Sergipe

Pró-Reitoria de Pós-Graduação e Pesquisa

Programa de Pós-graduação em Engenharia Elétrica

Ataíde Mateus Gualberto dos Santos

Um estudo sobre a extração de características em dados do *Twitter* na tarefa de detecção de depressão

SÃO CRISTÓVÃO - SE

2024

Ataíde Mateus Gualberto dos Santos

Um estudo sobre a extração de características em dados do *Twitter* na tarefa de detecção de depressão

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica - PROEE, da Universidade Federal de Sergipe, como parte dos requisitos necessários para a obtenção do título de Mestre em Engenharia Elétrica

Orientador: Jugurta Rosa Montalvão Filho

SÃO CRISTÓVÃO - SE

2024

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA-PROEE

TERMO DE APROVAÇÃO

“Um estudo sobre a extração de características em dados do *Twitter* na tarefa de detecção de depressão”

Discente:

Ataíde Mateus Gualberto dos Santos

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Sergipe, como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Aprovada pela banca examinadora composta por:

Prof. Dr. Elyson Ádan Nunes Carvalho (PROEE/UFS)
Presidente

Prof. Dr. Jânio Coutinho Canuto (PROEE/UFS)
Examinador Interno

Prof. Dr. Héctor Julián Tejada Herrera (DPS/UFS)
Examinador Externo

Ataíde Mateus Gualberto dos Santos
Discente

Cidade Universitária “*Prof. José Aloísio de Campos”,* 26 de agosto de 2024.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Santos, Ataíde Mateus Gualberto dos
S237e Um estudo sobre a extração de características em dados do *Twitter* na tarefa de detecção de depressão / Ataíde Mateus Gualberto dos Santos ; orientador Jugurta Rosa Montalvão Filho. - São Cristóvão, 2024.
55 f. : il.

Dissertação (mestrado em Engenharia Elétrica) – Universidade Federal de Sergipe, 2024.

1. Depressão mental. 2. Aprendizado do computador. 3. Redes sociais. 4. Levantamentos linguísticos. I. Montalvão Filho, Jugurta Rosa orient. II. Título.

CDU 621.3:616.89-008.454

Agradecimentos

Agradeço à CAPES por financiar a maior parte do desenvolvimento deste trabalho. Ao meu orientador, Jugurta, por toda a atenção, paciência, ideias e por seu jeito leve de conduzir todo o processo de pesquisa. Aos integrantes do BioChaves por me aceitarem e me fazerem sentir em casa. A Israel, por ter me ajudado a conseguir uma oportunidade de estágio internacional e estar sempre disposto a me ajudar. A Gabriel Francisco, pelas conversas no Labcom, pelas ajudas nas revisões textuais e por renovar meu interesse pela matemática. A Mayane, pelas conversas, cafezinhos e todo apoio técnico, burocrático e estrutural. Aos professores que fizeram parte da minha formação. À professora Jislane Menezes e aos colegas Lucas Macena e Viviane Andrade pela ajuda na rotulação dos dados.

Ao meu pai, Valdemar Gualberto, por todo suporte e preocupação. À minha avó, Maria Teresa, e seu santo refúgio, por onde incontáveis vezes me encontrei após estar perdido. À minha irmã, Mônica Gualberto (Tance), por todo apoio, amor e carinho, mesmo que de longe. À minha namorada e futura esposa, Bruna Silva, por toda dedicação, paciência, cuidado, carinho e amor, sempre me incentivando a ir mais longe.

E a você, mãe, Josefa de Jesus Santos, que mesmo não estando mais aqui, se faz presente em minhas lembranças e no meu coração.

Resumo

A depressão é uma condição mental que afeta milhões de pessoas em todo o mundo, manifestando-se por meio de sentimentos persistentes de tristeza, falta de interesse e alterações nos padrões de pensamento e comportamento. Com o aumento do uso das redes sociais, tornou-se possível identificar sinais de depressão através das postagens dos usuários, oferecendo novas oportunidades para o estudo de indicadores linguísticos associados a essa condição. Este estudo explora métodos de extração de características em dados de redes sociais, com o objetivo de identificar sinais de depressão em usuários do *Twitter*. A pesquisa começou com a criação de uma base de dados a partir de postagens públicas, seguida pela aplicação de técnicas de pré-processamento de dados. A Teoria Cognitiva Comportamental foi integrada ao referencial teórico, fornecendo a base para a extração manual de características. A seleção das características mais relevantes foi realizada por meio de testes de hipóteses combinados com o classificador AdaBoost. Entre os principais indicadores encontrados, destacou-se o uso frequente de palavras em primeira pessoa e o aumento nas postagens durante o período da noite por parte dos indivíduos rotulados com depressão em comparação com o grupo controle. Além disso, a análise dos dados revelou menor quantidade de palavras efetivas no vocabulário das pessoas rotuladas com depressão. Os resultados deste estudo têm potencial de contribuição para a sociedade. Profissionais da saúde podem utilizar ferramentas de triagem baseadas em dados e orientar na criação de políticas públicas de saúde mental. Além disso, essas técnicas podem auxiliar plataformas de redes sociais a identificar e apoiar usuários em risco.

Palavras-chave: depressão, extração de características, reconhecimento de padrões, aprendizado de máquina, redes sociais, indicadores linguísticos.

Abstract

Depression is a mental condition that affects millions of people worldwide, manifesting through persistent feelings of sadness, lack of interest, and changes in thought and behavior patterns. With the rise of social media, it has become possible to identify signs of depression through users' posts, offering new opportunities for the study of linguistic indicators associated with this condition. This study explores methods for feature extraction in social media data, aiming to identify signs of depression in Twitter users. The research began with the creation of a database from public posts, followed by the application of data preprocessing techniques. Cognitive Behavioral Theory was integrated into the theoretical framework, providing the basis for manual feature extraction. The selection of the most relevant features was carried out through hypothesis testing combined with the AdaBoost classifier. Among the key indicators found, the frequent use of first-person words and an increase in posts during nighttime by individuals labeled as depressed, compared to the control group, stood out. Additionally, data analysis revealed a lower number of effective words in the vocabulary of those labeled with depression. The results of this study have the potential to contribute to society. Health professionals can use data-driven screening tools and guide the creation of public mental health policies. Moreover, these techniques can assist social media platforms in identifying and supporting users at risk.

Keywords: depression, feature extraction, pattern recognition, machine learning, social media, linguistic indicators.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AdaBoost	Adaptive Boosting
CEP	Comitê de Ética em Pesquisa
CONEP	Comissão Nacional de Ética em Pesquisa
LIWC	Linguistic Inquiry and Word Count
NF	Negativo Falso
NV	Negativo Verdadeiro
OMS	Organização Mundial da Saúde
PA	Pensamento Automático
PAC	Probably Approximately Correct
PF	Positivo Falso
PV	Positivo Verdadeiro
TCC	Terapia Cognitiva-Comportamental
TF-IDF	Term Frequency-Inverse Document Frequency

Sumário

1. Introdução

Nos últimos anos, os transtornos mentais têm se tornado uma preocupação significativa na saúde pública. Segundo a OMS[1], 970 milhões de pessoas (cerca de 12,5% da população mundial) sofriam algum tipo de transtorno mental em 2019.

Existem diversos tipos de problemas de saúde mental. Esquizofrenia, transtornos de alimentação, estresse pós-traumático, transtorno bipolar, ansiedade e depressão são alguns exemplos de desordens mentais, sendo as duas últimas as mais frequentes com, respectivamente, 301 milhões e 280 milhões de pessoas acometidas em 2019, segundo a OMS[2].

No Brasil, um estudo recente [1] indicou que 11,3% dos 784 mil participantes foram diagnosticados com depressão, havendo maior prevalência no grupo das mulheres (14,7%) do que no dos homens (7,3%).

Transtornos como a depressão possuem consequências devastadoras tanto para os pacientes quanto para suas famílias e os transtornos depressivos estão entre os mais comuns entre as doenças mentais, caracterizados por tristeza, perda de interesse e prazer, sentimentos de culpa ou baixa autoestima, distúrbios do sono ou apetite, cansaço e falta de concentração[3].

Comumente, a depressão é detectada por um profissional da saúde mental, como um psicólogo, e seu diagnóstico é realizado por meio de sessões terapêuticas [2]. Nessas sessões, o profissional se atenta às características fisiológicas, motoras e linguísticas apresentadas pelos pacientes [3]. No caso do transtorno depressivo leve, por exemplo, os aspectos afetivos e cognitivos são reconhecidos principalmente por meio da linguagem [4].

De fato, dada a importância da linguagem na identificação de transtornos depressivos, tem sido crescente o interesse por indicadores linguísticos que possam ajudar no diagnóstico de pessoas com depressão. No estudo conduzido em [4], pessoas com depressão leve mostraram um uso maior de coloquialismos, repetições, e sentenças de cláusula única em linguagem escrita. Os autores também notaram que, apesar dos pacientes proverem respostas mais longas, elas continham expressões incompletas com omissão de palavras (elipses) e enunciados reduzidos.

Similarmente na pesquisa conduzida por [5], foi realizada uma análise quantitativa de indicadores linguísticos em pacientes que enfrentaram transtornos de depressão maior, tanto

[1] WORLD HEALTH ORGANIZATION. Mental disorders. Disponível em: https://www.who.int/news-room/fact-sheets/detail/mental-disorders. Acesso em: 04 jun. 2024.
[2] Ibidem.
[3] WORLD HEALTH ORGANIZATION. Depression. Disponível em: https://www.who.int/news-room/fact-sheets/detail/depression. Acesso em: 04 jun. 2024.

na linguagem falada quanto na escrita. Os resultados dos indicadores léxicos e sintáticos revelaram que os indivíduos diagnosticados com depressão maior usaram frases de estrutura mais simples, apresentando uma tendência a expressões mais curtas e truncadas, frequentemente acompanhadas por repetições.

Tanto nos dois estudos anteriores como em [6], o uso de pronomes em primeira pessoa demonstra ser um indicador determinante na classificação de pessoas com depressão. Os autores em [4] sugerem que o alto uso de pronomes nessa categoria está relacionado a perda de significados específicos na fala e também pode ser interpretado como uma manifestação de empobrecimento semântico.

Outro fator que se mostra importante é o sexo do paciente [1]. Na revisão conduzida por [7], foi verificado que, entre alunos do ensino superior, as mulheres estavam mais propensas a ter sintomas depressivos. Na mesma pesquisa ainda encontraram correlação entre sintomas de depressão e níveis padrões de depressão, traços de personalidade (neuroticismo e psicoticismo), eventos estressantes de vida, abuso na infância (incluindo abuso sexual) e pensamentos negativos ou ruminação.

Já em [8], os autores utilizaram dados provenientes da rede social *Twitter* para detectar depressão. Eles analisaram medidas como linguagem, emoção, estilo de escrita e engajamento do usuário. Os autores mostraram que os usuários depressivos mostraram ter um alto nível de foco em si mesmos, expressaram mais conteúdo religioso e engajaram em conversas com indivíduos pertencentes às mesmas redes de contatos do que o grupo de controle.

Nesse contexto, é notável que as redes sociais tornaram-se uma área relevante na pesquisa de saúde pública, permitindo a observação do bem-estar emocional das pessoas. Na pesquisa conduzida por [9], por exemplo, os autores utilizaram *tweets* e as atividades de usuário da rede social *Twitter* para desenvolver classificadores capazes de detectar prováveis usuários depressivos. Semelhantemente, em [10], os autores extraíram características psicolinguísticas de comentários do *Facebook* e utilizaram diversos algoritmos de classificação para a tarefa de detecção de depressão.

1.1 Motivação

É notável, nos trabalhos apresentados acima, que a extração de características linguísticas é considerada uma tarefa secundária, planejada de tal forma que sua execução em tempo real não seja viável, por conta de sua complexidade, ou necessita de ferramentas proprietárias a fim de realizar suas análises em tempo hábil. De fato, alguns dos indicadores

linguísticos utilizados em [9, 10] são extraídos pela ferramenta proprietária LIWC (*Linguistic Inquiry and Word Count*), enquanto que outros trabalhos [4, 5] realizam a extração manual de características.

Embora a extração manual de características possua vantagens, não há trabalhos que combinem extração automática e manual, baseada no conhecimento do fenômeno - neste caso, a depressão - de características de forma integrada.

Além disso, a maioria dos estudos se concentra na língua inglesa, apesar de haver um aumento na quantidade de estudos que utilizaram dados de redes sociais em português brasileiro [11 - 13]. Isso revela uma escassez de pesquisas que abordem a língua portuguesa e as particularidades da cultura brasileira.

Este trabalho propõe uma abordagem que combina a extração automática e manual de características linguísticas, integrando o conhecimento específico do fenômeno da depressão. Aplicando essa metodologia em dados de redes sociais em português brasileiro, busca-se preencher essa lacuna nas pesquisas atuais e melhorar a compreensão e detecção da depressão na língua portuguesa, considerando as particularidades culturais do Brasil. Além disso, realizar pesquisas na língua nativa, neste caso em português brasileiro, facilita o acesso e aumenta o interesse sobre o tema, uma vez que, a busca por artigos em outros idiomas pode ser mais demorada e trabalhosa devido à necessidade constante de tradução.

2. Objetivos

O objetivo deste trabalho é identificar as características mais relevantes na classificação de pessoas com depressão a partir de suas postagens em português brasileiro da rede social *Twitter* (atual X). Especificamente, este estudo propõe-se a responder às seguintes perguntas:

- É possível desenvolver um classificador capaz de imitar a tarefa de rotulação manual de postagens como (controle) escritas por pessoas sem depressão ou (positivos) escritas por pessoas sofrendo de depressão, a partir de características extraídas de *tweets*?
- Quais são as características mais importantes que podem ser extraídas de forma automática para realizar a classificação acima?
- A partir do estudo da teoria sobre a depressão, é possível encontrar de forma manual outros indicadores de depressão nos dados?

3. Referencial Teórico

3.1 Modelo Cognitivo de Beck

Aaron Temkin Beck, conhecido como o pai da teoria cognitiva, começou sua carreira como psicanalista. Contudo, por volta da década de 1950, percebeu que a teoria da psicanálise não explicava adequadamente suas observações clínicas no tratamento de pessoas com depressão [14]. Em 1961, Beck e colaboradores propuseram um inventário para aferir a gravidade dos sintomas depressivos por meio de um questionário, iniciando assim uma série de publicações que culminaram na criação da Terapia Cognitiva-Comportamental (TCC) e no desenvolvimento do modelo cognitivo da depressão [15, 16].

Ao conceber seu modelo cognitivo, Beck identificou padrões de pensamento recorrentes em seus pacientes, que os levavam a conclusões equivocadas sobre suas situações [17]. Esses padrões de pensamento influenciam a forma como estímulos externos são processados, resultando em erros sistemáticos, ou vieses de interpretação, como abstração seletiva, generalização excessiva e auto-atribuições negativas [17].

A cognição, entendida como o processo que envolve deduções sobre experiências e sobre a ocorrência e controle de eventos futuros, é central para entender os vieses de interpretação [18]. Existem três níveis de cognições identificadas pela TCC, sendo eles: pensamentos automáticos (PAs), crenças intermediárias e crenças centrais [19]. Os PAs fazem parte de um fluxo de processamento cognitivo subjacente ao processamento consciente. Geralmente, são particulares ao indivíduo e ocorrem de maneira rápida através da avaliação do significado de episódios da vida [20].

As crenças intermediárias são regras, atitudes ou suposições. São afirmações do tipo “se... então” que se apresentam de modo inflexível e imperativo [21]. Também podem ser chamadas de pressupostos subjacentes ou condicionais. Estas formam um conjunto de crenças, em geral, coerentes que oferecem apoio às crenças centrais com as quais apresentam relação [22]. De acordo com [23], todas as pessoas têm um conjunto de crenças condicionais que foram aprendidas e somadas umas às outras ao longo da vida, no intuito de dar significado ao mundo.

Dessa forma, uma mesma situação pode gerar reações distintas em diferentes pessoas, e uma mesma pessoa pode reagir de diversas maneiras a uma mesma situação [18]. Isso ocorre porque a interpretação individual das situações da vida é mais relevante para as emoções do que os próprios eventos, de acordo com [18]. Segundo Beck, para entender as

reações emocionais a um acontecimento, é necessário distinguir o significado público de uma ocorrência (definição formal e objetiva) de seu significado pessoal (interpretação individual) [24].

Para Beck, tais significados são estabelecidos e selecionados a partir de esquemas, ou crenças centrais, prévios que os sujeitos possuem sobre si mesmos, com a natureza do conteúdo do esquema sendo influenciada por eventos externos [25]. Na teoria cognitiva, o termo esquema refere-se a um padrão imposto à realidade ou à experiência para ajudar as pessoas a explicá-la, mediando a percepção e guiando as respostas. É uma estrutura cognitiva que molda as imagens, interpretações e percepções do indivíduo [16 - 18, 24].

Seguindo essa premissa, Young propõe que um esquema pode ser adaptativo, auxiliando positivamente o sujeito em suas situações de vida, ou desadaptativo, bloqueando parcial ou totalmente ações e comportamentos funcionais. Os esquemas podem ser rígidos ou flexíveis, formados em qualquer momento da vida, com diferentes níveis de gravidade e penetração [26].

Em [19], os autores ressaltam ainda que as crenças centrais representam os mecanismos desenvolvidos pelas pessoas para lidar com as situações cotidianas, ou seja, a maneira como os indivíduos percebem a si mesmos, aos outros e ao mundo, e ao futuro, sendo esta percepção chamada de tríade cognitiva.

Quando um pensamento se associa a um esquema, podem ocorrer distorções cognitivas, incluindo erros no conteúdo cognitivo (significado), no processamento cognitivo (elaboração de significado), ou em ambos [18]. Um esquema de fobia social poderia englobar, por exemplo, crenças centrais disfuncionais como as apresentadas no Quadro 1 a seguir:

Quadro 1 - Exemplo de crença central associada ao esquema de fobia social.

Esquema (fobia social)		
Crença central (sobre si)	Crença central (sobre os outros)	Crença central (sobre o mundo)
Sou incapaz de causar uma boa impressão	Os outros estão sempre me julgando negativamente	A sociedade é hostil

Fonte: Adaptado de [25].

As crenças centrais são compreensões duradouras, tidas como verdades, enraizadas, inquestionáveis e capazes de influenciar a percepção pessoal. Ao se deparar com algum evento que ative esta crença, o indivíduo começa a avaliar as situações às luzes desta crença.

Quando ocorre algo que contraria as verdades estabelecidas por ela, o sujeito tende a desconsiderar este acontecimento ou a deformá-lo, reafirmando as ideias trazidas pela crença central [27]. A Figura 1 apresenta uma representação do modelo cognitivo de Beck, o qual evidencia o processo de retroalimentação de uma crença central.

De fato, em [17], Beck comenta sobre como um pesquisador muito bem sucedido possuía uma esquema do tipo "eu sou um fracasso completo". Seus pensamentos automáticos gravitavam em torno de como ele se considerava inferior e mal sucedido. Quando foi pedido a essa pessoa para recordar um único acontecimento que ele não considerasse um fracasso, ele não conseguiu. No mesmo trabalho, Beck descreve um paciente que chorava toda vez que era elogiado. A crença central dessa pessoa era do tipo "eu sou uma fraude" e, assim, qualquer comentário positivo reforçava a ideia de que ele estava enganando outras pessoas.

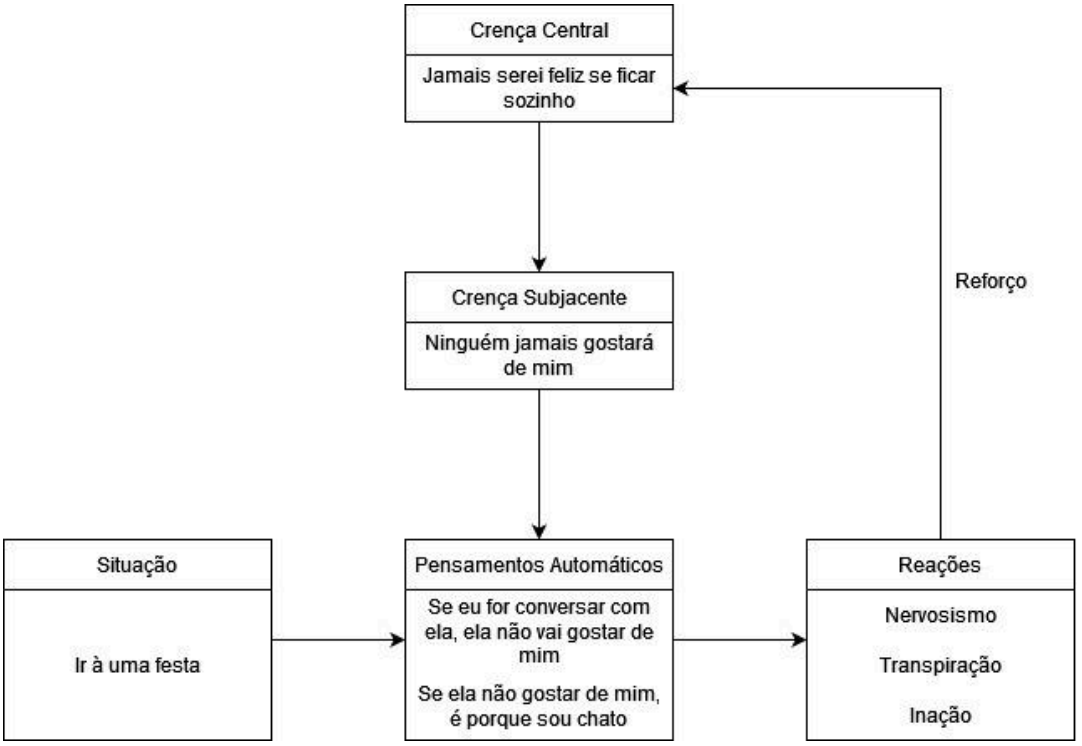

Figura 1 - Exemplo de ativação e reforço de um esquema da Teoria Cognitiva. Fonte: Adaptado de [21].

Conforme a Figura 1, a crença central (situada na parte superior do modelo) é a base que dá origem às crenças subjacentes (ou intermediárias), as quais postulam atitudes, regras e pressupostos que influenciam a forma como os sujeitos se comportam. As crenças subjacentes oferecem material para os pensamentos automáticos, os quais dão início ao esquema mental do indivíduo e ocorrem diante de uma situação que provoque um estímulo interno ou externo [24].

Para abordar essas questões, a TCC utiliza uma variedade de técnicas. Um dos métodos mais comuns é a reestruturação cognitiva, que envolve a identificação e a análise crítica dos pensamentos automáticos e das crenças subjacentes [18]. O terapeuta auxilia o paciente a questionar a validade desses pensamentos e a substituí-los por interpretações mais realistas e adaptativas [24].

Além disso, a TCC incorpora a exposição gradual a situações temidas, treinamento de habilidades sociais e outras intervenções comportamentais para promover mudanças duradouras. A meta é ajudar os pacientes a desenvolver um conjunto de crenças mais equilibradas e flexíveis, permitindo-lhes responder de maneira mais saudável aos desafios da vida [27].

Além dos aspectos cognitivos e comportamentais, a linguagem desempenha um papel crucial na manifestação e no tratamento da depressão [4]. A forma como os indivíduos expressam seus pensamentos e emoções pode refletir e influenciar seus estados mentais [4]. Estudos têm mostrado que a linguagem utilizada por pessoas com depressão tende a ser mais negativa, autorreferente e absolutista, reforçando padrões de pensamento disfuncionais [4 - 6]. Analisar a linguagem pode, portanto, ajudar na compreensão do estado emocional de um indivíduo e guiar intervenções terapêuticas mais eficazes.

As redes sociais, como plataformas de comunicação amplamente acessíveis e instantâneas, se tornaram espaços onde as pessoas se sentem encorajadas a expressar seus sentimentos de maneira direta e contínua. Nesse contexto, os sinais de alerta para a depressão se tornam mais evidentes, uma vez que esses comportamentos expressivos estão frequentemente acompanhados por padrões linguísticos negativos e autorreferentes e, assim sendo, ao analisar a frequência, o contexto e a forma como as pessoas se comunicam online, é possível não apenas detectar sinais precoces de depressão, mas também compreender melhor o impacto das interações sociais digitais na saúde mental.

3.2 Depressão e Linguagem

A linguagem é crucial para compreender e tratar a depressão, pois reflete e influencia os estados mentais dos indivíduos. Estudos indicam que a análise linguística pode revelar novas perspectivas sobre o estado emocional e a eficácia das intervenções psicoterapêuticas [28-30]. Indicadores linguísticos, como pensamentos automáticos, distorções cognitivas e padrões de estilo de linguagem e de gramática específicos, são fundamentais para identificar a depressão e avaliar mudanças ao longo do tratamento [5, 24, 29].

A expressão dos pensamentos e emoções dos indivíduos pode revelar padrões disfuncionais típicos da depressão. Pacientes depressivos tendem a usar uma linguagem mais negativa, reforçando a introspecção excessiva [4 - 6]. Esses padrões linguísticos podem ser identificados por meio da predominância de pronomes de primeira pessoa e palavras negativas, que são marcadores específicos do estado emocional do paciente, auxiliando na compreensão de seu quadro mental.

Além disso, o funcionamento do cérebro é alterado durante a depressão, com algumas áreas apresentando níveis de atividade significativamente diferentes em comparação a pessoas saudáveis [31, 32]. Essas mudanças neurofisiológicas influenciam diretamente a interpretação verbal, contribuindo para o "pensamento lento" característico da condição.

A literatura sugere que indivíduos deprimidos tendem a focar excessivamente em si mesmos, intensificando emoções negativas e autoacusação [33]. Estudos como o de [34] destacam que poetas suicidas utilizavam mais pronomes de primeira pessoa e menos palavras relacionadas ao coletivo social, refletindo um foco no eu e uma falta de integração social.

Esse foco excessivo em si mesmo e a expressão de emoções negativas também são observados em interações online. Nas redes sociais, há um crescente número de pesquisas sobre o uso de indicadores linguísticos para detectar sinais de depressão. Em [8], os autores coletaram dados do *Twitter* de pessoas diagnosticadas com depressão e desenvolveram um algoritmo capaz de prever o risco de uma pessoa ter depressão antes do diagnóstico. Além disso, descobriram que indivíduos com depressão apresentam redução nas atividades sociais, aumento na negatividade expressa e maior preocupação com relacionamentos e saúde.

De forma semelhante, em [9], os autores usaram vários indicadores extraídos de dados do *Twitter* para criar classificadores. Eles geraram uma matriz termo-documento onde cada elemento indica a frequência de uma palavra em uma postagem para cada usuário. A partir dessas matrizes, eles extraíram características como o uso de pronomes em primeira pessoa, TF-IDF (frequência do termo-inverso da frequência nos documentos), sentimentos das palavras (usando a ferramenta LIWC) e distribuições das palavras mais frequentes. Dessa forma, desenvolveram um classificador binário com acurácia de aproximadamente 80% e identificaram que a análise de sentimentos, atividades sociais e o agrupamento de palavras em sinônimos foram os indicadores mais importantes para a tarefa de classificação.

Em outro estudo, [9], os autores utilizaram indicadores emocionais, temporais e linguísticos de dados extraídos do *Facebook* para desenvolver um algoritmo capaz de detectar depressão. Com isso, criaram um classificador com revocação de 99% ao utilizar todos os

indicadores extraídos. Descobriram também que a maioria das pessoas depressivas publicava suas postagens entre meia-noite e meio-dia (durante a noite).

Todos os estudos acima realizam extração de características apenas de forma manual, com exceção de [9] que usa a técnica do ganho de informação. A extração manual pode ser limitada pela subjetividade e pelo volume de dados. Para superar essas limitações e melhorar a detecção de padrões linguísticos associados à depressão, é importante explorar também métodos automáticos. Técnicas de aprendizado de máquina podem ajudar nessa questão ao permitir a seleção automática de características. A próxima seção abordará como os métodos de *Boosting* podem ser úteis para enfrentar esses desafios.

3.3 Boosting

Com suas origens no modelo de aprendizado provavelmente aproximadamente correto (*Probably Approximately Correct* - PAC) [35], o termo *boosting*, também conhecido como reforço, se refere ao método de "combinar" regras de classificação fracas de forma a criar uma regra mais acurada, ou forte [35]. A essência do *boosting* está na ideia de que múltiplos modelos simples, quando integrados de forma adaptativa, podem superar suas limitações individuais, resultando em um sistema eficaz.

Um classificador, em aprendizado de máquina, é um sistema que associa vetores de características (dados de entrada) a rótulos de classes (saídas discretas) [36]. Para ser eficaz, ele deve ser treinado com exemplos prévios, aprendendo padrões que permitam generalizar para novos dados. Formalmente, dados um conjunto de treinamento $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, em que $x_i \in X$ e $y_i \in Y$ são exemplos dos espaços de características e de rótulos das classes, respectivamente, o objetivo do classificador é construir uma função $H: X \rightarrow Y$ que mapeie corretamente exemplos não vistos durante o treinamento, conforme o modelo PAC [36 - 38].

A distinção entre classificadores fortes e fracos é fundamental. Um classificador forte é definido pelos parâmetros ε (margem de erro aceitável) e δ (tolerância à incerteza). Com isso, um classificador forte é aquele que dados $\varepsilon, \delta > 0$, ambos reais, e acesso a exemplos de treinamento aleatórios, possui taxa de erro de no máximo ε com probabilidade de pelo menos $1 - \delta$ [35, 38, 40, 41].

Por outro lado, um classificador fraco possui desempenho apenas marginalmente melhor que uma escolha aleatória. No caso binário, por exemplo, um classificador fraco

satisfaz as mesmas condições do classificador forte, mas apenas para $\varepsilon \geq \frac{1}{2} - \gamma$ [34, 35]. O parâmetro $\gamma \in \mathbb{R}$ é uma pequena margem que garante vantagem sobre o acaso [35, 40, 41].

Continuando a discussão sobre *boosting*, existem infinitas formas de combinar classificadores. Uma delas, por exemplo, é realizar a soma das saídas de cada classificador de forma análoga a uma votação majoritária. No entanto, este modo possui a desvantagem de dar pesos iguais a todos os classificadores, independente de suas taxas de erro. Assim, a forma mais utilizada é uma soma ponderada das saídas dos classificadores [42]. Isso é equivalente a um sistema de votação em que os votos de alguns possuem peso maior do que de outros [35, 42].

Para compreender como o *boosting* é utilizado na literatura, o método AdaBoost foi estudado. O AdaBoost, assim como outros métodos de *boosting*, possui características e vantagens próprias, mas todos compartilham a ideia fundamental de melhorar a performance combinando múltiplos classificadores fracos.

3.3.1 Adaptive Boosting

O *Adaptive Boosting*, abreviado para AdaBoost, é um algoritmo de aprendizado de máquina capaz de gerar um classificador forte a partir de classificadores fracos [35]. Especificamente, o AdaBoost realiza reamostragens iterativas dos exemplos de treinamento, dando peso maior aos classificadores que acertam mais os exemplos que possuem os maiores números de erros nas iterações anteriores. Na Figura 2 está descrito o algoritmo de treinamento do AdaBoost [43].

Entradas: Dados um conjunto de treinamento $(x_1, y_1), \ldots, (x_M, y_M)$, em que $x_i \in X$, $y_i \in Y = \{-1, +1\}$, e um conjunto L de classificadores $\{h_1, \ldots, h_L\}$.

1. $D_1(i) = 1$, para i := 1 até M
2. para t := 1 até T, faça
    3. Selecione em L o classificador h_j que minimiza

$$D_j^e = \sum_{y_i \neq h_j(x_i)} D_t(i)$$

    e atribua $D_t^e = D_j^e$ e $h_t = h_j$.

    4. Atribua

$$\alpha_t = \frac{1}{2} ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

    ao classificador h_t, em que $\epsilon_t = \frac{D_t^e}{\sum_{i=1}^{m} D_t(i)}$.

    5. Atualize os pesos D_t de acordo com

$$D_{t+1}(i) = D_t(i) \cdot exp(-\alpha_t \cdot y_i \cdot h_t(x_i)).$$

6. fim para
7. Retorne $H_T(x) = sinal\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

Figura 2 - Algoritmo de treinamento do AdaBoost. Fonte: Adaptado de [43].

Uma das principais ideias do algoritmo é atribuir pesos para cada exemplo de treinamento. O valor do peso no exemplo i na iteração t é denotado por $D_t(i)$. Todos os pesos são inicializados igualmente com 1 e, a cada iteração, os pesos dos exemplos classificados incorretamente aumentam de forma que o algoritmo base (classificador fraco) seja forçado a focar nos exemplos mais difíceis de se rotular no conjunto de treinamento [35].

A função do algoritmo base é encontrar o melhor classificador $h_t: X \rightarrow Y$ a partir dos exemplos de treinamento amostrados a partir de D_t. Para problemas binários, o melhor classificador é encontrado ao se escolher um h_t que minimiza a taxa de erro

$$\epsilon_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i], \quad (1)$$

em que $i \sim D_t$ indica que o exemplo i será amostrado da distribuição D_t. É importante salientar que, pela sua construção, pode acontecer de $\sum_{i=1}^{M} D_t(i) \neq 1$, sendo uma simples normalização suficiente para resolver esse problema. Na prática, porém, ϵ_t é estimada como

$$\hat{\epsilon}_t = \frac{D_t^e}{\sum_{i=1}^{m} D_t(i)} \text{ [40].} \quad (2)$$

Vale ressaltar que essa taxa de erro não é a do AdaBoost, mas sim a do melhor classificador fraco na iteração t. Quando o classificador base tiver sido escolhido, o AdaBoost calcula um peso $\alpha_t \in \mathbb{R}$ que mede a importância de h_t. É possível mostrar que

$$\alpha_t = \frac{1}{2} ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \quad (3)$$

minimiza a função de custo do AdaBoost, descrita na Equação 4, para problemas de classificação binária [43]:

$$E_t = \sum_{i=1}^{m} exp\left(-y_i \cdot H_t(x_i)\right) \quad (4)$$

A prova encontra-se no Apêndice A. A distribuição D_t é, então, atualizada conforme a regra

$$D_{t+1}(i) = D_t(i) \cdot exp(-\alpha_t \cdot y_i \cdot h_t(x_i)) \quad (5)$$

e o classificador final H_T é uma votação ponderada dos T classificadores base, com α_t sendo o peso do classificador h_t [43]. Para ilustrar o funcionamento do AdaBoost, considere o seguinte problema brinquedo.

Problema Brinquedo 1: Utilize o AdaBoost para identificar quando um ponto está dentro do círculo descrito por C: $u^2 + v^2 \leq \frac{200}{\pi}$, utilizando apenas classificadores lineares. Considere que os pontos fora do círculo estejam dentro de um quadrado Q de lado 20, centralizado na origem, de forma que o círculo fique circunscrito dentro do quadrado e que a área de C seja metade da área de Q.

Solução: Inicialmente, é possível criar um conjunto de treinamento amostrando m pontos $x_i \in Q$. Em seguida atribui-se +1 caso $x_i \in C$ ou -1, caso contrário. Após isso, é possível amostrar L classificadores lineares do tipo $sinal(au + bv + c)$, em que $a, b, c \in [-10, 10]$. A escolha dos intervalos para a, b e c foi feita experimentalmente de

forma que a maioria das retas amostradas com parâmetros dentro desta faixa intersectam pelo menos um ponto de Q. Para esta solução foram utilizados $M = 2000$ pontos e $L = 10000$ classificadores. Na Figura 3 é possível observar os pontos amostrados assim como um dos classificadores.

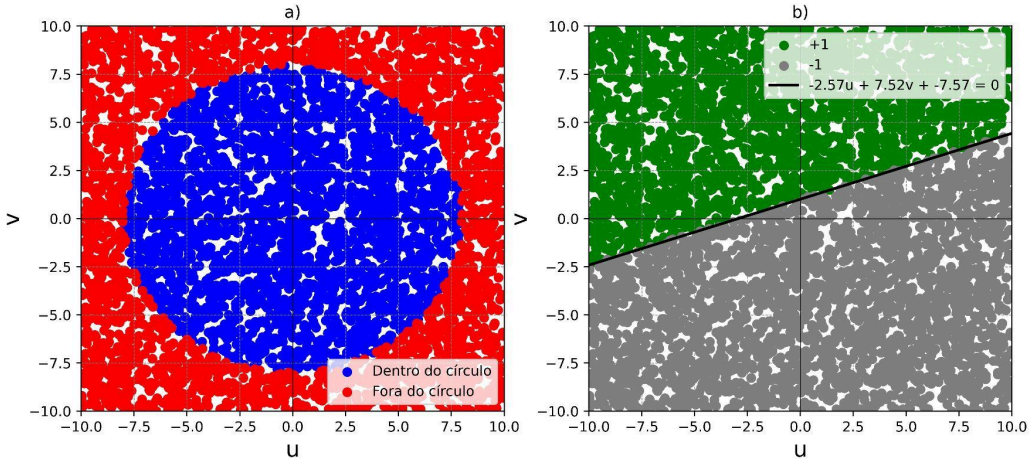

Figura 3 - a) Pontos amostrados uniformemente juntamente com seus rótulos e b) saída de um classificador linear amostrado aleatoriamente. Fonte: Autoria própria.

Com isso, o treinamento do AdaBoost começa atribuindo $D_1(i) = 1, \forall i \in [1, M]$. Em seguida, deve-se procurar qual classificador linear minimiza D_t^e. Fazer isso é equivalente à escolher o classificador fraco que minimiza a taxa de erro ϵ_t, pois D_t^e e ϵ_t são proporcionais. Logo após, calcula-se α_t e atualiza-se D_{t+1} utilizando as Equações 3 e 5, respectivamente. Repete-se este processo por um número fixo de T iterações ou até que a taxa de erro esteja baixa "o bastante". A Figura 4 demonstra a evolução do treinamento do algoritmo.

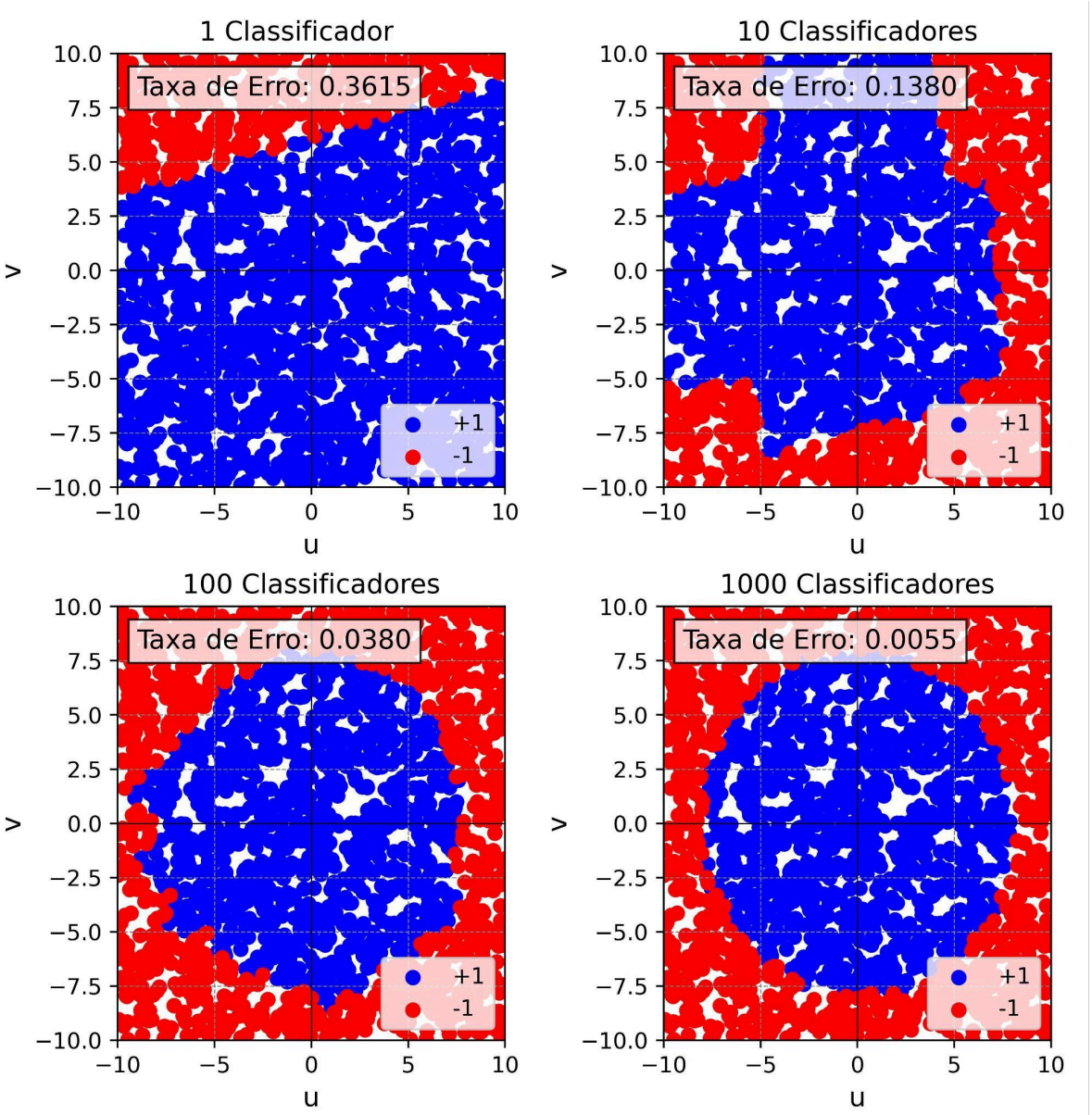

Figura 4 - Evolução da fronteira de classificação resultante e do erro de treinamento do AdaBoost. Fonte: Autoria própria.

□

Analisando a Figura 4, é perceptível que, ao combinar fronteiras de decisão lineares, é possível construir uma fronteira bastante não linear. Observa-se na Equação 3 que se $\epsilon_t = \frac{1}{2}$, o que no problema de classificação binária corresponde a um classificador aleatório, então $\alpha_t = 0$. Em outras palavras, a importância do classificador aleatório é nula. De fato, é possível verificar a importância de um classificador qualquer analisando apenas sua taxa de erro. A Figura 5 apresenta a variação da importância α_t versus a taxa de erro ϵ_t.

Durante a fase final do treinamento do problema ilustrativo, foi observado que as importâncias dos classificadores lineares se aproximaram de zero. Este resultado é esperado e indica que os exemplos de treinamento que continuavam a ser classificados incorretamente eram extremamente desafiadores para os *L* classificadores fracos disponíveis.

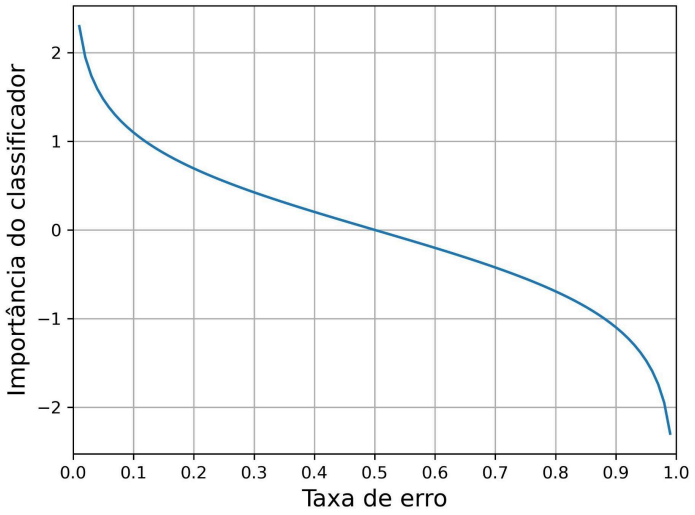

Figura 5 - Importância do classificador versus sua taxa de erro. Fonte: Autoria própria.

É possível observar que, como esperado, taxas de erro menores estão associadas a altos valores de importância. Por outro lado, taxas de erro superiores a 0,5 resultam em importâncias negativas. À primeira vista, isso pode parecer estranho, mas faz sentido teoricamente. Em problemas binários, um classificador com erro maior que 50% é, na prática, pior que um palpite aleatório. No entanto, seu "erro sistemático" pode ser convertido em uma vantagem: ao inverter suas previsões (ou seja, considerar o complemento do classificador), a taxa de erro efetiva passa a ser $1 - \epsilon$, onde ϵ é a taxa de erro original.

Para ilustrar, considere a situação de uma pessoa *A*, que é bastante azarada e acerta apenas 10% das vezes ao prever o próximo lançamento de uma moeda. Em outras palavras, *A* erra 90% das vezes. Nesse contexto, uma pessoa *B* pode se sair melhor, errando apenas 10% das vezes ao escolher a face oposta da moeda escolhida por *A*. Assim, a importância da classificação de *A* pode ser vista como negativa.

Além disso, é importante salientar que o AdaBoost, por focar nos exemplos mais difíceis de serem classificados, é sensível a *outliers*[4]. Isso ocorre porque o algoritmo atribui mais peso aos exemplos incorretamente classificados em cada iteração, o que pode levar a

[4] Um *outlier* é um exemplo nos dados que se desvia significativamente dos demais, podendo indicar variabilidade extrema, erros de medição ou fenômenos atípicos. Outliers podem influenciar análises estatísticas, tornando essencial a sua identificação e tratamento [41].

uma ênfase excessiva em dados ruidosos. Como resultado, a performance do modelo pode ser negativamente afetada por esses exemplos atípicos [40].

De fato, a Equação 3 pode ser considerada como a solução da descida mais íngreme do gradiente, onde α_t representa o passo que mais reduziria a função de custo exponencial definida na Equação 4. Essa abordagem busca minimizar a função de custo de forma agressiva, o que pode amplificar o impacto dos *outliers* no treinamento do modelo.

Para contornar esse problema, a literatura apresenta vários métodos que ajustam o processo de treinamento do AdaBoost para reduzir sua sensibilidade a *outliers*. Por exemplo, métodos como o Gentle AdaBoost, o LogitBoost e o AdaBoost.R2 introduzem modificações na função de custo ou na forma como os pesos são atualizados. Esses métodos visam limitar o impacto dos *outliers* ao suavizar a penalização associada aos exemplos difíceis, proporcionando uma abordagem mais robusta para o ajuste do modelo [37, 44, 45].

Além de sua aplicação principal, o AdaBoost pode ser utilizado, também, como uma ferramenta para seleção de características. Ao longo do treinamento, o AdaBoost ajusta a importância de cada característica indiretamente através da combinação dos classificadores fracos. Características que consistentemente ajudam a reduzir a taxa de erro em várias iterações tendem a ter mais influência no modelo final. Portanto, as características que aparecem frequentemente em classificadores que têm alta importância podem ser consideradas mais relevantes [46 - 48]. No entanto, esta não é a única forma de seleção. No tópico seguinte será abordada outra estratégia para esta tarefa.

3.4 Testes de Hipótesis

Em estatística, uma hipótese é uma afirmação sobre o valor de um parâmetro de determinada população, isto é, qualquer medida numérica calculada a partir de todos os indivíduos de uma população, como média, proporção, etc [49]. Já um teste de hipóteses é uma verificação, ou experimento, sobre a veracidade de uma afirmação, com um risco máximo associado [50]. O processo envolve quatro etapas principais: formulação das hipóteses, definição de um critério de decisão (nível de significância), cálculo de uma estatística de teste e tomada de decisão com base no p-valor.

Como inicialmente não se conhece a veracidade sobre uma hipótese, é necessário assumir uma posição cautelosa a fim de evitar erros [49]. A hipótese nula (H_0) cumpre esse papel ao assumir que a afirmação é falsa. Ela serve como referência para calcular a probabilidade dos dados observados sob a condição de inexistência do fenômeno investigado.

A hipótese alternativa (H_1), por sua vez, contradiz H_0 e só é aceita se a evidência amostral for suficientemente forte para rejeitar H_0 com um risco controlado de erro (p-valor) [51]. Para ilustrar esses conceitos, considere os dois próximos problemas brinquedo.

Problema Brinquedo 2: Um jogador profissional de jogos de azar afirma possuir a habilidade de sempre conseguir acertar qual será a face do próximo lançamento de uma moeda balanceada. Averigue se a afirmação desse jogador é verdadeira.

Solução: É necessário criar um experimento para testar a veracidade dessa afirmação. Um possível experimento é observar várias vezes o palpite e a face da moeda quando o jogador lançá-la. Se o jogador errar apenas uma vez, será possível dizer que a afirmação dele é falsa. No entanto, dado que a afirmação é bastante assertiva, com o jogador alegando acertar sempre, surgem problemas. Se forem observados n lançamentos, a probabilidade do jogador acertar todas as vezes por acaso é de 2^{-n}. Isso significa que existe a possibilidade de o jogador estar mentindo e essa mentira não ser detectada. É importante notar que 2^{-n} é uma função que decresce exponencialmente conforme n aumenta, no entanto nunca atinge zero.

Portanto, deve-se considerar a possibilidade de não se conseguir identificar um caso único de sorte (ou seja, o jogador pode estar acertando por acaso). Contudo, é possível reduzir arbitrariamente a probabilidade de erro, aumentando a quantidade de lançamentos observados. Assim, um experimento com n lançamentos pode ser conduzido com a hipótese nula (H_0) de que a afirmação do jogador é falsa até o término do experimento. Se o jogador acertar todas as n vezes, pode-se assumir que ele está dizendo a verdade (H_1) com uma probabilidade de erro de exatamente 2^{-n}.

Porém, se fossem realizados m lançamentos (com $m > n$), eventos ainda mais raros poderiam ocorrer. Por isso, o p-valor é uma medida que indica a probabilidade de se observar resultados tão extremos quanto, ou mais extremos que, os obtidos por um experimento com n lançamentos, assumindo que a hipótese nula é verdadeira. Para este problema, a expressão do p-valor é dada pela Equação 6. Por exemplo, para $n = 600$, com um tempo médio de 8 segundos entre os lançamentos, o experimento duraria uma hora e vinte minutos, e o p-valor seria de 2^{-599} (aproximadamente $4{,}82 \times 10^{-181}$).

$$p\text{-}valor = \sum_{k=n}^{\infty} 2^{-k} = 2^{1-n} \tag{6}$$

□

Problema Brinquedo 3: (Adaptado de [52]) Uma pessoa *A* diz ser capaz de, ao beber uma xícara de chá com leite, diferenciar se o chá ou o leite foi colocado primeiro na xícara. Desenvolva um experimento, monte hipóteses e verifique a veracidade da afirmação de *A*.

Solução: A solução proposta por Fisher, em [52], consiste em preparar 8 xícaras de chá, sendo metade com chá adicionado primeiro e a outra metade com leite adicionado primeiro, ordenando as xícaras de forma aleatória. A pessoa *A*, então, bebe de cada uma das 8 xícaras e deve separá-las em dois grupos, conforme qual ingrediente foi utilizado primeiro.

A hipótese nula considera que *A* não é capaz de discriminar as bebidas, enquanto a hipótese alternativa considera que *A* consegue. Ao calcular a probabilidade de discriminar as bebidas por acaso, obtém-se um valor de $\frac{1}{70}$. Isso ocorre porque existem 70 maneiras de escolher 4 elementos de um conjunto de 8 objetos sem considerar a ordem. Dentre essas 70 combinações, apenas uma delas resultaria em acertos para todas as xícaras. Formalmente, a probabilidade de acertar por acaso, p_{A_r}, é dada pelo inverso do número de combinações de 8 elementos tomados 4 de cada vez:

$$p_{A_r} = \frac{1}{\frac{8!}{4! \cdot 4!}} = \frac{1}{70}. \tag{7}$$

Assim, aceita-se a hipótese alternativa com probabilidade de erro de $\frac{1}{70}$ se *A* conseguir discriminar corretamente as bebidas. No entanto, novamente, ressalta-se que se o experimento fosse maior (com mais xícaras de chá), eventos mais raros que p_{A_r} poderiam acontecer. Para contabilizar as probabilidades desses eventos, a expressão do p-valor neste problema é dada pela Equação 8 (demonstração no Apêndice B), a qual, para $n = 4$, resulta num p-valor de aproximadamente 1,97%.

$$p\text{-}valor = \sum_{k=n}^{\infty} \frac{(k!)^2}{(2k)!} = \frac{2}{27}\left(18 + \sqrt{3}\pi\right) - \sum_{k=0}^{n-1} \frac{(k!)^2}{(2k)!} \tag{8}$$

□

Ao analisar os dois exemplos anteriores, nota-se que foi necessário calcular probabilidades a partir dos resultados obtidos (ou esperados) dos experimentos. Esses valores são chamados de estatísticas de teste e medem o grau de concordância entre uma amostra e a hipótese nula [53]. Ao mesmo tempo, escolheram-se os níveis de significância α de forma que fossem iguais às estatísticas de teste. No entanto, é prática comum na literatura utilizar um nível de significância de 5%, de forma que se o valor da estatística de teste for menor ou igual ao nível de significância, aceita-se a hipótese alternativa com risco de erro α [53].

É crucial definir α antes da coleta de dados para evitar vieses. Nos dois problemas brinquedos anteriores, os p-valores foram diretamente interpretados como probabilidades de erro, mas, na prática, α atua como um limiar predeterminado para controle de falsos positivos.

Ao tomar decisões baseadas nos resultados dos testes estatísticos, é fundamental estar ciente dos tipos de erros que podem ocorrer. Existem dois principais tipos de erro associados à tomada de decisões em testes de hipóteses: o erro Tipo I e o erro Tipo II [49]. A Tabela 1 demonstra como esses erros acontecem.

Tabela 1 - Tipos de erro que podem acontecer ao se aceitar ou rejeitar a hipótese nula.

	Realidade	
Decisão	H_0 verdadeira	H_0 falsa
Aceita H_0	Decisão correta	Erro Tipo II
Rejeita H_0	Erro Tipo I	Decisão correta

Fonte: Adaptado de [56].

O erro Tipo I, também conhecido como "positivo falso", acontece quando rejeita-se hipótese nula quando, na verdade, ela é verdadeira. A probabilidade de cometer esse erro é denotada por α, o nível de significância do teste. Por outro lado, o erro Tipo II, ou "negativo falso", ocorre quando não rejeitamos a hipótese nula e a hipótese alternativa é verdadeira, resultando em falhar em detectar um efeito ou diferença que realmente existe. A probabilidade de cometer um erro Tipo II é denotada por β. Esses erros estão intrinsecamente relacionados, e o equilíbrio entre α e β muitas vezes exige ajustes no tamanho da amostra e na escolha do nível de significância para otimizar a precisão e a confiabilidade dos resultados dos testes [50].

Em ambos os problemas brinquedos apresentados, foram utilizadas distribuições diferentes para modelar as probabilidades dos eventos raros – aqueles que, se ocorressem, fariam aceitar a hipótese alternativa. No primeiro caso, utilizou-se a distribuição binomial para modelar os lançamentos da moeda, enquanto no segundo, utilizou-se uma distribuição baseada nos coeficientes do binômio de Newton para determinar as chances de discriminação correta das xícaras. A escolha da distribuição a ser utilizada sempre dependerá da natureza do problema em questão. Em especial, neste trabalho, foi utilizado o teste com distribuição χ^2 (lê-se "qui-quadrado"), descrito no tópico seguinte, por conta da natureza dos dados coletados, descritos na seção 4.

3.4.1 Teste de hipótese χ^2

O teste qui-quadrado é uma ferramenta estatística usada para dois propósitos principais: (a) testar a independência entre duas ou mais variáveis categóricas e (b) verificar a adequação de uma distribuição observada em relação a uma distribuição esperada (teste de qualidade de ajuste) [54]. Este teste é adequado para dados categóricos, como grupos de indivíduos classificados em categorias (por exemplo, sexo, estado de depressão, etc), e também pode ser aplicado a variáveis discretas numéricas quando estas são agrupadas em categorias (por exemplo, idade ≥ 18 ou idade < 18) [55]. No entanto, não é apropriado para dados paramétricos contínuos, como altura ou peso, que requerem métodos estatísticos diferentes [56].

Para a aplicação válida do teste qui-quadrado, alguns pressupostos devem ser atendidos. Primeiramente, os dados devem ser coletados de maneira aleatória a partir da população, garantindo que a amostra seja representativa [54]. Além disso, as frequências esperadas em cada célula da tabela de contingência devem ser de pelo menos 5, e não deve haver células com contagens zero. Se esses requisitos não forem atendidos, o teste pode não ser válido e levar a resultados imprecisos [57]. O tamanho da amostra também deve ser suficientemente grande para evitar erros do tipo II, recomendando-se um mínimo que varia de 20 a 50 observações [55]. Finalmente, as variáveis devem ser mutuamente exclusivas, ou seja, cada variável deve ser contada apenas uma vez em uma categoria específica e não deve aparecer em outras categorias [54].

Dito isso, a tabela de contingência é uma matriz que exibe as frequências observadas de diferentes combinações de categorias de duas variáveis. Por exemplo, ao investigar a

relação entre fumantes e a presença de uma doença, a Tabela 2 de contingência pode ser montada da seguinte forma:

Tabela 2 - Exemplo de tabela de contingência.

	Doente	Saudável	Total_L
Fumante	30	70	100
Não fumante	10	90	100
Total_C	40	160	200

Fonte: Autoria própria.

Nesta tabela, as frequências observadas são contadas a partir dos dados coletados. Para calcular as frequências esperadas, assume-se que não há associação entre as variáveis e utiliza-se a fórmula

$$E_{ij} = \frac{Total_L(i) \cdot Total_C(j)}{Total\ Geral} \tag{9}$$

em que E_{ij} representa a frequência esperada para a célula na linha i e coluna j [58]. As frequências esperadas são então comparadas com as frequências observadas para calcular a estatística qui-quadrado, cuja qual é descrita pela Equação 10 [55].

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{\left(O_i - E_i\right)^2}{E_i} \right) \tag{10}$$

Na Equação 10, O_i são as frequências observadas e E_i são as frequências esperadas para a i-ésima célula da tabela de contingência. Essa fórmula calcula a soma dos quadrados das diferenças entre as frequências observadas e esperadas, normalizadas pelas frequências esperadas. A distribuição qui-quadrado é uma distribuição de probabilidade contínua que surge quando se somam os quadrados de variáveis aleatórias independentes com distribuição normal padrão $N(0,\ 1)$, tendo sua densidade de probabilidade com d graus de liberdade expressa por

$$f(x;\ d) = \frac{x^{(d/2)-1}\, e^{-x/2}}{2^{d/2}\, \Gamma(d/2)}, \tag{11}$$

em que $\Gamma(.)$ é a função gama e d é o número de graus de liberdade [53]. Esta distribuição é assimétrica e tem uma cauda longa à direita, especialmente para poucos graus de liberdade. À

medida que os graus de liberdade aumentam, a distribuição se aproxima de uma normal, o que facilita a interpretação dos resultados em grandes amostras ou amostras com muitas categorias [54].

Dessa forma, para testar a hipótese de independência entre duas variáveis, constrói-se uma tabela de contingência e calculam-se as frequências esperadas assumindo que as variáveis são independentes. A estatística do teste é obtida, então, utilizando a Equação 10 e o valor resultante é utilizado como entrada na distribuição qui-quadrado, dada pela Equação 11, juntamente com o número de graus de liberdade $d = (nº\ de\ linhas - 1) \cdot (nº\ de\ colunas - 1)$ [58].

Por outro lado, o teste de homogeneidade é usado para comparar a distribuição de uma variável categórica entre duas ou mais amostras independentes. Esse teste determina se as proporções de categorias são semelhantes entre as amostras. A estatística do teste e os pressupostos são idênticos aos usados no teste de independência [59].

Para interpretar os resultados dos testes qui-quadrado, comparamos a estatística calculada com o p-valor da distribuição qui-quadrado para um nível de significância específico. Se a estatística exceder o valor crítico, rejeita-se a hipótese nula, indicando uma diferença significativa entre as distribuições ou uma associação significativa entre as variáveis. Caso contrário, aceita-se a hipótese nula, sugerindo que não há evidências suficientes para afirmar uma diferença ou associação [55].

4. Metodologia

4.1 Coleta e Pré-Processamento de Dados

A linguagem *Python*[5] foi utilizada no desenvolvimento dos algoritmos que buscaram e armazenaram os *tweets*. Isto foi feito na plataforma de desenvolvimento *Visual Studio Code*[6] a qual permite criação de códigos de programação, gráficos, inserção de texto, equações matemáticas e *links* no formato de cadernos.

O processo de coleta de dados envolveu a busca por postagens de usuários brasileiros no *Twitter* que mencionaram explicitamente terem sido diagnosticados com depressão. Foram usados termos como “diagnosticado” (em ambos os gêneros) e “depressão”. A ferramenta de raspagem de dados *snscrape*[7] foi utilizada para realizar essa busca, retornando todas as

[5] PYTHON Software Foundation. *Python*. Disponível em: https://www.python.org/. Acesso em: 18 jul. 2024.
[6] MICROSOFT. *Visual Studio Code*. Disponível em: https://code.visualstudio.com/. Acesso em: 18 jul. 2024.
[7] JUSTANOTHERARCHIVIST. *Snscrape*. Disponível em: https://github.com/JustAnotherArchivist/snscrape. Acesso em: 18 jul. 2024.

ocorrências desses termos entre 11 de março de 2020 (início da pandemia de COVID-19 no Brasil) e 16 de outubro de 2022. Além do conteúdo dos tweets, também foram coletados diversos metadados, incluindo o horário de publicação, número de curtidas, compartilhamentos, menções e o tipo de dispositivo utilizado para realizar a postagem.

Os dados retornados pelo *snscrape* são estruturados em objetos do tipo dicionário. Para facilitar sua manipulação e análise, foi necessário convertê-los em tabelas com colunas correspondentes a cada tipo de dado coletado. A Tabela 3, abaixo, ilustra alguns dos principais campos desses objetos:

Tabela 3 - Alguns dos campos dos *tweets* retornados pelo *snscrape.*

Campo	Descrição
tweet.user.id	Número de identificação único do usuário que fez o *tweet*.
tweet.content	Texto do *tweet.*
tweet.date	Data e hora da postagem.
tweet.hashtags	Lista de hashtags incluídas na postagem.
tweet.retweetCount	Número de repostagens que a publicação recebeu.
tweet.replyCount	Número de respostas que o *tweet* recebeu.
tweet.likeCount	Número de curtidas.
tweet.sourceLabel	Aplicativo ou dispositivo de onde o *tweet* foi postado (por exemplo, "Twitter for iPhone").

Fonte: Autoria própria.

Com a identificação dos campos, foi desenvolvida uma rotina para automatizar a formatação e armazenamento dos dados num arquivo CSV. O código dessa rotina, ilustrado na Figura 6, utiliza o *snscrape* para coletar os tweets e a biblioteca Pandas para organizar os dados em um formato tabular.

```python
import snscrape.modules.twitter
import pandas as pd

def getTweetsSNS(search_words: str, since: str, until: str) -> pd.DataFrame:
    tweets = []
    for tweet in snscrape.modules.twitter.TwitterSearchScraper(f"{search_words} since:{since} until:{until}").get_items():
        tweets.append(
            [tweet.user.id, tweet.content, tweet.date, tweet.hashtags,
             tweet.retweetCount, tweet.replyCount, tweet.likeCount, tweet.sourceLabel]
        )

    tweets_df = pd.DataFrame(
        tweets, columns = ['userId', 'content', 'tweetCreated', 'hashtags', 'retweetCount', 'replyCount', 'likeCount', 'device']
    )

    return tweets_df


search_words = "diagnosticado depressão"
since = "2020-03-11" # Inicio da pandemia
until = "2022-10-16"
df = getTweetsSNS(search_words, since, until)
df.to_csv("dados_depressao.csv", index = False)
```

Figura 6 - Rotina para coletar os dados do *Twitter* utilizando a plataforma *Visual Studio Code*. Fonte: Autoria própria.

A função *getTweetsSNS* busca *tweets* com base nos termos definidos (*search_words*), além de coletar informações associadas como curtidas, *retweets* e *hashtags*, dentro do intervalo de tempo fornecido (entre *since* e *until*). Os dados são, então, organizados numa estrutura de dados do Pandas chamada de *DataFrame*, que possibilita a exportação dos dados para um arquivo CSV. O parâmetro *index=False* garante que a exportação não inclua uma coluna de índices, tornando o arquivo mais simples.

Em seguida, as declarações feitas pelos usuários foram verificadas manualmente por uma equipe de 4 colaboradores a fim de selecionar apenas os usuários que de fato disseram terem sido diagnosticados por um especialista da área da saúde mental, como psiquiatras ou psicólogos. Essa equipe foi formada por um engenheiro mecatrônico (o autor desta pesquisa), uma professora do ensino superior da área de tecnologia, um engenheiro civil e uma professora de inglês. Cada colaborador recebeu uma cópia do arquivo contendo os *tweets* e os metadados e todos os *tweets* foram analisados por todos os colaboradores.

Dessa forma, foram excluídos *tweets* que mencionaram terceiros, usavam ironia, humor, ou apresentavam indícios de golpes. Casos com discordância entre os avaliadores também foram descartados para evitar ambiguidades, e os tweets rejeitados foram marcados como positivos falsos. A Tabela 4 apresenta alguns exemplos de tweets removidos bem como o motivo da remoção.

Tabela 4 – Exemplos de *tweets* removidos.

Tweets	Motivo de retirada
Mas ele não só foi diagnosticado com depressão?	Falando de outra pessoa
Antes que você seja diagnosticado com depressão ou baixa auto estima, primeiro certifique-se que não esteja só cercado por idiotas.	Ironia
Já estava com vários sintomas, mas agora realmente fui diagnosticado com DPR: depressão pós reunião.	Brincadeira
@pessoa_famosa fui diagnosticado com depressão, mas não tenho dinheiro para comprar os remédios. Pode me ajudar?	Possível golpe

Fonte: Autoria própria.

As postagens removidas foram colocadas numa lista de falsos positivos, para serem utilizados como exemplos negativos na base de dados. Para tanto, foi utilizada uma amostra aleatória dos exemplos negativos, de forma que a base de dados ficasse balanceada (com o mesmo número de exemplos positivos e negativos). Depois disso, foi mantido apenas um *tweet* por usuário na base de dados. Vale ressaltar que os *user_id*s foram utilizados apenas para buscar as postagens públicas dos usuários, sendo, após a coleta, substituídos por um simples valor inteiro, sem relação com o *id* fornecido pelo *Twitter*.

Logo após a coleta e armazenagem dos *tweets*, os textos passaram por normalização com todas as letras tendo sido convertidas para minúsculas. Além disso, acentuações, pontuações, *emojis*, menções a usuários e links foram removidos. Em seguida, aplicou-se a técnica de radicalização (*stemming*), que reduz palavras a seus radicais (por exemplo, transformando “correndo” em “corr”), visando uniformizar termos com variações semânticas.

4.2 Representação Numérica dos Dados

A base de dados é multimodal, contendo dados textuais, numéricos e temporais, exigindo uma representação única para capturar essas características. Inspirado em [9], que usa uma matriz termo-documento para indicar a frequência de palavras em cada *tweet*, este trabalho adota uma abordagem similar, onde cada elemento da matriz indica a presença da k-ésima palavra no n-ésimo *tweet*.

A mesma abordagem foi utilizada para representar o tipo de aparelho utilizado para enviar as postagens. Já o horário de publicação foi convertido para um número inteiro no intervalo de 0 a 47, onde cada valor representa um intervalo de 30 minutos desde as 0h até às 23h59min59s. Por fim, os números de likes, compartilhamentos e menções, inicialmente em formato numérico, foram discretizados em categorias devido à grande variação de escala. Essa abordagem foi adotada para mitigar as distribuições assimétricas, onde a maioria dos

tweets possui, por exemplo, poucos *likes*, enquanto uma pequena amostra de usuários atinge milhares.

4.3 Seleção de Características

Para selecionar as características determinantes na detecção de depressão, foram utilizados testes de hipóteses do tipo qui-quadrado. O propósito é o de selecionar apenas as características cujas distribuições de probabilidade sejam estatisticamente diferentes entre os conjuntos de pessoas que disseram ter depressão e os que foram marcados como falsos positivos. Todos os pressupostos para utilização do teste χ^2 foram cumpridos, até mesmo para as variáveis numéricas como número de *likes*, compartilhamentos e menções, pois seus valores foram agrupados em categorias (ou *"bins"*).

Após a seleção, cada característica foi tratada como um classificador fraco de máxima verossimilhança, ou seja, um modelo simples que, individualmente, tem desempenho pouco superior ao acaso. Por exemplo, a presença de uma palavra específica ou uma faixa de curtidas poderia indicar uma leve tendência estatística de pertencer à classe positiva (pessoa diagnosticada com depressão). Para superar a limitação dos classificadores fracos, foi utilizado o algoritmo AdaBoost, que cria e ajusta uma combinação linear das respostas dos classificadores fracos.

De fato, o AdaBoost pode combinar um número arbitrariamente alto de classificadores, até mesmo repetindo alguns, de forma a memorizar todos os exemplos de treinamento, reduzindo, no entanto, sua capacidade de generalização. Esse fenômeno é comumente conhecido como *overfitting* e uma forma de evitá-lo é utilizando exemplos de validação [36].

Em especial, neste trabalho, foi utilizada a validação cruzada de k partições (*k-folds*), que divide os exemplos de treinamento em k subconjuntos (partições) e alterna quais deles são usados para treino e validação do AdaBoost em cada iteração, reduzindo o viés causado por divisões específicas dos dados [60]. O objetivo com isso foi encontrar o subconjunto de características mais relevantes na classificação de pessoas com depressão, sendo a importância medida a partir dos valores absolutos dos coeficientes atribuídos aos classificadores fracos pelo AdaBoost [46].

4.4 Métricas de Avaliação

Para avaliar o desempenho dos modelos criados é necessário utilizar métricas objetivas. Existem quatro números que podem ser observados diretamente da saída do

modelo e que representam de forma completa seu desempenho [10]. São eles os positivos verdadeiros (PV), positivos falsos (PF), negativos verdadeiros (NV) e negativos falsos (NF).

No entanto, na maioria das tarefas de classificação, são empregados índices mais fáceis de interpretar e que são calculados com os números acima. Estes são a acurácia, a precisão, a revocação e a medida F, cujas definições são apresentadas nas equações 12 a 15 [10, 11, 61].

$$Acurácia: A = \frac{PV+NV}{PV+NV+PF+NF} \tag{12}$$

$$Precisão: P = \frac{PV}{PV+PF} \tag{13}$$

$$Revocação: R = \frac{PV}{PV+NF} \tag{14}$$

$$Medida\ F: F = 2\frac{P \cdot R}{P+R} \tag{15}$$

A acurácia é a razão entre o número de acertos e o número de exemplos, a precisão é a taxa de exemplos positivos verdadeiros em relação ao número de exemplos que o modelo previu que eram verdadeiros, a revocação é a taxa de positivos verdadeiros sobre todos os exemplos que eram verdadeiros e a medida F é a média harmônica entre a precisão e a revocação [61].

4.5 Tamanho do Vocabulário

Existem várias formas diferentes de se verificar o tamanho do vocabulário de um *corpus*. Neste trabalho foram utilizadas duas, sendo elas, o número de palavras únicas e a quantidade efetiva de palavras. Embora ambas busquem mensurar a riqueza lexical do *corpus*, elas diferem em sua concepção e na forma como interpretam a diversidade do vocabularário.

A primeira contabiliza quantos termos distintos estão presentes no *corpus*, independentemente de quantas vezes cada palavra aparece. Já a segunda também leva em consideração a frequência relativa das palavras. Esse método é derivado do conceito de entropia na teoria da informação, adaptado para contextos linguísticos. Em específico, a Equação 16 contém a definição de quantidade efetiva de palavras utilizada nesta pesquisa [62]:

$$C = 2^{-\sum_{x \in V} p_d(x) \cdot log_2(p_d(x))} \quad (16)$$

Onde V é o vocabulário do *corpus* e $p_d(x)$ é a frequência relativa da palavra x ocorrer no discurso de uma pessoa com depressão. A Equação 16, conhecida como perplexidade no contexto de visualização de dados, quantifica a "surpresa" ou imprevisibilidade na distribuição das palavras [63]. Um valor baixo de C indica que o *corpus* é dominado por poucas palavras muito frequentes. Por outro lado, um valor alto de C sugere uma distribuição mais uniforme entre as palavras, com menor concentração de termos recorrentes.

4.6 Questões Éticas

Como mencionado anteriormente, os *ids* fornecidos pelo *Twitter* foram substituídos por valores numéricos arbitrários, de forma a não ser possível identificar os usuários que realizaram as postagens. Além disso, ao ingressar na rede social, os usuários assinam um termo em que concordam que suas postagens serão públicas[8], havendo a possibilidade de assim não sê-lo, caso desejem[9].

Dessa forma, por estar utilizando informações de acesso público, de acordo com artigo primeiro da Resolução nº 510, de 7 de abril de 2016, esta pesquisa não precisou ser registrada nem avaliada pelo sistema CEP/CONEP.

Ainda assim, apesar da natureza pública dos dados, a pesquisa foi conduzida com a devida consideração dos benefícios e riscos para os participantes. Entre os benefícios, é possível citar o aumento da literatura sobre como os transtornos depressivos se expressam em redes sociais e a possível criação de políticas públicas de prevenção e tratamento de casos depressivos. Enquanto risco, os participantes poderiam se sentir incomodados(as) por prestar informações acerca de seu estado emocional. No entanto, como mencionado anteriormente, a identidade dos participantes é preservada, especialmente após a criação da matriz termo-documento, que oculta a ordem das palavras nos tweets.

Dito isso, não foi oferecida qualquer forma de recompensa ou remuneração aos participantes. Os dados coletados foram utilizados exclusivamente para esta pesquisa, com a

[8] X CORP. *Política de Privacidade*. Disponível em: https://x.com/pt/privacy. Acesso em: 01 ago. 2024.
[9] X CORP. *Ajuda sobre Segurança e Privacidade: Postagens Públicas e Protegidas*. Disponível em: https://help.x.com/pt/safety-and-security/public-and-protected-posts. Acesso em: 01 ago. 2024.

divulgação dos resultados realizada por meio de artigos científicos em periódicos indexados e nesta dissertação e na publicação da base de dados em formato anonimizado[10].

5. Resultados e discussão

5.1 Representação Numérica dos Dados

Na construção da base de dados, foram coletados 5122 *tweets* com 4829 usuários diferentes. Destes, 1879 usuários foram incluídos como pessoas que disseram ter depressão. Foi coletada ainda uma amostra de 1879 pessoas que foram marcadas como falsos positivos na etapa de rotulação manual para que a base de dados ficasse balanceada.

Com a base montada, os dados multimodais vindos do *Twitter* foram, então, transformados para uma forma numérica única. Cada *tweet* foi transformado num vetor $\{0, 1\}^{5902}$, sendo o número de colunas igual ao número de palavras únicas da base. Já os dispositivos foram representados como vetores $\{0, 1\}^{20}$, ou seja, havia 20 tipos de dispositivos diferentes na base.

5.2 Seleção de Características

Para a seleção de características foram utilizados testes de hipóteses com $\alpha = 0{,}05$, que é um valor padrão na literatura. Foram selecionadas as palavras com os 100 menores p-valores, ou seja, as palavras com as maiores diferenças entre as distribuições de probabilidade (estimadas como frequências relativas) associadas a usuários que relataram ter depressão e aos falsos positivos. O Quadro 2 contém as dez palavras com os menores p-valores.

Quadro 2 - As 10 palavras com menor p-valor.

As dez palavras com menor p-valor
Fui
Eu
Ele
Foi
Pos
Me
Minha
Anos
Sou
Ansiedade

Fonte: Autoria própria.

[10] GUALBERTO, Ataíde. *Depression Indicators in Twitter*. 2025. Disponível em: https://data.mendeley.com/datasets/s25h5tzgyf/1. Acesso em: 12 mar. 2025.

Vale ressaltar que, por serem os termos de consulta utilizados no momento de obtenção dos dados, não houve diferença significativa entre as distribuições de frequência das palavras “diagnosticado” e “depressão”, pois ambas as palavras apareceram em todos os *tweets*. Além disso, nota-se que as palavras encontradas são, em sua maioria, referentes a conjugações em primeira e terceira pessoa, além de palavras de teor negativo como ‘ansiedade’. Este resultado é consistente com os achados de [4 – 6], que identificaram diferenças estatisticamente significativas no uso de termos em primeira pessoa entre indivíduos com depressão e o grupo controle. De acordo com os autores em [4], o uso frequente de pronomes dessa categoria pode indicar uma redução na riqueza de significados na fala, além de ser interpretado como um indicativo de empobrecimento semântico.

No caso dos dispositivos, apenas os do tipo *Android* passaram no teste de hipóteses com um p-valor < 2%, tendo como outros candidatos os sistemas *iPhone*, *Mac* e *iPad*, além de outros sites que conseguiam publicar *tweets* como *Instagram* e *WordPress*. A variável horário de publicação também apresentou relevância estatística com um p-valor de 2%. Finalmente, o número de compartilhamentos teve um p-valor < 0,02%. As outras características (número de *likes* e de menções) não apresentaram diferença estatística relevantes, com p-valores > 5%.

Dessa forma, foram criados conjuntos de treinamento e teste com 103 características aprovadas nos testes de hipóteses. O conjunto de treinamento recebeu 80% dos dados (3006 exemplos) e o de teste 20% (752 exemplos). O conjunto de treinamento foi dividido em 4 partes para validação cruzada, buscando evitar o *overfitting*. A Figura 7 apresenta a evolução da acurácia na partição de validação conforme aumenta o número de classificadores combinados pelo AdaBoost.

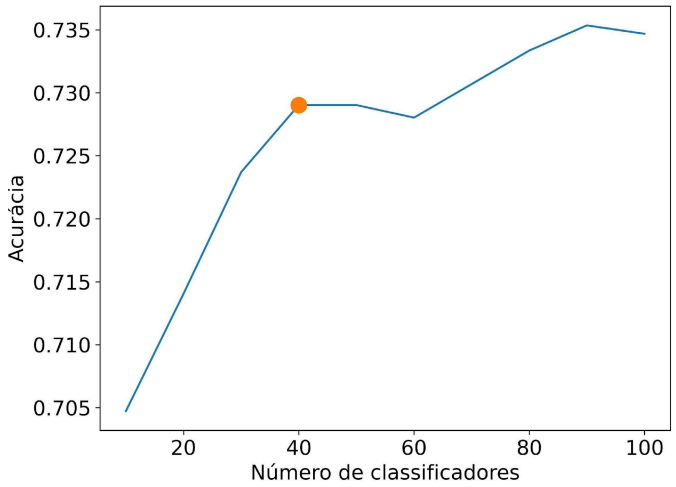

Figura 7 - Acurácia do AdaBoost no conjunto de validação versus o número de classificadores combinados.

Fonte: Autoria própria.

Nota-se que a partir de 40 classificadores a acurácia começa a diminuir no conjunto de validação antes de voltar a subir. Porém, a velocidade de aumento é muito menor, indicando uma diminuição na capacidade de generalização do modelo mesmo com o aumento do número de parâmetros. Com isso, foi treinado outro AdaBoost, com todos os exemplos de treinamento, com exatamente 40 classificadores e seu desempenho foi aferido no conjunto de teste. Os resultados encontram-se na Tabela 5.

Tabela 5 - Métricas de desempenho do AdaBoost no conjunto de teste.

Acurácia	**Revocação**	**Precisão**	**Medida F**
0,731	0,739	0,728	0,733

Fonte: Autoria própria.

Pode-se observar que todas as métricas giram em torno de 73%, sugerindo que o modelo comete praticamente a mesma quantidade de erros dos tipos I e II. Estudos correlatos [9 - 11] revelaram resultados similares em relação à medida F, variando de 73% a 79%. Entretanto, as demais métricas empregadas nessas pesquisas apresentaram variações significativas, com algumas delas (precisão e revocação) superiores às encontradas aqui neste trabalho, indicando que os algoritmos utilizados acarretaram em erros de um dos tipos em maior proporção que do outro.

Uma das características do AdaBoost é que ele atribui coeficientes lineares a cada uma das respostas dos classificadores. Observando a magnitude desses coeficientes, é

possível inferir suas importâncias. A Tabela 6 apresenta alguns dos classificadores mais importantes do AdaBoost com 40 classificadores.

Tabela 6 - Alguns dos melhores classificadores utilizados pelo AdaBoost com 40 classificadores.

Classificador	Posição	Importância
Fui	1º	10,39%
Eu	2º	6,92%
Pos	3º	5,84%
Ansiedade	4º	4,74%
Sou	5º	4,22%
Ele	6º	3,92%
Foi	7º	3,81%
Anos	8º	3,49%
[horário de publicação]	29º	1,47%
[dispositivo]	30º	1,37%

Fonte: Autoria própria.

Semelhante ao encontrado em outros trabalhos, nota-se que alguns dos indicadores mais importantes foram palavras em tempo verbal conjugados em primeira e terceira pessoa [4, 33]. Vale lembrar que os classificadores do tipo textual verificam se a presença de determinada palavra está mais associada a exemplos positivos ou negativos, mas só podem ser utilizados quando o texto contiver as palavras “diagnosticado” e “depressão”.

De forma semelhante, o classificador do tipo dispositivo verifica se a maioria dos *tweets* de treinamento foi enviada por um aparelho *Android*. Já o classificador referente ao horário de publicação da postagem verifica se num horário específico o número de pessoas que disseram ter depressão era maior do que os de falsos positivos. A Figura 8 mostra a saída do classificador do tipo horário para cada um dos horários do dia, enquanto a Figura 9 mostra as frequências absolutas utilizadas para criar o classificador do tipo dispositivo.

Apesar de parecer surpreendente, verificar a presença de certas palavras em um texto curto no *Twitter*, o horário da publicação e o dispositivo utilizado (como *Android*) permite criar um algoritmo que identifica se uma pessoa declarou ter depressão. O AdaBoost consegue combinar classificadores fracos, como a presença da palavra "fui", em um classificador forte, que atinge uma precisão maior do que qualquer um desses classificadores fracos isoladamente.

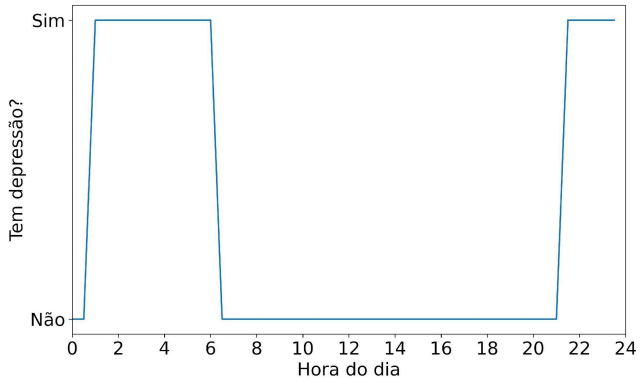

Figura 8 – Evolução da classificação de pessoas depressivas ao longo do dia. Fonte: Autoria própria.

Nota-se que das 21h30min até às 6h30min praticamente todos os intervalos são associados com a classe de pessoas depressivas, exceto das 0h às 0h30min, possivelmente devido a alguma especificidade da amostra coletada para este trabalho. Um intervalo similar, das 23h às 6h, foi encontrado em [64], que utilizou dados do *Twitter* na língua espanhola. Uma explicação para isso pode ter relação com o fato da população com depressão sofrer com distúrbios do sono[11]. Porém, é necessário lembrar que este classificador contribuiu apenas 1,47% nas decisões do AdaBoost. De fato, a acurácia do classificador horário é de apenas 52,9%, sendo levemente superior ao classificador binário aleatório.

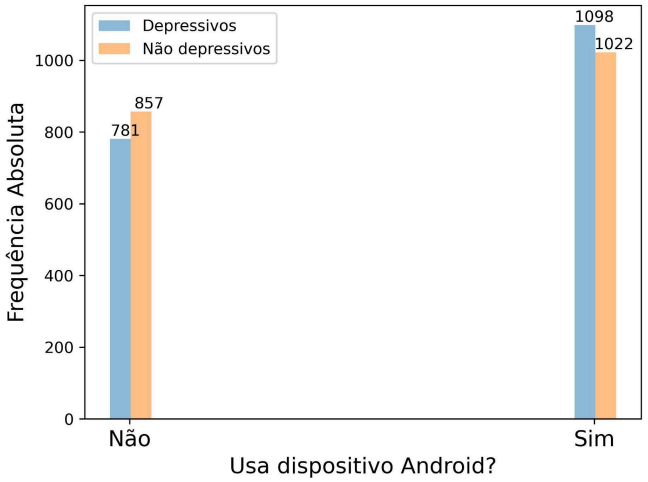

Figura 9 – Frequências absolutas utilizadas para construir o classificador do tipo dispositivo. Fonte: Autoria própria.

Ao analisar a Figura 9, é possível perceber como funciona a construção de um classificador de máxima verossimilhança. Dada uma entrada, neste caso binária, verifica-se

[11] Idem, nota 3.

qual é o rótulo mais frequente associado a essa entrada no conjunto de exemplos de treinamento e essa será a saída do classificador. Especificamente, no caso do classificador do tipo dispositivo, nota-se um número maior de usuários que publicaram uma postagem a partir de um dispositivo *Android* rotulados como depressivos, sendo essa a saída do classificador ao avaliar novos exemplos.

Uma hipótese para esse achado é que a condição socioeconômica dos usuários de dispositivos *Android* é pior do que a de usuários de outros sistemas operacionais, como o *IOS*. É necessário reiterar que este classificador possui apenas 1,37% de importância no AdaBoost e, dessa forma, deve-se ter cuidado com afirmações categóricas sobre este resultado.

5.3 Tamanho do Vocabulário

Na seção anterior, foi mostrado que o número de palavras únicas na base de dados é 5902. Ao realizar a contagem de palavras únicas das pessoas com depressão e sem depressão, foram encontradas, respectivamente, 3755 e 4382 palavras. Esses números, apesar de serem simples de obter, assumem que todas as palavras possuem a mesma probabilidade de ocorrer no texto, o que não é verdade.

Para contornar esse problema, foi utilizada uma métrica baseada na entropia de Shannon, descrita na Equação 16, para calcular o tamanho efetivo do vocabulário dos usuários com e sem depressão. Esse cálculo retornou 367 palavras para os depressivos e 430 para os não depressivos. Esses valores podem ser entendidos como uma média ponderada da quantidade de palavras utilizadas pelas pessoas com e sem depressão, respectivamente, refletindo a probabilidade ajustada de ocorrência no texto. A relevância estatística deste resultado foi aferida por um teste de hipóteses qui-quadrado que verificou se as distribuições de frequência de uso das palavras das pessoas com e sem depressão são iguais. Este teste retornou um p-valor $< 10^{-12}$, indicando que as distribuições são diferentes.

Esses resultados reforçam aqueles encontrados em [5], onde pessoas com transtorno depressivo maior demonstraram maior uso de frases curtas, truncadas, áridas, além de tautologias e elipses, em comparação com o grupo controle. Similarmente, em [4], os autores encontraram maior uso de repetições lexicais e semânticas, além do uso predominante de frases com apenas uma oração, no grupo de pessoas com depressão leve. Já em [64], os autores encontraram que o número de caracteres utilizados nas sentenças do grupo com depressão é menor do que o controle.

Uma explicação para o fato de diversos estudos encontrarem resultados similares, apesar das diferenças metodológicas, se dá por conta de que o cérebro de pessoas com depressão funcionar de forma diferente de uma pessoa saudável e isso influencia, entre outras coisas, as formas de comunicação da pessoa depressiva [31, 32]. De fato, o estudo de [5] descreve como mudanças linguísticas estão associadas a déficits na memória de trabalho, alternância de tarefas, planejamento estratégico, atenção e velocidade psicomotora, além de mostrar que essas mudanças na linguagem refletem a gravidade e a natureza dos déficits cognitivos presentes na depressão.

5.4 Limitações

Este estudo apresenta algumas limitações que devem ser claramente apontadas. A primeira delas é que a base de dados foi obtida inteiramente por pessoas que disseram ter sido diagnosticada com depressão e, apesar do esforço do autor e da equipe de colaboradores em separar os exemplos verdadeiros positivos dos falsos positivos manualmente, nenhum teste clínico foi aplicado. Dessa forma, não é possível saber se os usuários escolhidos realmente tinham depressão ou não. A segunda está relacionada com o fato de o número de termos utilizados para buscar postagens de pessoas com depressão foi limitado a apenas dois ("diagnosticado" e "depressão"), o que pode ter reduzido a cobertura de *tweets* retornados. A terceira se deve a falta de dados sócio-econômicos, idade, gênero e escolaridade, as quais sabidamente influenciam no diagnóstico de depressão. A quarta tem relação com a possibilidade de que os indivíduos selecionados para a base possam ter outras doenças além da depressão. Por fim, não foi verificado quais tipos de depressão os usuários possuíam.

6. Conclusões

Esta pesquisa evidenciou visibilizar as características relevantes na classificação de pessoas com depressão, utilizando dados do *Twitter* incluindo suas postagens, número de *likes*, menções e compartilhamentos, além do horário de publicação do *tweet* e do tipo de dispositivo pelo qual a postagem foi realizada. Testes de hipóteses e o AdaBoost foram utilizados para selecionar os melhores indicadores presentes nos dados.

Foi mostrado que, verificando a presença de determinadas palavras nos tweets, bem como o horário de publicação e se o dispositivo utilizado foi do tipo *Android*, é possível desenvolver um classificador capaz de simular a rotulação manual de exemplos verdadeiros positivos contra falsos positivos com acurácia e medida F de 73%, o que está próximo de resultados encontrados na literatura para detectores de depressão.

Baseado em trabalhos relacionados, investigou-se o tamanho do vocabulário de pessoas com depressão usando duas abordagens: contagem direta de palavras únicas nos *tweets* e entropia de Shannon, que considera a frequência das palavras. Constatou-se que o vocabulário é menor no grupo com depressão em comparação ao grupo controle em ambos os métodos analisados.

As contribuições principais deste trabalho estão nos fatos de que (i) apenas juntando classificadores fracos é possível simular a rotulação manual realizada por humanos e (ii) o tamanho do vocabulário de pessoas rotuladas manualmente com depressão é menor do que o da população rotulada como controle.

Essas descobertas têm implicações para a prática da saúde mental, já que, a partir de dados levantados, é possível criar ferramentas que detectam sinais de depressão com base em padrões identificáveis em postagens de redes sociais, permitindo que equipes de saúde pública possam tomar decisões informadas sobre como o dinheiro público pode ser utilizado. Portanto, a pesquisa não só promove um acréscimo da literatura sobre o reconhecimento automático de indicadores de depressão, mas também abre possibilidades para estratégias de intervenção mais direcionadas e efetivas, possibilitando um apoio direcionado para aqueles que mais necessitam.

Trabalhos futuros podem tentar encontrar características específicas para diferentes tipos de depressão. Outra possibilidade é a utilização de representações semânticas, muito utilizados em ferramentas de geração de texto e imagens como o ChatGPT[12] ou DALL-E[13], de forma a representar os dados numa dimensão que reflete mais fielmente a complexidade real desses dados. Por fim, é possível aumentar a quantidade de dados, utilizando a mesma plataforma (*Twitter*), ou buscando dados de outras redes sociais.

[12] OPENAI. *ChatGPT*. Disponível em: https://chat.openai.com/. Acesso em: 18 jul. 2024.
[13] OPENAI. *DALL-E*. Disponível em: https://openai.com/dall-e. Acesso em: 18 jul. 2024.

7. Referências

[1] BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Vigilância de Doenças Não Transmissíveis. *Vigitel Brasil 2020: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico: estimativas sobre frequência e distribuição sociodemográfica de fatores de risco e proteção para doenças crônicas nas capitais dos 26 estados brasileiros e no Distrito Federal em 2020*. Brasília: Ministério da Saúde, 2021. 124 p. ISBN 978-65-5993-122-4.

[2] AMERICAN PSYCHIATRIC ASSOCIATION. *Diagnostic and statistical manual of mental disorders*. 5. ed. Arlington, VA: American Psychiatric Publishing, 2013.

[3] SCHERER, K. R. Vocal communication of emotion: a review of research paradigms. *Speech Communication*, v. 40, n. 1-2, p. 227-256, 2003.

[4] SMIRNOVA, Daria et al. Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in psychiatry*, v. 9, p. 105, 2018.

[5] TRIFU, Raluca Nicoleta et al. Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. *Journal of Evidence-Based Psychotherapies*, v. 17, n. 1, 2017.

[6] RUDE, Stephanie; GORTNER, Eva-Maria; PENNEBAKER, James. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, v. 18, n. 8, p. 1121-1133, 2004.

[7] LIU, Yan et al. Predictors of depressive symptoms in college students: a systematic review and meta-analysis of cohort studies. *Journal of Affective Disorders*, v. 244, p. 196-208, 2019.

[8] DE CHOUDHURY, Munmun et al. Predicting depression via social media. In: *Proceedings of the international AAAI conference on web and social media*. 2013. p. 128-137.

[9] ALSAGRI, Hatoon S.; YKHLEF, Mourad. Machine learning-based approach for depression detection in Twitter using content and activity features. *IEICE Transactions on Information and Systems*, v. 103, n. 8, p. 1825-1832, 2020.

[10] ISLAM, Md Rafiqul et al. Depression detection from social network data using machine learning techniques. *Health information science and systems*, v. 6, p. 1-12, 2018.

[11] SANTOS, Wesley Ramos dos; DE OLIVEIRA, Rafael Lage; PARABONI, Ivandré. SetembroBR: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, v. 58, n. 1, p. 273-300, 2024.

[12] MENDES, Augusto R.; CASELI, Helena. Identifying Fine-grained Depression Signs in Social Media Posts. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024. p. 8594-8604.

[13] MANN, Paulo; PAES, Aline; MATSUSHIMA, Elton H. See and read: detecting depression symptoms in higher education students using multimodal social media data. In: *Proceedings of the International AAAI Conference on Web and social media*. 2020. p. 440-451.

[14] BECK, Judith S.; FLEMING, Sarah. A brief history of Aaron T. Beck, MD, and cognitive behavior therapy. *Clinical psychology in Europe*, v. 3, n. 2, 2021.

[15] BECK, Aaron T. et al. An inventory for measuring depression. *Archives of general psychiatry*, v. 4, n. 6, p. 561-571, 1961.

[16] KNAPP, Paulo; BECK, Aaron T. Fundamentos, modelos conceituais, aplicações e pesquisa da terapia cognitiva. *Brazilian Journal of Psychiatry*, v. 30, p. s54-s64, 2008.

[17] BECK, Aaron T. Thinking and depression: II. Theory and therapy. *Archives of general psychiatry*, v. 10, n. 6, p. 561-571, 1964.

[18] BECK, Aaron T.; ALFORD, Brad A. *O poder integrador da terapia cognitiva*. Porto Alegre: Artmed, 2000.

[19] BECK, A. T.; RUSH, A. J.; SHAW, B. F.; EMERY, G. *Terapia cognitiva da depressão*. Porto Alegre: Artmed, 1997.

[20] WRIGHT, J. H.; BASCO, M. R.; THASE, M. E. *Aprendendo a terapia cognitivo-comportamental: um guia ilustrado* (M. G. Armando, Trad.). Porto Alegre: Artmed, 2008.

[21] LEAHY, R. L. *Técnicas de terapia cognitiva: manual do terapeuta*. Trad. M. A. V. Veronese; L. Araújo. Porto Alegre: Artmed, 2006.

[22] KUYKEN, W.; PADESKY, C. A.; DUDLEY, R. *Conceitualização de casos colaborativa: o trabalho em equipe com clientes em terapia cognitivo-comportamental*. Trad. S. M. M. da Rosa. Porto Alegre: Artmed, 2010.

[23] WHITE, J. R. Introdução. In: WHITE, J. R.; FREEMAN, A. S. (Eds.). *Terapia cognitivo-comportamental em grupo para populações e problemas específicos*. São Paulo: Roca, 2003. p. 03-30.

[24] BECK, Aaron T. *Cognitive therapy and the emotional disorders*. Penguin, 1979.

[25] BECK, Aaron T.; HAIGH, Emily A. P. Advances in cognitive theory and therapy: the generic cognitive model. *Annual Review of Clinical Psychology*, v. 10, n. 1, p. 1-24, 2014.

[26] YOUNG, Jeffrey E.; KLOSKO, Janet S.; WEISHAAR, Marjorie E. *Terapia do esquema: guia de técnicas cognitivo-comportamentais inovadoras*. Porto Alegre: Artmed Editora, 2009.

[27] BECK, Judith S. *Terapia cognitivo-comportamental*. Porto Alegre: Artmed Editora, 2013.

[28] BROCKMEYER, Timo et al. Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in Psychology*, v. 6, p. 1564, 2015.

[29] CACIOPPO, John T.; VON HIPPEL, William; ERNST, John M. Mapping cognitive structures and processes through verbal content: the thought-listing technique. *Journal of Consulting and Clinical Psychology*, v. 65, n. 6, p. 928, 1997.

[30] SMIRNOVA, D. A. S30-01 Clinical Linguistics as the Basic Element of Methodological Knowledge in Psychotherapy. *Asian Journal of Psychiatry*, n. 4, p. S27, 2011.

[31] DISNER, Seth G. et al. Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience*, v. 12, n. 8, p. 467-477, 2011.

[32] ABDULLAEV, Yalçin; KENNEDY, Barbara L.; TASMAN, Allan. Changes in neural circuitry of language before and after treatment of major depression. *Human Brain Mapping*, v. 17, n. 3, p. 156-167, 2002.

[33] PYSZCZYNSKI, Tom; GREENBERG, Jeff. Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological Bulletin*, v. 102, n. 1, p. 122, 1987.

[34] STIRMAN, Shannon Wiltsey; PENNEBAKER, James W. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, v. 63, n. 4, p. 517-522, 2001.

[35] SCHAPIRE, Robert E. The boosting approach to machine learning: an overview. In: *Nonlinear estimation and classification*. 2003. p. 149-171.

[36] DOMINGOS, Pedro. A few useful things to know about machine learning. *Communications of the ACM*, v. 55, n. 10, p. 78-87, 2012.

[37] THEODORIDIS, Sergios; KOUTROUMBAS, Konstantinos. *Pattern recognition*. Elsevier, 2006.

[38] HAUSSLER, David; WARMUTH, Manfred. The probably approximately correct (PAC) and other learning models. In: *The Mathematics of Generalization*. 2018. p. 17-36.

[39] SCHAPIRE, Robert E. The strength of weak learnability. *Machine Learning*, v. 5, p. 197-227, 1990.

[40] SCHAPIRE, Robert E. et al. A brief introduction to boosting. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 1999. p. 1401-1406.

[41] BISHOP, Christopher M.; NASRABADI, Nasser M. *Pattern recognition and machine learning*. New York: Springer, 2006.

[42] ZHANG, Cha; MA, Yunqian. *Ensemble machine learning*. New York: Springer, 2012.

[43] ROJAS, Raúl et al. AdaBoost and the super bowl of classifiers: A tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep.*, v. 1, n. 1, p. 1-6, 2009.

[44] SHAHRAKI, Amin; ABBASI, Mahmoud; HAUGEN, Øystein. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost

and Modest AdaBoost. *Engineering Applications of Artificial Intelligence*, v. 94, p. 103770, 2020.

[45] DRUCKER, Harris. Improving regressors using boosting techniques. In: *ICML*. 1997. p. e115.

[46] WANG, Ruihu. AdaBoost for feature selection, classification and its relation with SVM, a review. *Physics Procedia*, v. 25, p. 800-807, 2012.

[47] LIU, Xiaoming; YU, Ting. Gradient feature selection for online boosting. In: *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007. p. 1-8.

[48] LIU, Huawen; LIU, Lei; ZHANG, Huijie. Boosting feature selection using information metric for classification. *Neurocomputing*, v. 73, n. 1-3, p. 295-303, 2009.

[49] LEHMANN, E. L.; ROMANO, J. P. *Testing statistical hypotheses*. 3. ed. Springer, 2005.

[50] RICE, John A.; RICE, John A. *Mathematical statistics and data analysis*. Belmont, CA: Thomson/Brooks/Cole, 2007.

[51] CASELLA, George L.; BERGER, R. L. *Statistical inference*. 2001.

[52] FISHER, Ronald Aylmer et al. *The design of experiments*. Edinburgh: Oliver and Boyd, 1966.

[53] WASSERMAN, Larry. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[54] AGRESTI, Alan. *Categorical data analysis*. John Wiley & Sons, 2012.

[55] MOORE, David S.; MCCABE, George P. *Introduction to the practice of statistics*. WH Freeman/Times Books/Henry Holt & Co, 1989.

[56] CONOVER, William Jay. *Practical nonparametric statistics*. John Wiley & Sons, 1999.

[57] MCHUGH, Mary L. The chi-square test of independence. *Biochemia Medica*, v. 23, n. 2, p. 143-149, 2013.

[58] EVERITT, Brian S. *The analysis of contingency tables*. CRC Press, 1992.

[59] SIEGEL, Sidney. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, 1956.

[60] BERRAR, Daniel. Cross-validation. In: *Encyclopedia of Bioinformatics and Computational Biology*, v. 1, n. April, p. 542-545, 2019.

[61] HOSSIN, Mohammad; SULAIMAN, Md Nasir. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, v. 5, n. 2, p. 1, 2015.

[62] MONTALVÃO, Jugurta et al. On the Representation of Sparse Stochastic Matrices with State Embedding. Available at SSRN 4605637.

[63] VAN DER MAATEN, Laurens; HINTON, Geoffrey. Visualizing data using t-SNE. *Journal of Machine Learning Research*, v. 9, n. 11, 2008.

[64] LEIS, A. et al. Detecting signs of depression in tweets in Spanish: behavioral and linguistic analysis. *J Med Internet Res.*, 2019; 21(6): e14199.

[65] SURY, B.; WANG, Tianming; ZHAO, Feng-Zhen. Identities involving reciprocals of binomial coefficients. *J. Integer Seq*, v. 7, n. 2, p. 04.2, 2004.

8. Apêndices

8.1 Apêndice A: Prova que α_t minimiza a função de custo do AdaBoost

A função de custo do AdaBoost é dada por:

$$E_t = \sum_{i=1}^{m} exp\left(- y_i \cdot H_t(x_i)\right)$$

O objetivo deste apêndice é mostrar que o valor de α_t que minimiza essa função de custo em cada iteração do algoritmo é:

$$\alpha_t = \frac{1}{2} ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right)$$

Prova

Dado um conjunto de treinamento $\{(x_1, y_1), ..., (x_m, y_m)\}$, com $y_i \in Y = \{-1, 1\}$, e um conjunto de classificadores fracos $\{h_1, h_2, ..., h_L\}$, em que $h_i(x_j) \in Y$, suponha que após a (t-1)-ésima iteração de treinamento do AdaBoost a saída do classificador final é

$H_{t-1}(x_i) = \alpha_1 h_1(x_i) + \alpha_2 h_2(x_i) + ... + \alpha_{t-1} h_{t-1}(x_i)$

em que o rótulo da classe é dado pelo sinal de H_{t-1}. Na t-ésima iteração, é esperado que H_t seja melhor do que H_{t-1} por ter sido adicionado mais um classificador fraco h_t com peso α_t:

$$H_t(x_i) = H_{t-1}(x_i) + \alpha_t \cdot h_t(x_i).$$

Para escolher α_t, derivamos a função de custo exponencial e igualamos a zero:

$$E_t = \sum_{i=1}^{m} exp\left(- y_i \cdot \left(H_{t-1}(x_i) + \alpha_t \cdot h_t(x_i)\right)\right)$$

Fazendo $D_1(i) = 1$ e $D_t(i) = exp\left(- y_i \cdot H_{t-1}(x_i)\right)$, obtém-se:

$$E_t = \sum_{i=1}^{m} D_t(i) \cdot exp\left(- y_i \cdot \alpha_t \cdot h_t(x_i)\right)$$

É possível dividir o somatório entre os exemplos corretamente e incorretamente classificados de forma que $y_i \cdot h_t(x_i) = 1$ e $y_i \cdot h_t(x_i) = -1$, respectivamente. Assim,

$$E_t = \sum_{y_i = h_t(x_i)} D_t(i) \cdot exp\left(-\alpha_t\right) + \sum_{y_i \neq h_t(x_i)} D_t(i) \cdot exp\left(\alpha_t\right)$$

$$E_t = \sum_{i=1}^{m} D_t(i) \cdot exp\left(-\alpha_t\right) + \sum_{y_i \neq h_t(x_i)} D_t(i) \cdot \left(exp\left(\alpha_t\right) - exp\left(-\alpha_t\right)\right).$$

Nesta última Equação, é possível perceber que o termo $\sum_{y_i \neq h_t(x_i)} D_t(i)$ é o único que depende de h_t, sendo este o motivo de que o classificador que minimiza E_t é aquele que minimiza $\sum_{y_i \neq h_t(x_i)} D_t(i)$. Para determinar α_t que minimiza E_t, deriva-se E_t em relação à α_t.

$$\frac{dE_t}{d\alpha_t} = -\sum_{y_i = h_t(x_i)} D_t(i) \cdot exp\left(-\alpha_t\right) + \sum_{y_i \neq h_t(x_i)} D_t(i) \cdot exp\left(\alpha_t\right) = 0$$

$$exp\left(\alpha_t\right) \cdot \sum_{y_i \neq h_t(x_i)} D_t(i) = exp\left(-\alpha_t\right) \cdot \sum_{y_i = h_t(x_i)} D_t(i)$$

$$\alpha_t + ln\left(\sum_{y_i \neq h_t(x_i)} D_t(i)\right) = -\alpha_t + ln\left(\sum_{y_i = h_t(x_i)} D_t(i)\right)$$

$$2\alpha_t = ln\left(\frac{\sum_{y_i = h_t(x_i)} D_t(i)}{\sum_{y_i \neq h_t(x_i)} D_t(i)}\right)$$

$$\alpha_t = \frac{1}{2} \cdot ln\left(\frac{\sum_{y_i = h_t(x_i)} D_t(i)}{\sum_{y_i \neq h_t(x_i)} D_t(i)}\right)$$

Sabendo que $\epsilon_t = \frac{\sum_{y_i \neq h_t(x_i)} D_t(i)}{\sum_{i=1}^{m} D_t(i)}$, obtém-se:

$$\alpha_t = \frac{1}{2} ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right)$$

E assim conclui-se a prova.

8.2 Apêndice B: Demonstração da expressão do p-valor do Problema Brinquedo 3

A probabilidade da pessoa *A* discernir corretamente ao acaso todas as 8 xícaras de chá é dada pelo inverso da quantidade de combinações de 8 elementos tomados 4 de cada vez. De forma geral, para um experimento com $2n$ xícaras, sendo metade preparada com leite primeiro e a outra com chá primeiro, a probabilidade de *A* discernir ao acaso será de

$$p_A = \frac{1}{\frac{(2n)!}{n! \cdot n!}} = \frac{(n!)^2}{(2n)!}.$$

No entanto, sabe-se que, caso realize-se experimentos maiores, eventos mais raros do que o associado a p_A poderiam acontecer. Dessa forma, para encontrar a expressão do p-valor deste problema, somam-se as probabilidades tão raras quanto, ou mais raras que, p_A. Assim, para um experimento com $2n$ xícaras a expressão do p-valor é dada por

$$p\text{-}valor = \frac{(n!)^2}{(2n)!} + \frac{((n+1)!)^2}{(2(n+1))!} + \frac{((n+2)!)^2}{(2(n+2))!} + \ \ldots \ = \sum_{k=n}^{\infty} \frac{(k!)^2}{(2k)!},$$

restando mostrar que $\sum_{k=n}^{\infty} \frac{(k!)^2}{(2k)!} = \frac{2}{27}\left(18 + \sqrt{3}\pi\right) - \sum_{k=0}^{n-1} \frac{(k!)^2}{(2k)!}$.

Em [65], o autor mostra que $f(x) = \frac{2x}{\sqrt{1-x^2}} sin^{-1}(x) = \sum_{k=1}^{\infty} \frac{(2x)^{2k}}{k} \cdot \frac{(k!)^2}{(2k)!}$. Diferenciando f em relação a x, obtém-se:

$$f'(x) = 2\left(x\sqrt{1-x^2} + sin^{-1}(x)\right)\left(1-x^2\right)^{\frac{-3}{2}} = \sum_{k=1}^{\infty} 2^{(2k+1)}\, x^{(2k-1)}\, \frac{(k!)^2}{(2k)!}$$

Para $x = \frac{1}{2}$, obtém-se:

$$f'\left(\frac{1}{2}\right) = 4\sum_{k=1}^{\infty} \frac{(k!)^2}{(2k)!} = 2\left(\frac{1}{2}\sqrt{\frac{3}{4}} + sin^{-1}\left(\frac{1}{2}\right)\right)\left(\frac{3}{4}\right)^{\frac{-3}{2}}$$

$$f'\left(\frac{1}{2}\right) = 4\sum_{k=1}^{\infty} \frac{(k!)^2}{(2k)!} = \left(\frac{4}{3} + \frac{8\pi}{9\sqrt{3}}\right)$$

Portanto

$$\sum_{k=1}^{\infty} \frac{(k!)^2}{(2k)!} = \left(\frac{1}{3} + \frac{2\pi}{9\sqrt{3}}\right)$$

Ao somar o termo correspondente a $k = 0$, neste caso, $\frac{(0!)^2}{(0)!} = 1$, obtém-se:

$$\sum_{k=0}^{\infty} \frac{(k!)^2}{(2k)!} = \left(\frac{4}{3} + \frac{2\pi}{9\sqrt{3}}\right) = \frac{2}{27}\left(18 + \sqrt{3}\pi\right)$$

Esta última expressão possui todas as parcelas de zero a infinito. No entanto, a expressão do p-valor começa a partir de n, sendo necessário apenas subtrair todas as parcelas do somatório de 0 a $(n - 1)$ para encontrar a fórmula do p-valor:

$$p\text{-}valor = \sum_{k=n}^{\infty} \frac{(k!)^2}{(2k)!} = \frac{2}{27}\left(18 + \sqrt{3}\pi\right) - \sum_{k=0}^{n-1} \frac{(k!)^2}{(2k)!}$$

E assim conclui-se a demonstração.